Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Segmentación de clientes activos de una entidad financiera empleando el algoritmo de K-means y árbol de decisión

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

César Enrique CISTERNA MOLLOCCO

ASESOR

Dra. Ofelia ROQUE PAREDES

Lima, Perú

2021

Referencia bibliográfica

Cisterna, C. (2021). *Segmentación de clientes activos de una entidad financiera empleando el algoritmo de K-means y árbol de decisión*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	César Enrique Cisterna Mollocco
Tipo de documento de identidad	DNI
Número de documento de identidad	72561586
URL de ORCID	
Datos de asesor	
Nombres y apellidos	Ofelia Roque Paredes
Tipo de documento de identidad	DNI
Número de documento de identidad	06243124
URL de ORCID	https://orcid.org/0000-0001-8280-021X
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Helfer Joel Molina Quiñones
Tipo de documento	DNI
Número de documento de identidad	40014631
Miembro del jurado 1	
Nombres y apellidos	Hugo Marino Rodriguez Orellana
Tipo de documento	DNI
Número de documento de identidad	40162362
Datos de investigación	
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad (Empresa, Instituciones, Poblaciones locales, regionales y nacionales)

Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Torre de Interbank País: Perú Departamento: Lima Provincia: Lima Distrito: La Victoria Avenida: AV. Carlos Villaran 140 Latitud: -12.0888 Longitud: -77.0199
Año o rango de años en que se realizó la investigación	Junio 2021 - Julio 2021
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACIÓN DE TRABAJO DE SUFICIENCIA PROFESIONAL EN LA MODALIDAD VIRTUAL PARA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA

En Lima, siendo las 16:00 horas del domingo 03 de octubre del 2021, se reunieron los docentes designados como Miembros del Jurado del Trabajo de Suficiencia: Dr. Helfer Joel Molina Quiñones (PRESIDENTE), Mg. Hugo Marino Rodriguez Orellana (MIEMBRO) y la Dra. Ofelia Roque Paredes (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**SEGMENTACIÓN DE CLIENTES ACTIVOS DE UNA ENTIDAD FINANCIERA EMPLEANDO EL ALGORITMO DE K-MEANS Y ÁRBOL DE DECISIÓN**", presentado por el señor **Bachiller César Enrique Cisterna Mollocco**, para optar el Título Profesional de Licenciado en Estadística.

Luego de la exposición del trabajo de suficiencia, el Presidente invitó al expositor a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del Jurado Evaluador, el expositor mereció la aprobación de **BUENO,** con un calificativo promedio de **DIECISEIS (16)**.

A continuación, los miembros del Jurado dan manifiesto que el participante **Bachiller César Enrique Cisterna Mollocco** en vista de haber aprobado la sustentación del Trabajo de Suficiencia Profesional, será propuesto para que se le otorgue el Título Profesional de Licenciado en Estadística.

Siendo las 16:30 horas se levantó la sesión firmando para constancia la presente Acta.

Dr. Helfer Joel Molina Quiñones
PRESIDENTE

Mg. Hugo Marino Rodriguez Orellana
MIEMBRO

Dra. Ofelia Roque Paredes
MIEMBRO ASESOR

La Vicedecana de la Facultad de Ciencias Matemáticas, Mg. Zoraida Judith Huamán Gutiérrez, certifica virtualmente la participación del Jurado Evaluador, el titulando, el acto de instalación y el inicio, desarrollo y término del acto académico de sustentación, dejando constancia en el acta respectiva.

Índice

Índice de tablas

Índice de figuras

A. Resumen

Actualmente la Institución Financiera ha identificado a clientes según su interacción con los canales físicos y digitales, entre clientes activos (42%) y clientes inactivos (58%), por lo cual es fundamental poder realizar acciones comerciales diferenciadas sobre este universo de clientes.

Se define como cliente activo a aquel cliente que realizó operaciones monetarias y no monetaria por canales digitales del banco dentro de los últimos seis meses o que realizan sus operaciones en canales físicos dentro de los últimos seis meses.

Debido a ello las áreas de negocio encargadas de realizar las campañas, decidieron priorizar la acción comercial en los clientes activos, lo cuales son alrededor de un millón setecientos mil clientes de manera mensual. Sin embargo, se desea realizar diferentes acciones comerciales según el perfil de los clientes activos puesto no todos tienen el mismo perfil.

Por lo cual, el presente trabajo consiste en la segmentación de clientes activos, el cual se desarrolló dentro del área de Business Analytics, área encargada del perfilamientos y segmentaciones de los clientes. Y mediante la segmentación, los responsables del negocio podrán realizar acciones comerciales que permitan gestionar los KPI's establecidos, que son el cross, el uso de tarjetas de crédito o débito y el aumento del uso de los canales digitales.

Esta segmentación permitirá conocer de manera acertada el perfil de los clientes activos, lo que permitirá ofrecer productos que calcen con las necesidades de los clientes activos, permitiendo incrementar sus KPI's.

B. Abstract

Currently, the Financial Institution has identified clients according to their interaction with physical and digital channels, between active clients (42%) and inactive clients (58%), which is why it is essential to be able to carry out differentiated commercial actions on this universe of clients.

An active customer is defined as a customer who carried out monetary and non-monetary operations through the bank's digital channels within the last six months or who carried out their operations through physical channels within the last six months.

Due to this, the business areas in charge of carrying out the campaigns decided to prioritize the commercial action in active clients, which are around one million seven hundred thousand clients on a monthly basis. However, you want to carry out different commercial actions according to the profile of active clients, since not all have the same profile.

Therefore, this work consists of the segmentation of active clients, which was developed within the Business Analytics area, the area in charge of profiling and segmentation of clients. And through segmentation, those in charge of the business will be able to carry out commercial actions that allow managing the established KPIs, which are the cross, the use of credit or debit cards and the increase in the use of digital channels.

This segmentation will allow to know in a correct way the profile of the active clients, which will allow to offer products that match the needs of the active clients, allowing to increase their KPI’s.

I. Introducción

La Entidad Financiera en la cual se realizó la segmentación de los clientes activos, es una de las cuatro entidades principales del Perú. Y dentro del área de Business Analytics se realizó la segmentación, esta área es de suma importancia debido a que se encarga de los conceptos y definiciones principales dentro del banco. Área en la cual me encuentro laborando año y medio, siendo mi principal función la elaboración de indicadores, perfiles de clientes y segmentación de clientes.

La Entidad Financiera busca conocer mejor a sus clientes para poder ofrecer productos que permitan satisfacer las necesidades de los clientes. Debido a ello, la entidad ha adoptado técnicas de minería de datos que ayudan a comprender patrones y comportamientos en grandes volúmenes de datos. Siendo la segmentación una herramienta apropiada que permitirá distribuir mejor a los clientes y conocer de forma precisa el perfil de los clientes, permitiendo realizar una gestión diferenciada en las campañas mensuales.

Debido a la coyuntura actual de pandemia, los responsables de negocio han establecido en mejorar o incrementar ciertos KPI's que son el cross, uso de tarjeta de crédito y débito además del uso de medios digitales, por ello requieren enviar campañas que estén personalizadas de acuerdo con las necesidades de los clientes.

Actualmente las campañas son enviadas sin distinguir el perfil que tienen estos clientes. Es por ello por lo que la elaboración de esta segmentación es de suma importancia para las principales áreas de negocio.

