Artículo

Empleo de algoritmos KNN en metodología multicriterio para la clasificación de clientes, como sustento de la planeación agregada

Multi-Criteria KNN Algorithms for Customer Classification as the Base of Aggregate Planning

Carlos Jesús Madariaga Fernández[1]* https://orcid.org/0000-0001-8194-2216
Yosvani Orlando Lao León[2] https://orcid.org/0000-0001-7491-3548
Dagnier Antonio Curra Sosa[3] https://orcid.org/0000-0001-5361-6536
Rafael Lorenzo Martín[4] https://orcid.org/0000-0001-6852-5725

[1]Departamento de Desarrollo de Sistemas, Universidad de Holguín, Cuba
[2]Facultad de Ciencias Empresariales, Universidad de Holguín, Cuba
[3]Facultad de Ingeniería, Universidad de Holguín, Cuba
[4]Dirección de Ciencia, Tecnología e innovación, Universidad de Holguín, Cuba
*Autor para la correspondencia: carlosjmadariaga@gmail.com

RESUMEN

Objetivo: Proponer una metodología multicriterio para la clasificación de clientes, considerando algoritmos KNN como sustento para la planeación agregada, a partir de una modificación al modelo RFM (recencia [o actualidad], frecuencia y monetario [valor momentario]).

Métodos y técnicas: se mostró una metodología para clasificar los clientes atendiendo a seis variables: fidelidad con la empresa, frecuencia con que realiza sus compras, valor patrimonial de los clientes, variedad de productos que compran, cercanía física y horizonte temporal de compras. **Principales resultados:** Una escala jerárquica de variables para la clasificación de clientes; igualmente se brindó una clasificación general de clientes añadida a la clasificación de estos según las

seis variables del estudio, lo que permite el análisis individualizado de su comportamiento.

Conclusiones: La aplicación de la herramienta metodológica en la empresa ACINOX Holguín comercializadora, validó su efectividad para resolver problemas de clasificación de clientes. Como derivado de su aplicación se proporcionó a los directivos de dicha institución, un conjunto de conglomerados individuales de cada variable que sustentó la planeación agregada, facilitó la toma de decisiones y optimizó el proceso de venta desde una visión general.

Palabras claves: algoritmos KNN, clasificación de clientes, planeación agregada

ABSTRACT

Aim: To propose a multi-criterion methodology for customer classification, considering the utilization of KNN algorithms as the base of aggregate planning, from an adjusted RFM model (recency [or currentness], frequency, and monetary [momentary value]).

Methods and techniques: A methodology was used to classify customers according to six variables: faithfulness to the company, purchasing frequency, customer's assets, variety of purchased items, physical nearness, and temporary purchasing horizon.

Main results: A hierarchical scale of variables to classify customers was designed. A general customer classification was added to their classification according to the six variables of the study, which allowed for a customized analysis of their performance.

Conclusions: The utilization of this methodological tool in ACINOX Holguin sales company validated its effectiveness to solve customer classification issues. Upon the application of this methodology, the executives of the institution had access to several individual clusters of each variable, which supported aggregate planning, enabled decision-making, and optimized the sales process under a general vision.

Keywords: KNN algorithms, customer classification, aggregate planning

Recibido: 15/03/2021

Aceptado: 28/02/2022

INTRODUCCIÓN

El complejo dinamismo en el que se desenvuelven las empresas comercializadoras y en específico las mayoristas (también conocidas como B2B), las ha llevado a buscar una mejor comprensión del comportamiento de sus clientes a través de formas innovadoras de almacenar y analizar la información sobre estos (Khajvand, Zolfaghar, Ashoori, y Alizadeh, 2011).

En este sentido uno de los principales desafíos de la gestión de las relaciones con los clientes (CRM, por sus siglas en inglés) es establecer relaciones más duraderas y rentables con clientes (Kord y Tavoli, 2015), a través de la recopilación, almacenamiento y análisis de los datos del cliente para aumentar su lealtad, valor y los beneficios organizacionales (Kandeil, Saad y Youssef, 2014).

Con el surgimiento de las tecnologías de la informática y las comunicaciones han ocurrido cambios importantes en la capacidad de las organizaciones para recopilar, almacenar y analizar grandes conjuntos de datos. Por lo tanto, se pueden almacenar más de miles de datos para cada cliente, lo que permite el análisis su historial de compras (Miguéis, Poel, Camanho y Cunha, 2012). Para comprender e identificar a los clientes según el valor que representan para las organizaciones, estas deben segmentarlos según su comportamiento.