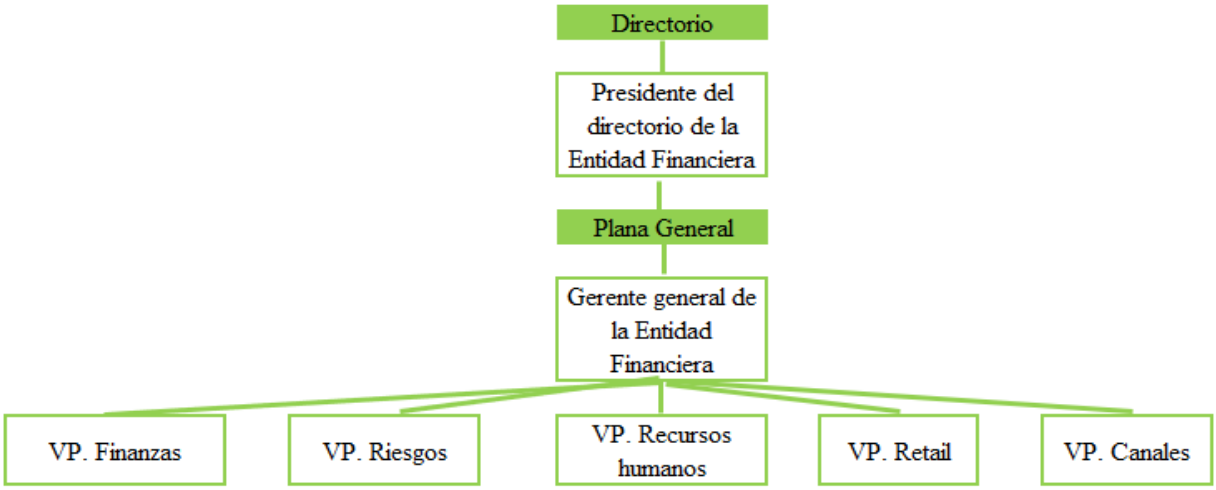
II. Descripción de la Actividad

2.1 Datos de la empresa

2.1.1 Nombre de la Institución: Institución Financiera
2.1.2 Periodo de duración del TSP: Mayo y junio 2021
2.1.3 Dirección: La Victoria
2.1.4 Correo electronico: confidensial
2.1.5 Área donde se realizó el TSP: Business Analytics
2.1.6 Organización de la empresa:

Figura 1

Organización de la empresa del año 2021.



Nota: Elaboración Propia

2.2 Reseña histórica

La Institución Financiera peruana fue creada el 1 de mayo de 1897. Iniciando actividades el mismo año. Siendo Elías Mujica quien asumió la presidencia del directorio. Cuyo primer local que entro en funcionamiento se ubicaba en la calle Espaderos, hoy Jirón de la Unión. En 1934 comenzó el proceso de expansión en las provincias de Chiclayo, Arequipa y Piura.

En el año 1942, la entidad tomo mayor fuerza de expansión, donde adquirió propiedad es en la Plazuela de la Merced y en la calle Lescano, en estos lugares se construyó la sede de “La Merced", y debido al buen acabado arquitectónico, el Instituto Nacional de Cultura lo catalogó a la sede como monumento histórico.

Carlos Rodríguez Pastos junto a un grupo financiero, entre ellos Nicholas Brady, el 20 de julio de 1994, adquirió el 91% de las acciones que se encontraban disponibles en aquel momento.

La entidad financiera cambio de nombre en el año de 1996, año en el cual la banca se renovó, con la finalidad que los clientes se sientan en un lugar confiable, lugar donde los clientes puedan encontrar productos de calidad, y serán bien atendidos por los representantes financieros.

En el año 2001, se inauguró la sede principal, ubica en la avenida Javier Prado y Paseo de la República, inició una nueva era, brindando mejores servicios y con tecnología de punta.

En el año 2012, inició lazos comerciales con representantes de Sao Paulo de Brasil, con el objetivo de asesorar a empresarios peruanos y brasileros, cabe recalcar que esto permitirá identificar oportunidades de inversión.

En la actualidad la entidad es una de las principales entidades financieras del Perú, que brinda productos que satisfacen las necesidades de los clientes, los cuales son mas de 4 millones, cabe recalcar que, debido a la pandemia, los clientes ahora interactúan más por el aplicativo del banco.

a) Misión

Mediante un servicio ágil, amigable en las diferentes sedes, se busca brindar la mejor calidad para los clientes.

b) Visión

A partir de los mejores colaboradores, ser el banco top.

c) Objetivos

- Lograr la preferencia de los clientes mediante una relación transparente y confiable.
- Que las expectativas de los clientes sean cubiertas por los productos y servicios financieros.
- Facilitar la vida financiera de los clientes mediante servicios de alta calidad, cordial y eficiente.
- Promover los valores en los colaboradores, promover el trabajo en equipo permitiendo desarrollar nuevas habilidades para mejorar los procesos, productos y servicios.

2.3 Descripción de la actividad

El presente trabajo consiste en la segmentación de clientes considerados como activos de la entidad financiera, que mediante esta técnica se podrá conocer mejor el perfil de los clientes dentro de cada clúster, lo que permitirá ofrecer un mejor producto que calce con la necesidad de los clientes.

2.4 Finalidad y objetivos

Actualmente las campañas destinadas a los clientes del banco no han dado los resultados esperados por parte de los responsables de las campañas, esto se debe principalmente a que el perfil de los clientes es totalmente diferente, donde se ha identifica a clientes activos e inactivos.

Donde se define como cliente activo a la recurrencia de uso de los diferentes canales físicos como digitales dentro de los últimos seis meses, donde las operaciones que realicen pueden ser monetarias y no monetarias. Debido a esta necesidad, se optó realizar una segmentación sobre los clientes activos de tal manera que se pueda incrementar KPI's claves establecidos por el área de campañas.

a) Objetivo general

   Realizar una segmentación de los clientes considerados como activos para realizar acciones comerciales diferenciadas.

b) Objetivos específicos

   1. Consolidar la base de datos con información de los clientes activos para realizar la segmentación.
   2. Estabilizar la segmentación para futuras campañas.
   3. Identificar los perfiles de los diferentes clústers para alinear los KPI's establecidos por el área de campañas para cumplir el objetivo comercial de la empresa.

c) Objetivo Comercial

   1. Elaborar acciones comerciales personalizadas según los perfiles encontrados en cada clúster alineado a los KPI's de negocio.
   2. Incremento los KPI's de cross, el uso de tarjetas de crédito y/o débito y el aumento del uso de los canales digitales en los diferentes clústers identificados.

d) Herramientas

Los diferentes softwares empleados para poder realizar la segmentación fueron:

- Python
- SQL
- Teradata
- AWS
- Office
- Alletion

2.5 Metodología

El primer paso es entender la necesidad de los responsables de negocio que realizan las diferentes campañas para los clientes de la entidad financiera. Luego de ello, se realiza la extracción de información correspondiente a los clientes activos de fuentes gobernadas disponibles en las bases de datos. La construcción de estas variables tiene una ventana de tiempo de un año hacia atrás, de las cuales se procede a la construcción de variables como promedios, desviaciones entre otras. Luego se procede a la limpieza de variables e identificar a las variables que se encuentren no correlacionadas para poder ser empleadas en el algoritmo de K-Means.

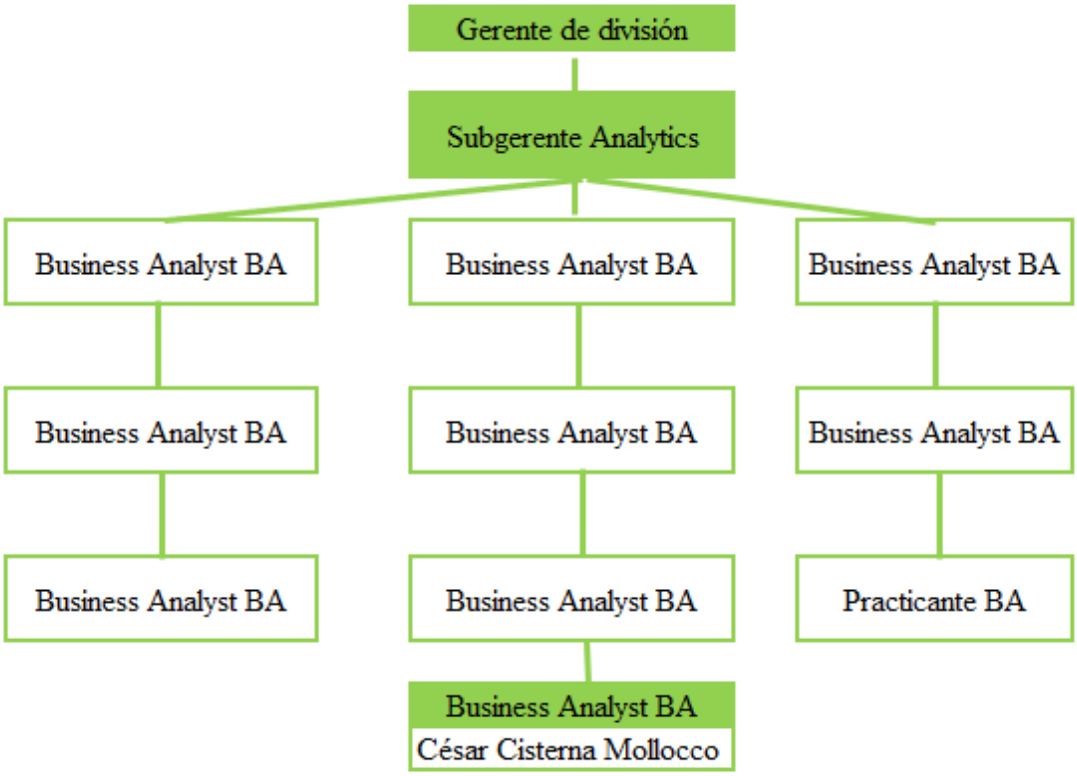
La metodología empleada es la Crisp - DM. Se identificará la combinación de variables con mejores indicadores en Silueta e Inercia. Posterior a ello se valida si la segmentación es estable en periodos previos. Luego de obtener el número de clúster se realizará un perfilamiento en cada clúster. Y finalmente la segmentación es presentada a los responsables de negocio, quienes son los encargados de realizar las acciones comerciales en las diferentes campañas futuras.