A fin de aumentar la satisfacción del cliente y evitar que abandonen la organización, esta debería centrarse en la segmentación y satisfacción de las necesidades individuales de ellos (Tsai, Hu y Lu, 2015).

En general la segmentación del cliente es el proceso de dividirlos en diferentes grupos según información geográfica, demográfica y etnológica, para adoptar estrategias adaptadas a cada grupo basadas en el consumo de bienes y servicios y

compra de clientes (Tsai *et al.,* 2015). El proceso de segmentación continua potencia la armonía en el valor de los clientes; junto con la detección y correcta colocación de estos en grupos relacionados, y es uno de los factores más esenciales en CRM y el éxito empresarial (Kandeil *et al*., 2014; Tsai *et al*., 2015). Dado que no existe un enfoque preferido para la segmentación de clientes, el mejor modelo es aquel que dibuja un adecuado conocimiento de clientes actuales y potenciales, y ayuda a las organizaciones para lograr mercados efectivos y apropiados.

Como contribución a dicha problemática este artículo se planteó como objetivo: proponer una metodología multicriterio para la clasificación de clientes, considerando algoritmos KNN como sustento para la planeación agregada, a partir de una modificación al modelo RFM.

DESARROLLO

El modelo RFM

El modelo más utilizado en la segmentación del cliente es el denominado RFM, que consta de tres variables de comportamiento: R (actualidad), F (frecuencia) y M (monetario) (Kandeil *et al*., 2014). Su método general se expresa a continuación:

a) Se establece una escala de 5 elementos para cada indicador (desde 1 hasta 5) siendo 1 el mejor caso y 5 el peor. La combinación ideal de cliente será el que cumpla con la combinación 111 y el peor 555.

b) Se construye una tabla en la que se muestren los elementos clasificados en 1 y 2 los mejores, 4 y 5 los peores y 3 los promedios.

No obstante, el modelo RFM en la mayoría de los casos requiere de mejoras para adecuarse a condiciones específicas, por lo que múltiples estudios intentan mejorar y desarrollar el modelo al incluirle otras variables (Moghaddam, Abdolvand y Harandi, 2017). Este modelo intenta identificar a los clientes en función del comportamiento de sus características (Chen, Kuo, Wu y Tang, 2009; Noori, 2015;

Silva, Varela, Borrero y Rojas, 2019). Este comportamiento es modelado con el empleo de tres criterios de transacciones de clientes, que son entonces interpretados por herramientas y algoritmos de minería de datos.

Por lo tanto, el modelo RFM predice los próximos movimientos del cliente basados en su comportamiento (Asllani y Halstead, 2015). Además, no solo se considera como un modelo de segmentación, sino también incluye el concepto de valor del cliente. Es por eso que muchos estudios usan este modelo para el descubrimiento y análisis del valor del cliente basado en el comportamiento de compra pasado (Chen *et al*., 2009; Silva *et al*., 2019). Sobre la base del comportamiento de sus compras, el RFM revela que los clientes valiosos son aquellos que se manifiestan más recientemente, con la mayor frecuencia y valor monetario. A pesar de su simplicidad de comprensión, interpretación e implementación, el modelo tiene defectos, como la falta de atención a características personalizadas y variables demográficas de los clientes (Hosseini, Maleki, y Gholamian, 2010).

En este sentido, Chen *et al*. (2009) basaron su estudio en el modelo RFM aplicado a clientes segmentados de una empresa de servicios electrónicos en Taiwán. La clasificación se hizo con el propósito de agregar al modelo, la variable crédito de clientes. Después del agrupamiento de clientes con su algoritmo propuesto LEM2, extrajeron reglas importantes para la toma de decisiones de marketing y así potenciar las estrategias de la empresa. Por otra parte, Khajvand *et al.* (2011) en su estudio en la industria cosmética, han clasificado y valorado a clientes según la cantidad de productos comprados por estos, junto con otras variables de comportamiento. En este mismo contexto Noori (2015) agregó la variable del depósito del cliente al modelo RFM para clasificar dispositivos móviles de los usuarios bancarios. Indicó que identificar clientes por una puntuación de comportamiento facilita la asignación de estrategias de marketing.