2.6 Organigrama

A continuación, se muestras la organización del área de Business Analytics en la cual me encuentro laborando y se realizó la segmentación de los clientes activos.

Figura 2

Organigrama del área de Business Analytics



Nota: Elaboración propia

III. Marco teórico

3.1 Minería de datos

Actualmente, las empresas almacenan grandes volúmenes de información que son generados de manera constante, los cuales están almacenado en diferentes repositorios de datos, estos al ser tratados de manera correcta permitirán tomar mejores decisiones de manera anticipada. Por ello es fundamental conocer estas técnicas que van a permitir descubrir patrones, comportamientos y tendencias en grandes volúmenes de información (Laude, 2017, p. 15 -18).

La minería de datos tiene como base principal a la estadística clásica, la cual contempla conceptos de estadística descriptiva, regresión, distribuciones, probabilidades, entre otros. Cabe recalcar que la estadística proporciona los métodos y procedimientos que permiten recolectar, organizar, analizar y presentar los datos (Mitacc, 2011, p. 1).

El segundo campo que considera es la inteligencia artificial, que nace en el campo de la informática, que consiste en la creación de programas y mecanismos que replican el comportamiento humano en diferentes tareas (Logreira, 2011, p. 12-13).

Como tercera área considerada es el aprendizaje automatizado o conocido de también como machine learnig. Esta rama toma los conceptos de la estadística y la inteligencia artificial que busca que los programas o máquina aprendan de bases de entrenamiento donde internamente encuentran patrones o tendencias, los cuales no requieren de hipótesis previas (Logreira, 2011, p. 68-71),

3.1.1 Correlación de Pearson

Sea (x1;y1), … (xk;yk), valores de las variables estadísticas bidimensional (X;Y), con frecuencias absolutas f1, f2, … , fk. El coeficiente de correlación muestral entre las variables X e Y se define como:

La correlación de Pearson se define como:

$$\text{Corr}[X;Y] = r = \frac{Sxy}{SxSy} = \frac{\sum_{i=1}^{k} f(xi-\bar{X})(yi-\bar{Y})}{\sqrt{\sum_{i=1}^{k} f(xi-\bar{X})^2}\sqrt{\sum_{i=1}^{k} f(yi-\bar{Y})^2}} \quad (1)$$

$$\text{Donde } S_{xy} = \text{Cov}[X;Y] = \frac{\sum_{i=1}^{k} f(xi-\bar{X})(yi-\bar{Y})}{n} \quad (2)$$

$$S_x = \sqrt{Sx^2} = \sqrt{\frac{\sum_{i=1}^{k} f(xi-\bar{X})^2}{n}} \qquad S_y = \sqrt{Sy^2} = \sqrt{\frac{\sum_{i=1}^{k} f(yi-\bar{Y})^2}{n}} \quad (3)$$

El coeficiente de correlación cuantifica la asociación entre variables, cuyo número está comprendido entre -1 a 1.

Cuando se obtiene un valor positivo, este indica que las dos variables aumentan o disminuyen al mismo tiempo, mientras que valores negativos indican que una variable aumenta y la otra disminuye o viceversa.

Si el valor es -1 ó a +1, significa que hay una perfecta asociación entre las dos variables, lo que significa, por cada unidad que se incrementa o se reduce una variable, la otra variable se ve afectada igual al número de unidades (Mitacc, 2011, p. 389).

3.1.2 Decil

Esta definición la encontramos en la estadística descriptiva, considerado como un estadístico de posición, el cual consiste en particionar en diez grupos iguales al conjunto de datos, la cual se define de la siguiente manera:

La fórmula del decil se define como:

$$D_k = L_i + \frac{\frac{K.N}{10} - F_{i-1}}{f_i} . a_i \quad (4)$$

Donde:

- Li es el límite inferior.
- N es la suma de las frecuencias absolutas.
- F_{i-1} es la frecuencia acumulada anterior a la clase del decil iésimo.
- a_i es la amplitud de la clase o longitud del intervalo correspondiente a la clase del decil i - ésimo.

Los deciles son muy importantes al momento de saber la posición de los datos (Mitacc, 2011, p. 71).

3.1.3 Técnicas de minería de datos

La principal tarea que realiza es la de encontrar patrones, perfiles y tendencias en grandes volúmenes información empleando técnicas como árboles de decisión, redes neuronales, algoritmos genéticos, random forest, entre otros (Laude, 2017, p. 24).

Según la tarea que realicen, estas se clasifican en descriptivas y predictivas.

Tabla 1

Técnicas de Minería de Datos

Técnicas Predictivas	Clasificación
	Regresión
Técnicas Descriptivas	Clustering
	Asociación

Nota: Elaboración propia

3.1.3.1 Modelos Supervisados o técnicas predictivas

Se encargan de predecir la clase a la cual pertenece un elemento o un valor numérico a la cual pertenece en base a ciertas variables propias de los elementos considerados en el estudio. Dependiendo de la tarea a realizar pueden ser de clasificación o de regresión (Bonaccorso, 2006, p. 14).

a. Clasificación

Se encarga de encontrar un modelo el cual pueda predecir comportamientos futuros a partir de datos ya conocidos, donde la tarea principal es la de encontrar grupos mutuamente excluyentes.

La clasificación se define como la tarea de predecir una categoría en un problema de clase discreta, por lo general son problemas de clasificación binaria. Cuando se trata de problemas de más de dos clases se denomina clasificación múltiple (Azoumana, 2013, p 4151).

Regresión

Consiste en predecir una variable continua a partir de la de otras variables de naturaleza continua. Esto es muy aplicado en temas correspondientes a ingresos, llamadas, ganancias, costos, etc (Azoumana, 2013, p 4151).

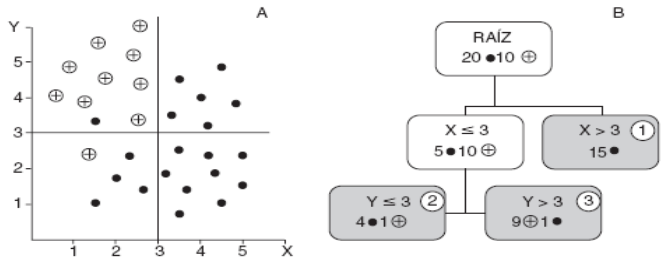
1. ÁRBOL DE DECISIÓN

Un árbol de clasificación representa una partición recursiva, la cual se basa en un conjunto de individuos. Los árboles de clasificación se encuentran dentro de los métodos de clasificación supervisada, por lo que se tendrá una variable dependiente. Todo árbol de clasificación comienza con un nodo al que pertenecen todos los datos de la muestra que se quieren clasificar, a este nodo se le conoce como nodo padre, mientras que el resto de los nodos son los nodos intermedios y nodos terminales. (Bonaccorso, 2006, p. 242).

Cabe recalcar que los árboles de clasificación se basan en ciertas reglas, los cuales se representan de manera gráfica y debido a su fácil comprensión son muy usados. Además, este algoritmo es usado en tareas de clasificación y regresión.

Figura 3

Árbol de Decisión



Nota: Google

Los árboles de decisión a medida que crecen generan diferentes nodos, los cuales tienen que ser evaluados, y esto se realiza mediante el índice de Gini y Ganancia de información, esta última se basa en la Entropía.

a. Índice de Gini

Es un indicador de "impureza" de los nodos, el cual busca identificar cuan mezclados se encuentran los nodos que están divididos. (Raschka, 2015, p 82 – 85)

b. Ganancia de información

Se emplea para atributos que son categóricos, el cual busca estimar la información que aporta los atributos el cual se basa en la teoría de la información. La medición de la aleatoriedad de incertidumbre de un valor aleatorio de una determinada variable se le denominada entropía. (Raschka, 2015, p 82 – 85)

c. Accuracy o exactitud

Este indicador cuantifica el porcentaje de aciertos que ha tenido y para poder realizar el cálculo se realiza mediante una matriz de doble entrada o mediante la matriz de confusión. (Bonaccorso, 2006, p. 168-173).

Tabla 2

Matriz de Confusión

		Predicción	
		0	1
Realidad	0	TN	FP
	1	FN	TP

Nota: Elaboración propia

Donde:

- TN: Verdadero negativo
- TP: Verdadero positivo
- FP: Falso positivo
- FN: Falso negativo

Cuya fórmula para calcular el accuracy es:

$$Accuracy: \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Existen diversos algoritmos que contribuyen a la creación de un árbol de clasificación, entre ellos tenemos al algoritmo CHAID y CARD

CHAID

El algoritmo CHAID (Chi Square automatic interaction detector), emplea el estadístico chi-cuadrado con la finalidad de identificar divisiones óptimas. Genera árboles binarios y no binarios. Este algoritmo puede trabajar con variables de naturaleza continua y categórica. Esto para las variables dependientes e independientes. Esto significa que este algoritmo realiza tareas de regresión y clasificación. (Bonaccorso, 2006, p. 242-258).

CART

Este algoritmo predice o clasifica observaciones hacia un futuro. Realiza particiones reiteradas excluyendo los registros de entrenamiento en grupos, los cuales minimizan las impurezas en cada paso donde un nodo w3, donde un nodo se considera puro si el 100% de las observaciones dentro del nodo corresponden a una categoría específica. Los elementos de entrada y objetivo pueden ser continuos. Este algoritmo solo genera particiones binarias, cabe

recalcar que este algoritmo realiza tareas de clasificación y regresión (Bonaccorso, 2006, p. 242-258).

3.1.3.2 Modelo no supervisado

En el campo laboral, los objetos no tienen asignado marcas de clases, por lo que mediante los algoritmos se buscará asignar o clasificar de manera correcta a los elementos que se encontraran en estudio (Bonaccorso, 2006, p. 17).

Estos algoritmos permiten identificar grupos que guardan parecido entre sí y diferencias respecto al resto de los grupos. El análisis de clúster conocido como segmentación de data tiene el objetivo de segmentar o agrupar a una colección de elementos o individuos en pequeños grupos de tal manera que al interior de cada clúster sean homogéneos y heterogéneos.

1. K-Means

Es un algoritmo que se emplea variables de naturaleza cuantitativas, donde se emplea la distancia euclidiana como medida de disimilitud [8]. Este algoritmo consiste en agrupar por vecindad, la cual parte de un conjunto de variables que pertenecen a un conjunto de individuos u objetos sin clasificación previa.

El algoritmo de K-means se encuentra dentro del aprendizaje no supervisada, que reune los objetos u elemento en k grupos, basándose en sus características. Este agrupamiento se desarrolla minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Para lo cual se usa la distancia cuadrática o distancia Euclidiana (Bonaccorso, 2018, p. 298-313).

El algoritmo consta de tres pasos:

a. Inicialización: Una vez elegido el número de grupos, se establecen *K* centroides en el espacio de los datos, que usualmente son elegidos de manera aleatoria.

b. Asignación: Cada objeto de los datos es asignado al centroide más cercano.

c. Actualización de centroides: Se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Así como en el aprendizaje supervisado se emplea indicadores como el GINI, cuando se emplea el algoritmo de K – means se emplea los indicadores de silueta e inercia, los cuales nos indicaran que tan homogéneos y heterogéneos son los clústers (Bonaccorso, 2018, p. 298-313).

a. Índice de la silueta

El índice tiene como tarea fundamental el de identificar el óptimo número de clúster en un algoritmo de aprendizaje no supervisado. Lo que permitirá concluir cuan similar es un objeto al resto de elementos que se encuentran dentro de un clúster.

Este índice está comprendido entre los valores de -1 a 1, donde un valor cercano a 1 indica que los clústers se encuentran homogéneos (Bonaccorso, 2018, p. 298-313).

$$S(i) = \frac{b-a}{\max(a,b)} \quad (6)$$

Donde:

- a es la distancia media entre el elemento y el resto de los elementos del grupo.
- b es la distancia media entre el elemento y todos los otros elementos del clúster más cercano.

El valor de S(i) es obtenido de la combinación de los valores de a(i) y b(i) como se muestra en la siguiente manera:

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, si\ a(i) < b(i) \\ 0, si\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1,\ si\ a(i) > b(i) \end{cases} \quad (7)$$

b. Índice de la inercia

Este indicador se calcula luego de haber empleado el algoritmo de K Means, el cual calcula suma de distancias al cuadrado de cada clúster a su respectivo centroide. (Bonaccorso, 2018, p. 298-313)

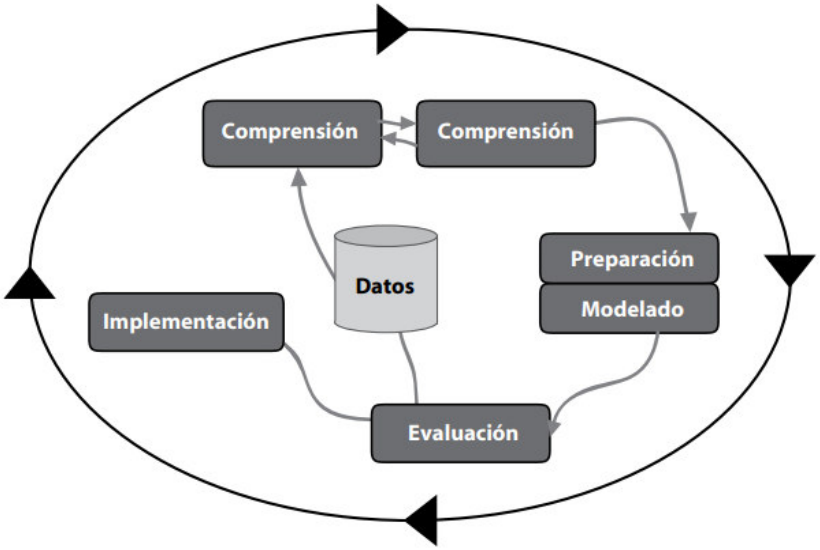
$$Inercia = \sum_{i=0}^{N} ||xi - \mu||^2 \quad (8)$$

c. La Metodología CRISP – DM

(Galan, 2015, p 21-34). En el año 1993 los representantes de SPSS, NCR y AG construyeron el acrónimo CRISP – DM (CROSS – INDUSTRY STANDARD PROCESS FOR DATA MINING) fue construido en base a experiencias de personas que trabajan en proyectos relacionados con minería de datos. Es la más empleada teniendo visibilidad de cada paso que se realiza, los que son conocidos por el negocio, conocimiento de los datos, preparación de los datos, modelado, evaluación y despliegue.

Figura 4

Metodología Crisp - DM



Nota: Google

1. **Comprensión del negocio:** esta parte consta en la comprensión de los objetivos del proyecto y los requisitos de la empresa.

2. **Comprensión de los datos:** consiste en la recolección de datos de diferentes fuentes, en los cuales se puede identificar datos atípicos, datos ausentes, esto permitirá tener un primer contacto con el caso de negocio.

3. **Preparación de los datos:** luego de haber realizado un análisis exploratorio de los datos con los que se cuenta se procede a dar el tratamiento correspondiente a los datos atípicos y a los datos ausentes mediante técnicas de imputación, esto se realiza con la finalidad de tener una base de datos limpia.

4. **Modelado:** luego de tener la base de datos lista se procede a emplear las diferentes técnicas de modelado.

5. **Evaluación:** luego de haber empleado diversas técnicas de modelado se procede a identificar si lo planificado se ha logrado.

6. **Implementación:** luego de haber evaluado los resultados de los modelos se procede a su posterior uso.

Actualmente diversas empresas e instituciones han incorporado la minería de datos al momento de tomar decisiones, entre los que destacan gobierno, empresas bancarias, universidades, hospitales, clubes deportivos entre otros. Cabe recalcar que la metodología CRISP es la más usada, y compañías como instituciones bancarias emplean esta metodología cuando van a afrontar problemas de índole analítica. Por se destacan algunas ventajas y desventajas.

Ventajas

1. Permite analizar grandes volúmenes de datos encontrando patrones y tendencias que por las técnicas tradicionales resultaría imposible hallar.
2. La minería de datos proporciona a los altos mandos de las empresas información importante que no sabían que existía y que puede ser aprovechada.
3. Mejora la relación de la empresa con los clientes.