Esos estudios emplearon el enfoque minorista (B2C); sin embargo, otros han realizado estudios en el campo mayorista (B2B). Una de las características más interesantes en el campo de B2B lo constituye que la empresa obtendrá grandes ganancias del gran volumen de compras al realizar ventas en la mayoría de los casos por lotes. Por lo tanto, entender a los clientes en este tipo de organizaciones sería la clave del éxito.

Otros como Hosseini *et al*. (2010) han agregado el ciclo de vida del producto/duración del producto como variables al modelo RFM y clasificaron los clientes en 34 categorías en una empresa B2B. Por su parte, Kandeil *et al.* (2014) emplearon técnicas de agrupamiento para segmentar clientes de un distribuidor B2B. Más recientemente Moghaddam *et al*. (2017) adicionaron la variable de variedad de productos como una nueva variable de comportamiento en el modelo RFM, y para clasificar a los clientes de la empresa de tipo B2B.

En este artículo se muestra una modificación al modelo RFM con el fin de incluir la variable tiempo de concurrencia para determinar qué clientes solicitan productos en el mismo horizonte de tiempo, cuestión fundamental para poder desagregar desde la planeación recursos de las organizaciones de tipo B2B.

Diferentes técnicas y formas del análisis de datos se pueden usar para la segmentación de clientes; aunque los más comunes son el uso de técnicas de minería de datos y variables de comportamiento del cliente. La minería de datos es el proceso de descubrir y extraer patrones ocultos de grandes cantidades de datos (Kord y Tavoli, 2015). Uno de sus algoritmos de clasificación más eficientes lo constituye el *K-Nearest Neighbor*.

Clasificador *K-Nearest Neighbor*

El algoritmo *K-Nearest Neighbor* (KNN) es uno de los métodos más simples para resolver problemas de clasificación; a menudo produce resultados competitivos y tiene ventajas significativas sobre otros métodos de minería de datos. Se ha demostrado que es capaz de superar algunos de los problemas asociados con otros algoritmos disponibles. Según Adeniyi, Wei y Yongquan (2016) entre sus características se encuentran:

- Superar el problema de escalabilidad común a muchos métodos de minería de datos existentes, como la técnica del árbol de decisión, a través de su capacidad para manejar datos de entrenamiento que son demasiado grandes para caber en la memoria.

- El uso de la distancia euclidiana simple para medir las similitudes entre las tuplas[1] de entrenamiento y las tuplas de prueba en ausencia de conocimiento previo sobre la distribución de datos, por lo tanto, facilita su implementación.
- Reducción de la tasa de error causada por la inexactitud en los supuestos hechos para el uso de otra técnica, como la técnica de clasificación Bayesiana Naive, como la independencia condicional de clase y la falta de datos de probabilidad disponibles, que generalmente no es el caso cuando se utiliza el método KNN.
- Proporcionar una recomendación más rápida y precisa al cliente con cualidades deseables, como resultado de la aplicación directa de similitud o distancia para fines de clasificación.

El algoritmo KNN se describió por primera vez a principios de la década de 1950. KNN es aplicable en muchos campos como el patrón reconocimiento, extracción de texto, finanzas, agricultura y medicina, entre otros (Imandoust y Bolandraftar, 2013). Es un algoritmo no paramétrico. No requiere ningún conocimiento previo sobre el conjunto de datos y supone que las instancias en los conjuntos de datos están distribuidas de forma independiente e idéntica, por lo tanto, las instancias cercanas entre sí tienen la misma clasificación (Pooja, 2017).

KNN también se conoce como estudiante perezoso porque en su fase de aprendizaje simplemente almacena todas las tuplas de entrenamiento que se le dan como entrada sin realizar ningún cálculo o solo hace un poco de procesamiento y espera hasta que se le dé una tupla de prueba para clasificar. Todos los cálculos o procesos se aplican al momento de la clasificación de una tupla de prueba. La forma más común es clasificar la tupla desconocida comparándola con tuplas de entrenamiento que son similares a ella. Cuando se le da una tupla desconocida, un clasificador de k-vecino más cercano busca en el espacio del patrón las t-vecinas más cercanas o las tuplas más cercanas a la tupla desconocida. La cercanía se define en términos de una distancia métrica (Han y Kamber, 2006). Para encontrar

[1] Una tupla es una secuencia de valores agrupados que sirve para agrupar, como si fueran un único valor, varios valores que, por su naturaleza, deben ir juntos.

las tuplas de entrenamiento K más cercanas a la tupla desconocida se usa como medida la distancia euclidiana. La distancia euclidiana estándar $d(x, y)$ es a menudo utilizada como la función de distancia.