Desventajas

1. Se requiere de tecnología de punta y personal capacitado.
2. Según el tamaño de las fuentes de datos, el procesamiento puede demorar.

Luego de haber obtenido los valores de la inercia, se procede a representarlos de manera gráfica lineal, la cual es conocido como el método del codo, en esta gráfica se debe de apreciar un cambio brusco en la inercia, el cual nos dirá el número óptimo de los clústers.

3.2 Segmentación de clientes activos

Debido al comportamiento que tienen los clientes, la entidad financiera ha priorizado realizar una gestión sobre los clientes activos, con la finalidad de poder incrementar KPI's importantes para los responsables de las campañas, los clientes activos se definen según su interacción de los clientes con los diferentes canales tanto físico y digitales en los últimos seis meses, donde se realizan operaciones monetarias y no monetarias.

Donde se ha identificado de manera mensual, que en promedio se ha registrado alrededor de un millón setecientos mil clientes activos. Debido a ello, al ser un número voluminoso de clientes, poder conocer de manera acertada a los clientes es fundamental, donde cada cliente tiene diferentes necesidades y esto se puede detectar mediante una segmentación de clientes lo que permitirá realizar acciones comerciales diferenciadas que cubran las necesidades de los clientes, siendo en esta oportunidad la segmentación de los clientes activos.

3.3 Antecedentes

3.3.1 Antecedentes nacionales

Según Salazar Gebol, Jimmy Stain(2015) en la tesis titulada "Segmentación de clientes en base a su comportamiento de consumo a través del modelo de segmentación K-Means en una entidad Bancaria" plantea como objetivo segmentar a los clientes de acuerdo al comportamiento de los consumos lo que permitirá un mejor direccionamiento de las campañas. Sin embargo, el tener una gran variedad de información disponible no permitía dificultaban el inicio de la segmentación, por ello se realizó un análisis factorial para la reducción de estas con la técnica de componentes principales. Se logró identificar a los clientes que realizan consumos por los

canales digitales, por lo cual se realizó la segmentación con el algoritmo de K-means, donde la segmentación final presento siete grupos.

Según Marchán Manay, Gasdaly Edith(2017) en la tesis titulada “Implementación de un sistema web utilizando algoritmo k-means para mejorar el proceso de reclutamiento y selección del capital humano en la empresa M. y C. Pariñas S.A., Talara” M. y C. Pariñas S.A”, menciona que es una empresa líder en el rubro de instalaciones industriales, centrándose en el rubro petrolero, por lo cual requiere de un sistema web que se encargue del proceso de reclutamiento de colaboradores empleando el algoritmo de k – means, esto permitirá administrar de manera optima las contrataciones, puesto que el algoritmo puede trabajar con grandes volúmenes de información. Y debido a los costos que genera que el personal se dedique al proceso de selección se opto por emplear este algoritmo, dicho algoritmo se encargará de identificar a los candidatos que calcen con las expectativas que tienen los contratistas de la empresa.

Según Lira Flores, A. L. A. (2018), en su tesis titulada “Determinación de patrones de comportamiento de consumo de agua potable con algoritmos de clusterización en la provincia de Andahuaylas”. Menciona en la de tesis, que el objetivo es identificar patrones de comportamiento de consumo de agua potable correspondiente de una entidad prestadora de servicios agua potable en el municipio de Chaka en la provincia de Andahuaylas, en el cual se emplea el algoritmo de k-means, empleando las variables de consumo de agua potable y la categoría a la que pertenece un determinado usuario, con el objetivo de que en cada clúster se identifique los diferentes patrones de comportamiento de consumo de agua, lo que permitirá

tomar mejores decisiones para optimizar el consumo de agua potable en la ciudad de Andahuylas.

3.3.2 Antecedentes internacionales

Según Sairy Fernanda Chamba Jiménez (2015, Ecuador) en la tesis titulada “Mineria de datos para segmentación de clientes en la empresa tecnológica Master” menciona que el agrupamiento se basa en pequeños segmentos donde cada uno contiene datos similares entre si y diferentes entre grupos, donde se empleó la metodología CRISP – DM dentro de tareas de minería de datos, donde el objetivo es clasificar a los clientes post venta para poder identificar el nivel de lealtad de los clientes, para lo que se realizó primero un análisis usando el modelo RFM, posterior a ello se emplea el algoritmo de K- means, SOM (Selft Organizing), y se elegirá cuyo algoritmo brinde mejores grupos los cuales sean de calidad.

IV. METODOLOGÍA

Debido al gran volumen de información con el cual se va a trabajar, nos vamos a apoyar en la metodología CRISP – DM, metodología popular en diferentes proyectos de minería de datos. Teniendo como punto de partida comprensión del negocio, siendo esta etapa fundamental antes de iniciar la elaboración de la segmentación, etapa en la cual se define la necesidad por parte de los responsables de las campañas. Luego, mediante un conjunto de pasos se procede a elaborar la segmentación de los clientes donde se identificará el perfil en cada clúster.

El perfilamiento es la parte final y más importante para los responsables de negocio, puesto permitirá identifica que clústers presentan buenos KPI's y cuales requieren de acciones comerciales más agresivas para revertir la situación.

Figura 5:

Metodología

Nota: Elaboracion propia

4.1 Comprensión del negocio

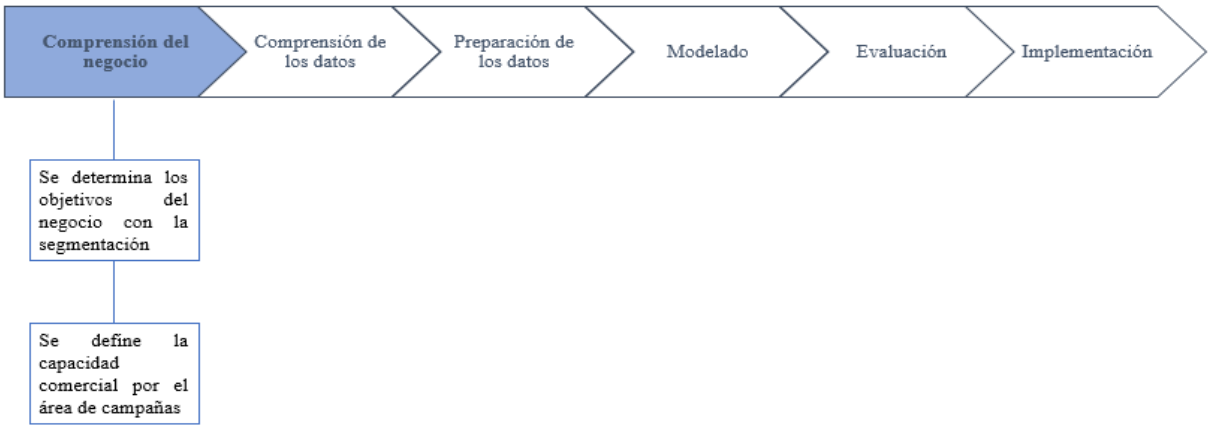
El primer paso es entender las necesidades que tienen los responsables del negocio, sus expectativas y conocer su capacidad comercial que se tiene para las futuras campañas mensuales.

Esta etapa consiste en una reunión entre el área de campañas, quienes manifiestan su necesidad de poder realizar acciones comerciales personalizadas al área de Business Analytics, área responsable de realizar las diferentes segmentaciones de la entidad financiera.

Los responsables de las campañas trasladan los objetivos que se desean alcanzar con la segmentación, la cual es la de poder incrementar KPI's claves para la entidad financiera, los cuales son el cross, uso de tarjetas de débito o crédito y el uso de los canales digitales. Adicional a ello, el área de Business Analytics considera fundamental conocer la capacidad comercial que tiene el área de campañas para futuras ventas al finalizar la segmentación.

Figura 6

Comprensión del negocio



Nota: Elaboración propia

Finalmente, en esta etapa se procede a construir en conjunto a los responsables de negocio un calendario de actividades.

Tabla 3

Organización de Actividades

N°	Actividad	Descripción	Días	Avance	Estado
1	Entendimiento de las Fuentes		5	0%	Pendiente
2	Construcción de features/ PO	2.1. Consolidación (Teradata)	6	0%	Pendiente
		2.2. Limpieza (Teradata o AWS)	2	0%	Pendiente
		2.3. Imputación/Vacios (Teradata o AWS)	1	0%	Pendiente
		2.4. Horizonte Temporal de features	2	0%	Pendiente
3	Validación de correlación de variables		2	0%	Pendiente
4	Segmentación	4.1. Selección de features significativos	3	0%	Pendiente
		4.2. Ejecución del modelo para las combinaciones	3	0%	Pendiente
5	Backtest	Validación de estabilidad en 3 cosechas	2	0%	Pendiente
6	Modelo Predictivo (Marca Cluster)		2	0%	Pendiente
7	Perfilamiento Excel		3	0%	Pendiente
8	Presentación PPT		2	0%	Pendiente
9	Revisión con el jefe de Business Analytics		2	0%	Pendiente
10	Revisión Gerente de Business Analytics		2	0%	Pendiente
11	Revisión Negocio		1	0%	Pendiente

Nota: Elaboración propia

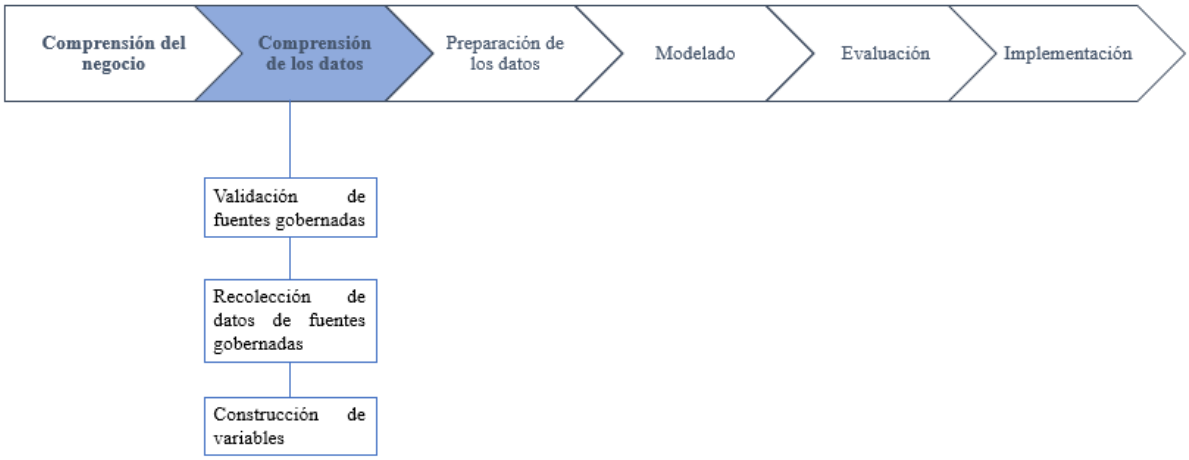
4.2 Comprensión de los datos

Esta etapa consiste en la recolección de los datos que se encuentran disponibles en las diferentes bases de datos de la entidad financiera correspondientes a los clientes activos de la entidad financiera. La información recolectada corresponde a consumos realizados con tarjetas de crédito y débito en diferentes rubros, operaciones realizadas por los diferentes canales físicos y digitales, información del sistema financiero, tenencias de productos, entre otros.

Las variables recolectadas tienen una ventana de un año de historia, de las cuales se puede generar otras variables derivables que pueden ser promedios, desviaciones, variaciones entre otros. Siendo alrededor de 250 variables que fueron construidas.

Figura 7

Comprensión de los datos



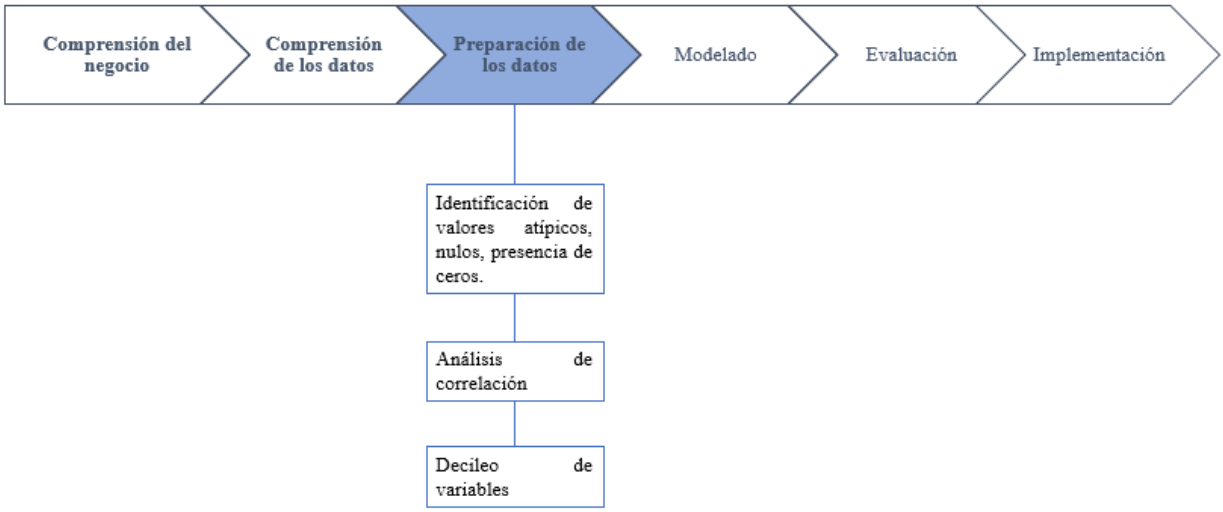
Nota: Elaboración propia

4.3 Preparación de los datos

Esta etapa consta de una serie de pasos, antes de realizar la segmentación que son la limpieza de datos, detección de valores atípicos, presencia de valores nulos, variables que presenten un 70% de valores ceros, los cuales seran excluidos al momento de emplear el algoritmo de segmentación, análisis de correlación y decileo de variables.

Figura 8

Preparación de los Datos



Nota: Elaboración propia

Como segundo paso dentro de esta etapa, se realiza un análisis de correlación entre las variables que están disponibles, se permite que las variables tengan como máximo una correlación hasta 0.30, las cuales serán empleadas en las diferentes combinaciones de variables que posteriormente generaran los clústeres mediante el algoritmo de k – means.

Se muestra las variables a emplear y su naturaleza:

Tabla 4

Descripción de las Variables

Variable	Tipo de variables	Escala
Número de operaciones totales realizadas en el aplicativo de la entidad financiera.	Cuantitativa	Ordinal
Monto promedio en consumos de plataformas digitales en los últimos seis meses.	Cuantitativa	Continua
Monto promedio de ingreso a las cuentas de los clientes en los últimos seis meses.	Cuantitativa	Continua

Nota: Elaboración propia

Solo se mostrará la correlación existente entre las variables que contribuyen a la construcción de la segmentación.

Tabla 5

Matriz de Correlación

	V1	V2	V3
V1	100%	14%	0%
V2	14%	100%	2%
V3	0%	2%	100%

Nota: Elaboración propia

Donde:

V1: Número de operaciones totales realizadas en el aplicativo de la entidad financiera.

V2: Monto promedio en consumos de plataformas digitales en los últimos seis meses.

V3: Monto promedio de ingreso a las cuentas de los clientes en los últimos seis meses.

Antes de emplear el algoritmo de K - means se procede a deciliar las variables, esto con la finalidad de que las variables que no estén altamente correlacionadas tengan las mismas unidades de medida al momento del cálculo de distancias empleado en uno de los pasos del

algoritmo de K means, por otro lado, aquellas variables que presenten valores 0, se le asignará la categoría de cero.

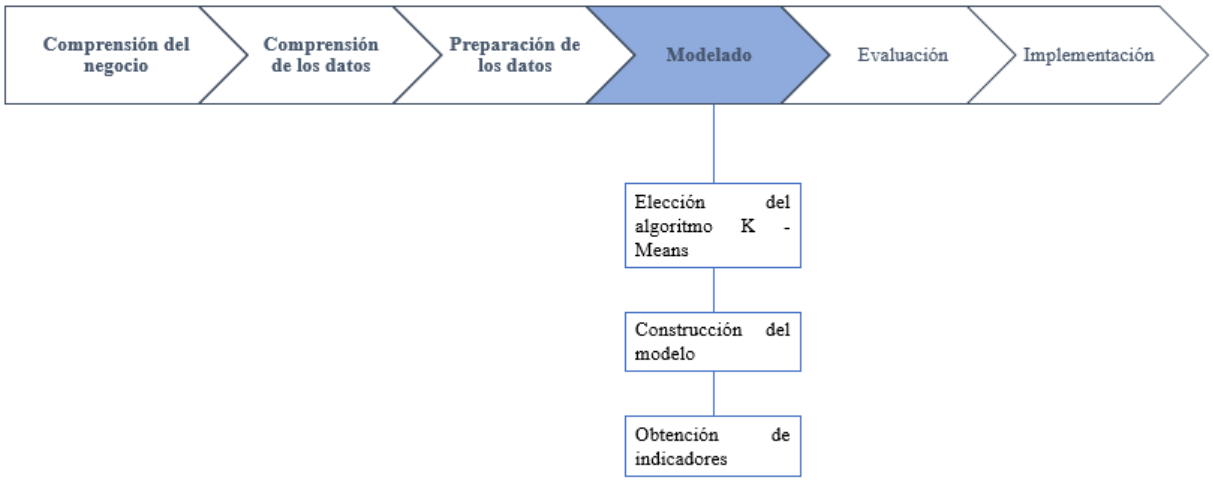
4.4 Modelado

En esta etapa, se selecciona el algoritmo que nos ayudara a la construcción de la segmentación de clientes activos, en esta ocasión se empleó el algoritmo de K – means, posteriormente se usará los indicadores de inercia y silueta que nos indican cuan homogéneo y heterogéneo son los clústers obtenidos.