A continuación, se proporciona la función de distancia euclidiana que calcula la distancia entre dos tuplas x e y.[2]

(I) $$d(x,y) = \sqrt{\sum_{i=1}^{n}\big(a_i(x) - a_i(y)\big)^2}$$

Donde *n* es el número de atributos totales y a es el valor del atributo en las instancias x (tupla de prueba) e y. Para la clasificación en este algoritmo se puede ponderar la contribución de cada vecino de acuerdo a la distancia entre el vecino y el ejemplar a ser clasificado (en nuestro caso x) y dando mayor peso a los vecinos más cercanos. Este cálculo se muestra en la siguiente ecuación.

(II) $$c(x) = argmax \sum_{c\varepsilon C=1}^{k} \delta\big(c, c(y_i)\big)$$

Donde $y_i = y_1,\ y_2 \ldots y_k$ son los k vecinos más cercanos de x, k es el número de vecinos totales, de los cuales se busca los mejores valores ($argmax$), y en concordancia con que los valores se encuentren en vecindad como se expresa por,$\delta\big(c, c(y_i)\big)$ 1 si $c\ c(y_i)$ y $\delta\big(c, c(y_i)\big)$ 0 de lo contrario.

La elección de K es muy importante en la construcción del modelo KNN. Es uno de los factores más importantes de este modelo que puede influir fuertemente en la calidad de las predicciones. Para cualquier problema, un pequeño valor de K conducirá a una gran variación en las predicciones. Alternativamente, establecer K en un valor grande puede conducir a un sesgo de modelo grande. Por lo tanto, K debe establecerse en un valor lo suficientemente grande como para minimizar la probabilidad de clasificación errónea y lo suficientemente pequeña (con respecto al número de tuplas en el conjunto de datos) para que el punto K más cercano esté lo suficientemente cerca del punto de consulta. Cuando K = 1, a la tupla desconocida se le asigna la clase de la tupla de entrenamiento que está más cerca de él en el

[2] La distancia euclidiana o euclídea, es la distancia ordinaria entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras y se calcula en un espacio bidimensional, como la distancia entre dos puntos P_1 y P_2, de coordenadas cartesianas (x_1, y_1).

espacio del patrón (Han y Kamber, 2006). El valor de K suele ser números impares, es decir, K = 1, K = 3, K = 5, esto es concebido para evitar los lazos[3] (Adeniyi *et al*., 2016).

Su empleo en la clasificación es ampliamente reconocido como efectivo y como se muestra en Geler, Kurbalija, Ivanović y Radovanović (2020) puede emplearse con éxito en la clasificación de series temporales. Por otra parte, en la regresión logística y combinado con el algoritmo *Random Forest* se evidenció su empleo exitoso en Shah, Patel, Sanghvi y Shah (2020). Igualmente se utiliza con éxito como método de detención de anomalías como demostraron (Sarmadi y Karamodin, 2020). Más recientemente siguen aportes que atestiguan su vigencia y su evolución; de los que se consideran como relevantes pueden citarse a Jasmir, Nurmaini y Tutuko, 2021; Harinadha (2022) y Yikun, Zhibin, Dong (2022).

El procesamiento de grandes cantidades de datos para respaldar los procesos de decisión, requiere, aparte de algoritmos de clasificación, de metodologías para su aplicación industrial. En tal sentido la más aceptada es la CRISP-DM (Huber, Wiemer, Schneider y Ihlenfeldt, 2019).

Metodología CRISP-DM

CRISP-DM es una metodología de proceso independiente de la industria para la minería de datos (Huber *et al.,* 2019), la cual, según Acosta, Janeth, Cuesta, Umaña y Coronado *(*2022) tiene gran aceptación para estudios de análisis de datos al no estar restringida a ninguna tecnología. Consta de seis fases iterativas de entendimiento del negocio hasta el despliegue (Tabla 1). CRISP-DM sigue un enfoque orientado a objetivos; se trata de un enfoque maduro que continúa siendo ampliamente aceptado en proyectos de minería de datos a través de algoritmos de aprendizaje automático de datos (Schröer, Kruse y Marx, 2021).

Entre sus reconocidas ventajas según (Huber *et al*., 2019) se encuentran:

- gran capacidad para el descubrimiento de conocimiento a través del procesamiento de bases de datos;

[3] Esto se refiere a la necesidad de comparación de los elementos de los conglomerados, que, si considera números pares, no tendría criterio preponderante para decidir.

- prevé la especialización de acuerdo con un contexto predefinido.

Tabla 1. Fases de la metodología CRISP-DM

Fase	Explicación
Comprensión del negocio	En ella se evalúa la situación del negocio para obtener una visión general de los recursos disponibles y necesarios.
Comprensión de datos	En esta fase se efectúa la recopilación de datos de las fuentes de datos, su exploración, descripción y la verificación de su calidad.
Preparación de datos	Se seleccionan los datos a utilizar atendiendo a criterios de inclusión y exclusión.
Modelado	Consiste en seleccionar la técnica de modelado, construir el caso de prueba y el modelo.
Evaluación	En esta fase los resultados se comparan con los objetivos comerciales definidos.
Implementación	Se despliegan los resultados obtenidos.

Fuente: Elaboración propia

Consecuentemente, en el presente artículo se empleó el algoritmo KNN para modelar y clasificar los comportamientos de compra de los clientes. Además, se utiliza la metodología CRISP-DM, que es uno de los mejores métodos analíticos para proyectos de minería de datos. El algoritmo KNN y la medida de la calidad de agrupación de siluetas se utiliza para medir y determinar el número de grupos. Esta operación se llevó a cabo en 739 registros normalizados de clientes con el uso del *software* SPSS.21. Luego se utiliza, el método de análisis jerárquico para ponderar las variables R, F, M, V, C, T y calcular el valor de cada cliente. Finalmente se ejecuta la prueba ANOVA en los grupos obtenidos, para determinar la significación de los datos conseguidos en ellos, utilizando el *software* SPSS.21.

RESULTADOS

Basados en las seis fases con que consta la metodología CRISP-DM: comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación y despliegue (Moro, Laureano y Cortez, 2011), se procede a mostrar el resultado de su aplicación en este estudio.

Fase 1. Comprensión empresarial

Descripción: en esta fase, se realiza la descripción general del tipo de negocio, en función del cual se realiza la investigación.

El estudio se ejecutó para la empresa ACINOX UEB Holguín comercializadora, la cual es de tipo B2B y su objeto es la comercialización de metales y derivados. Tiene como objetivo principal agrupar a los clientes en *clusters*, lo que contribuirá a poder tomar mejores decisiones sobre la planeación agregada que se requiere en dicha institución.

Con este fin los *clusters* se ordenarán según la ponderación jerárquica de las seis variables que componen el modelo RFMVCT, las cuales determinan agrupaciones sobre su fidelidad con la empresa, la frecuencia con que realiza sus compras, el valor patrimonial de los clientes, la variedad de productos que compran, la cercanía física a los almacenes de la empresa y el horizonte temporal en el cual concurren sus compras. Se hará especial énfasis en la variable, horizonte temporal en el cual realizan sus compras (T), la cual contribuye con mayor notoriedad a la agregación de los recursos que requiere una planeación agregada más eficiente, que permita satisfacer necesidades comunes por grupos de clientes.

Fases 2. Comprensión de datos

Descripción: la segunda fase consiste en recopilar, describir y evaluar la calidad de los datos.

En general, el objetivo de esta fase es seleccionar la fuente de datos apropiada para alcanzar la meta (Moghaddam *et al.,* 2017). Por lo tanto, se empleó la información disponible de la base de datos del sistema de gestión de la empresa ACINOX UEB Holguín. De esta se obtuvieron los datos de 53 clientes.

De estos datos se obtuvieron lo referente a las seis variables del modelo, las cuales se determinaron como se muestra en la Tabla 2. Y todo el conjunto de datos de salida se procesaron con el *software Microsoft Excel* versión 2019.

Tabla 2. Variables consideradas en la clasificación

Variable	Forma de medida
Fidelidad con la empresa (R)	Tiempo, expresado en meses que lleva siendo cliente.
Frecuencia con que realiza sus compras (F)	Cantidad de compras que realiza, expresada en unidades.
Valor patrimonial de los clientes (M)	Valor monetario que reportan sus compras, expresado en unidades.
Variedad de productos que compran (V)	Cantidad de productos que compra, expresado en unidades.
Cercanía física (C)	Distancia requerida para la transportación de la compra, expresado en kilómetros.
Horizonte temporal de compras. (T)	Momento en el año que realiza sus compras, expresado en cuartos*.