Un punto importante es la elección del periodo que se va emplear para realizar el entrenamiento del algoritmo K – means, para este trabajo se trabajó con información correspondiente al mes de mayo de 2021.

Figura 9

Modelado de la Segmentación

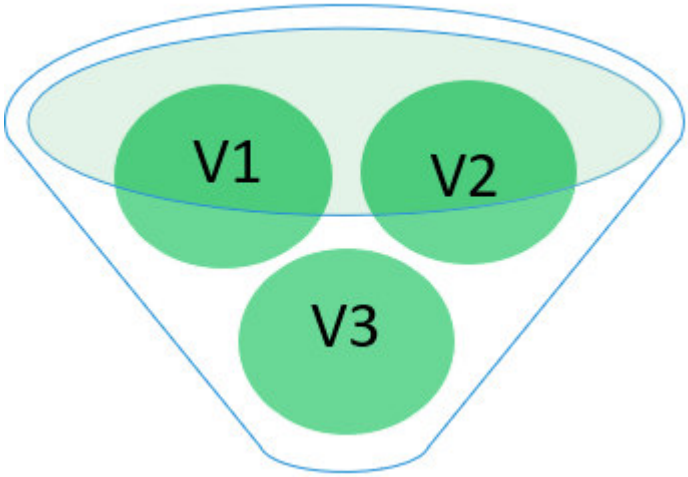


Nota: Elaboración propia

Por otro lado, se han probado diferentes combinaciones de variables para la obtención de los clústers, se han probado 450 combinaciones, donde la combinación de variables que resultó la más idónea para obtener la segmentación de cliente fue:

Figura 10

Variables de la Segmentación



Nota: Elaboración propia

Donde:

V1: Número de operaciones totales realizadas en el aplicativo de la entidad financiera.

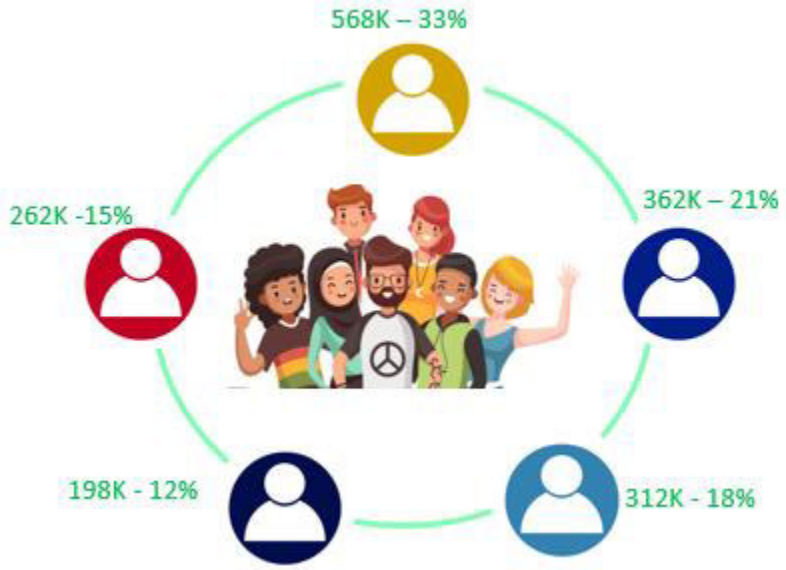
V2: Monto promedio en consumos de plataformas digitales en los últimos seis meses.

V3: Monto promedio de ingreso a las cuentas de los clientes en los últimos seis meses.

Donde se obtuvieron cinco clústeres correctamente distribuidos.

Figura 11

Distribución de los Clústers



Nota: Elaboración propia

Los indicadores obtenidos al emplear el algoritmo de k – means son:

Tabla 6

Indicadores de la Segmentación

Cosecha	# Clu	Inercia	Silueta	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
May-21	5	12,777,820	0.56	34%	22%	18%	15%	11%

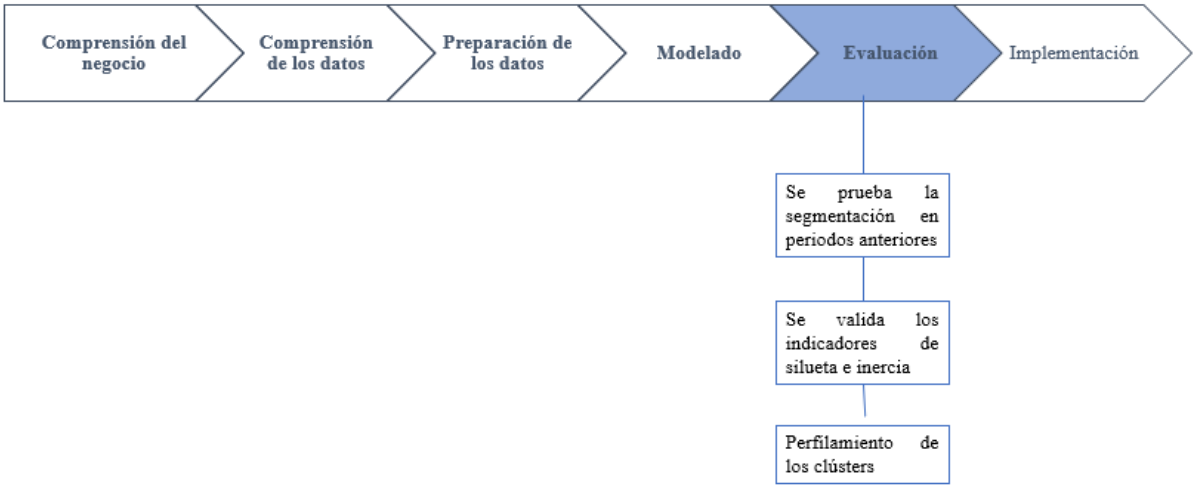
Nota: Elaboración propia

4.5 Evaluación

Como se ha obtenido la segmentación de los clientes activos correspondiente al mes de mayo de 2021, se procede a evaluar si esta segmentación es estable en periodos anteriores, esto con la finalidad de poder emplear la segmentación en próximas campañas con nuevos públicos objetivos.

Figura 12

Evaluación



Nota: Elaboración propia

Para ello, se realiza el cálculo de los indicadores de silueta, inercia y distribución de lo clústers, los cuales deben de tener una similitud en sus cifras al ser aplicado en los meses de enero, febrero, marzo y abril de 2021.

Tabla 7

Validación de la Segmentación en Diferentes Cosechas

Cosecha	# Clu	Inercia	Silueta	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
Ene-21	5	12,534,640	0.59	34%	21%	18%	15%	11%
Feb-21	5	12,852,590	0.57	34%	21%	19%	15%	11%
Mar-21	5	12,777,675	0.52	34%	22%	18%	15%	11%
Abr-21	5	12,946,375	0.56	34%	22%	17%	15%	11%
May-21	5	12,777,820	0.56	34%	22%	18%	15%	11%

Nota: Elaboración propia

Como se puede apreciar, los indicadores de inercia y silueta correspondientes a los periodos de enero, febrero, marzo y abril son muy parecidos a los que se han obtenido al mes de mayo 2021. Verificando de esta manera que la segmentación que se ha propuesta es estable y está lista para poder ser empleada en futuras campañas con nuevos públicos objetivos.

Luego que hemos verificado que la segmentación propuesta es estable, se procede a identificar el perfil de los clústers obtenidos, cabe recalcar que este perfilamiento permitirá a los responsables de las campañas gestionar las acciones comerciales para poder gestionar los KPI's que son prioritarios para los responsables de negocio.