*La planeación agregada requiere de un mínimo de tres meses para su concreción(Krajewski, Malhotra y Ritzman, 2018); en tal sentido se divide el año en cuatro trimestres y se asigna numéricamente el rango correspondiente al equivalente de una unidad (0,25; 0,5; 0,7; 1).

Fuente: Elaboración de los autores

Fases 3. Preparación de datos

Descripción: en esta fase se preparan los datos para el modelado.

La preparación de los datos incluye el proceso de exclusión de valores atípicos y normalización de datos. En este estudio se recopilaron 927 registros de los cuales, después de eliminar duplicados y datos incompletos, se redujeron a 739 registros de datos. Se utilizó el *software* SPSS.21 para identificar valores atípicos. Posteriormente, basados en el método MIN-MAX y con el uso de las fórmulas (1), (2), (3), (4), (5) y (6) se normalizan los índices en el rango de cero y uno para prepararse para el modelado (siguiente fase). Con este fin, estas fórmulas,$MAX_R, MAX_M, MAX_F, MAX_V, MAX_C$ y MAX_Trepresentan el valor más alto de cada variable; y $MIN_R, MIN_M, MIN_F, MIN_V, MIN_C$ y MIN_T representa el valor más bajo en el conjunto de datos, que posteriormente se normalizan a los valores finales de R, M, F, V, C y T.

$$R = \frac{(R - MIN_R)}{(MAX_R - MIN_R)} \quad (1)$$

$$M = \frac{(M - MIN_M)}{(MAX_M - MIN_M)} \quad (2)$$

$$F = \frac{(F - MIN_F)}{(MAX_F - MIN_F)} \quad (3)$$

$$V = \frac{(V - MIN_V)}{(MAX_V - MIN_V)} \quad (4)$$

$$C = \frac{(C - MIN_C)}{(MAX_C - MIN_C)} \quad (5)$$

$$T = \frac{(T - MIN_T)}{(MAX_T - MIN_T)} \quad (6)$$

Fase 4. Modelado

Descripción: en esta fase se conforma y aplica el modelo.

Basado en el estudio de Moghaddam *et al.* (2017) —el cual desarrolla una solución similar a la requerida en este problema—, se determina el uso de tres *clusters* para resolver el problema de clasificación.

Valoración y determinación del modelo RFMVCT

Cada variable se pondera con el empleo de un análisis jerárquico el cual debe determinarse en orden para calificar y agrupar clientes. En este estudio se emplea un método de análisis jerárquico sustentado en Saaty (2001), para determinar los

pesos de las variables R, F, M, V, C y T. Para este fin se pidió a los encuestados que hicieran parejas de comparación, y les asignaran un valor de 1 a 9 a cada índice.

Se consultaron a los especialistas comerciales de la empresa objeto de estudio y a su director. De acuerdo con la fórmula (7), se considera el valor total de 1 y un índice de inconsistencia de 0,1. El peso total de las variables se muestra en la Tabla 3.

$$1 = wR + wF + wM + wV + wC + wT \quad (7)$$

Entonces, el valor de cada cliente (VC) es calculado en base a la fórmula (8).

$$VC_i = R_i * wR + F_i * wF + M_i * wM + V_i * wV + C_i * wC + T_i * wT \quad (8)$$

Tabla 3. Peso resultante de cada variable

Variable	Peso (%)
R	4,6
F	5,0
M	43,6
V	5,6
C	25,3
T	15,9

Fuente: Elaboración de los autores

En lo adelante se muestra todo el procedimiento para el conglomerado de la variable VC que es el resultado de la suma ponderada de todas las variables del modelo RFMVCT (Tabla 4). No obstante, el mismo procedimiento se desglosa para cada variable individual y se muestra en los resultados de la Fase 6.

Tabla 4. Centros iniciales de los conglomerados

	Conglomerado		
	1	2	3
Puntuación R	1,706 55	-1,836 05	3,035 02
Puntuación F	3,064 54	-1,224 25	1,127 67
Puntuación M	-,3079 1	-,502 81	3,650 11
Puntuación V	1,647 54	-1,298 83	1,647 54
Puntuación C	1,877 15	1,433 32	-,792 97
Puntuación T	-,333 81	1,435 38	-,333 81
Puntuación VC	-,306 69	-,501 97	3,650 06

Fuente: Elaboración de los autores

Después de determinar los pesos de las variables y el valor de todos los clientes, se procede a determina, de acuerdo con el valor medio de cada *cluster*, los rangos de clientes. Cada índice es determinado a partir de los datos extraídos en la primera fase.