Cabe recalcar que los valores mostrados corresponden a el número de clientes y el porcentaje que representa del total de clientes activos que se encuentran en un clúster, donde el símbolo K, hace referencia a miles,

Clúster 1 (262K, 15%)

- Presenta un ingreso mayor a 5 mil soles.
- Del total de clientes, el 91% son clientes multiproductos.
- Presenta altos consumos con sus tarjetas de crédito y débito.

Este clúster se caracteriza por ser el mejor clúster en comparación al resto de clústers.

Clúster 2 (362K, 21%)

- Es un clúster donde la mayoría de los clientes tienen productos pasivos.
- En este clúster se concentra el mayor saldo en los productos de ahorro.

Este clúster se caracteriza por concentrar el mayor saldo en monto de los productos de ahorro en comparación al resto de clústers.

Clúster 3 (198K, 13%)

- Los clientes presentan un ingreso promedio de S/2,500.
- El 65% se encuentra sobre endeudado.

Este clúster se caracteriza porque está sobre endeudado en el sistema financiero y presenta menor ingreso en comparación al resto de clústers.

Clúster 4 (312K, 18%)

- La edad promedio es de 33 años.
- El 60% son clientes multiproductos.
- El 70% se encuentra vinculado a la entidad bancaria
- El 90% realiza operaciones por los canales digitales.

Este clúster se caracteriza por ser el más joven y porque realiza operaciones por los medios digitales del banco que en comparación al resto de clústers sobresale en ello.

Clúster 5 (568K, 33%)

- La edad promedio es de 50 años.
- El 80% tiene el producto de plazo fijo.

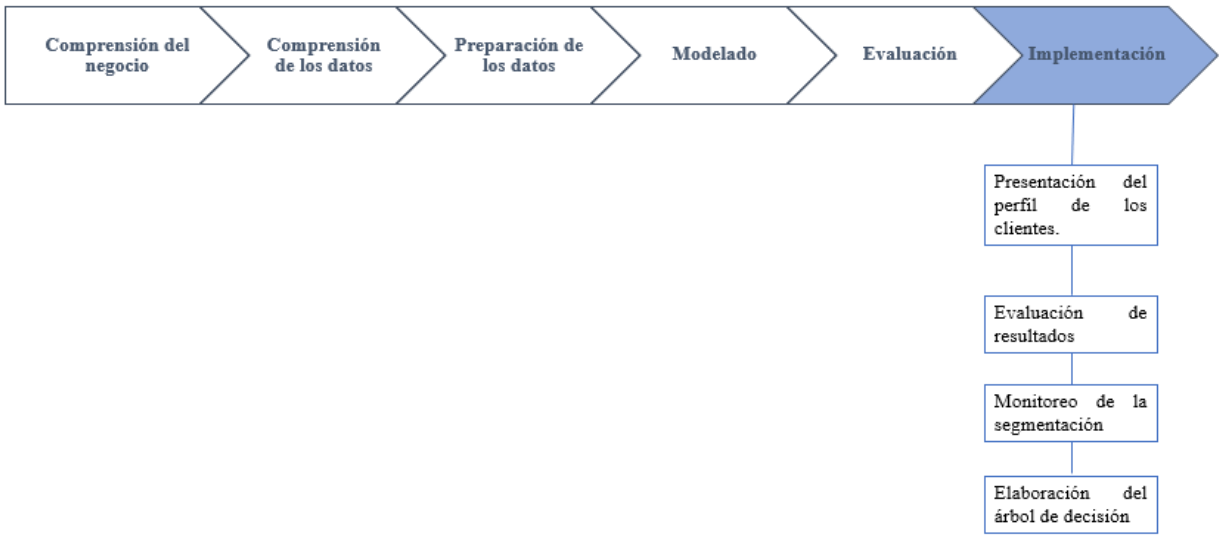
Este clúster se caracteriza por ser el más mayor de los cinco clústers y concentra el mayor monto en el producto de plazo fijo en comparación al resto de clústers.

4.6 Implementación

Una vez que se ha obtenido el perfilamiento de los clústers, se procede a presentar los resultados del perfilamiento a los encargados de campañas, quienes serán los encargados de elaborar las propuestas comerciales que calcen con los perfiles de los clientes de tal manera que se puedan incrementas los KPI's establecidos.

Figura 13

Implementación



Nota: Elaboración propia

Posterior a la reunión con el negocio, donde manifiestan su conformidad con la segmentación obtenida, se procede a la elaboración de un árbol de decisión, el cual será entrenado con la información del mes de mayo 2021. Esto se realiza debido a que el algoritmo de K – means al scorear nuevos públicos objetivos asigna diferentes marcas de clúster, pero se mantiene los perfiles de los clientes. Por ello se emplea el árbol de decisión multi clase para evitar este inconveniente.

Al evaluar el desempeño del árbol mediante la matriz de confusión, se observa la concentración de los valores en la diagonal en la matriz de doble entrada.

Figura 14

Matriz de Confusión

```
array([[170859,      0,      1,      0,      0],
       [     0,  91723,      0,      0,      0],
       [     0,      0,  60538,      0,      0],
       [     0,      0,      0,  79601,      0],
       [     0,      0,      0,      0, 108455]], dtype=int64)
```

Nota: Elaboración propia

En base a estos resultados se puede calcular el indicador del Accuaracy.

Figura 15

Accuracy

```
Accuaracy: 0.9999980437304495
```

Nota: Elaboración propia

Y como vemos, en base a los resultados obtenidos, podemos concluir que el algoritmo del árbol de decisión tiene un buen desempeño, el cual será empleado al momento de scorear nuevo públicos objetivos de manera mensual.

V. Conclusiones

Al concluir el presente trabajo correspondiente a la segmentación de los clientes activos de la institución financiera se ha llegado a las conclusiones siguientes:

- Se realizó la segmentación de los clientes activos identificando cinco grupos o clusters que permitiran a los responsables de campañas realizar acciones comerciales diferenciadas.
- Se consolidó la base de datos con información correspondiente a los clientes activos, esto se desarrollo gracias a la calidad de los datos que se encontraban en las bases de datos gobernadas.
- Se estabilizó la segmentaciones para ser empleada en futuras campañas, lo cual se logra concluir al observar el desempeño de la segmentación en periodos previos.
- Se ha obtenido cinco clústers, en los cuales cada grupo presenta características propias, que permitirá a los responsables de campañas identificar a que grupo realizar acciones comerciales más agresivas que ayuden a incrementar los KPI's de cross, uso de productos de Tarjeta de crédito o débito y uso de medio digitales.
- Mediante la segmentación se podrá conocer mejor las necesidades de los clientes, lo que permitirá mejorar los vinculas entre la entidad financiera y los clientes.
- Luego de obtener los cinco grupos o clústering, se implemento la solución analítica para ser empleada en futuras campañas.

VI. Recomendaciones

- Definir los objetivos que se quieren lograr con la segmentación junto a los responsables que realizan las campañas.
- Antes de ejecutar algún algoritmo de minería de datos, es fundamental validar y verificar que las fuentes de información provienen de fuentes de producción las cuales tienen que ser gobernadas.
- Se debe de realizar un preprocesamiento lo más profundo posible, lo cual contempla limpieza de variables, análisis de correlación.
- Si bien es cierto que existen diferentes variables disponibles, es importante que las variables candidatas para la elaboración de la segmentación, sean gestionables, es decir que sean conocidas por parte de los responsables de negocio.
- Cuando se realiza un modelo sea supervisado o no supervisado, se sugiere validar que el modelo sea estable en el tiempo para poder ser empleado en futuras campañas.

VII. Bibliografía

Henri Laude, (2017). *Data Scientist y lenguaje R,* pp15-18, 24.

Máximo Mitacc Meza (2011), *Tópicos de estadística descriptiva y probabilidad*, pp 1, 71, 389.

C. Logreira (2011), *Minería de datos y su incidencia en la, Ingeniería Solidaria,* vol. 7, nº 12-13, pp. 68-71.

K. Azoumana, (2013), *Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos, Pensamiento Americano*, p. 4151.

Giuseppe Bonaccorso, (2018), *Machine Learning Algorithms*, p 14-17, 168-173,282, 242-258, 298-313

Sebastian Raschka (2015), *Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive Analytics,* p 82 -85.

Víctor Galan Cortina (2015), *Aplicación de la metodología Crisp DM a un proyecto de Mineria de datos en el entorno universitario*, p 21 – 34.