A continuación se corrigen los centros de cada *cluster*. Esto se obtiene después de cinco iteraciones como se muestra en la Tabla 5 y de ahí, los conglomerados finales como aparece en la Tabla 6.

Tabla 5. Historial de iteraciones

Iteración	**Cambio en los centros de los conglomerados**		
	1	2	3
1	3,193	2,654	2,921
2	,295	,096	,000
3	,261	,104	,000
4	,112	,045	,000
5	,000	,000	,000

Fuente: Elaboración de los autores

Tabla 6 Centros de los conglomerados finales

Iteración	**Conglomerado**		
	1	2	3
Puntuación (R)	,387 30	-,393 52	1,042 31
Puntuación (F)	,897 09	-,553 47	,937 44
Puntuación (M)	-,217 71	-,448 66	2,177 28
Puntuación (V)	,509 17	-,543 97	1,480 13
Puntuación (C)	-,360 48	,230 91	-,411 78
Puntuación (T)	-,186 38	,121 89	-,223 24
Puntuación (VC)	-,217 96	-,44 857	2,177 29

Fuente: Elaboración de los autores

Una vez determinados los *clusters*, se procede a clasificar los clientes dentro de cada clúster, como se muestra en la Tabla 7, donde se visualiza el resultado de los conglomerados de la variable VC.

Tabla 7 Pertenencia a los conglomerados de la variable VC

Cliente	Conglomerado VC	Distancia		Cliente	Conglomerado VC	Distancia
1	1	1,633		28	3	1,375
2	1	1,631		29	2	1,938
3	1	1,968		30	1	1,686
4	2	1,955		31	2	,913
5	3	2,921		32	2	1,678
6	2	1,678		33	2	2,505
7	2	1,403		34	2	1,143
8	1	2,075		35	2	2,179
9	1	2,103		36	1	1,233
10	2	1,632		37	3	2,749
11	2	1,698		38	3	2,329
12	2	1,385		39	3	2,796
13	1	3,575		40	2	1,287
14	3	2,822		41	2	1,236
15	1	1,776		42	2	1,287
16	2	3,083		43	2	1,181
17	2	1,117		44	2	1,811
18	2	1,979		45	2	1,460
19	2	1,851		46	2	1,517
20	2	1,133		47	1	,613
21	2	1,837		48	3	1,460
22	2	3,791		49	2	1,296
23	1	1,335		50	2	1,523
24	2	2,281		51	3	2,655
25	2	1,673		52	2	1,170
26	2	1,437		53	2	1,807
27	1	,785				

Fuente: Elaboración de los autores

Fase 5. Evaluación

Descripción: en esta fase se prueba estadísticamente el modelo realizado y se muestran los resultados de su aplicación en el conjunto de datos.

Después de agrupar y analizar los resultados, la diferenciación de grupos creados por el algoritmo KNN y su resolución, se evalúa a través de la prueba ANOVA. Como se indica en la Tabla 8, los niveles de significación (sig) de todos variables: R (sig) = 0,000; F (sig) = 0,000; M (sig) = 0,000; V (sig) = 0,000; C (sig) = 0,095 y T (sig) = 0,030 son inferiores a 0,05, por lo tanto, la media de homogeneidad de las

poblaciones se rechaza, y se muestra que los grupos tienen diferencias significativas.

Tabla 8 ANOVA

	Conglomerado		**Error**		**F**	**Sig.**
	Media cuadrática	gl	Media cuadrática	gl		
Puntuación (R)	7,801	2	,728	50	10,716	,000
Puntuación (F)	13,398	2	,504	50	26,580	,000
Puntuación (M)	22,568	2	,137	50	164,395	,000
Puntuación (V)	15,201	2	,432	50	35,192	,000
Puntuación (C)	2,338	2	,946	50	2,470	,095
Puntuación (T)	6,653	2	,114	50	6,644	,030
Puntuación (VC)	22,567	2	,137	50	164,362	,000

Fuente: Elaboración propia

Fase 6. Despliegue

Descripción: en esta fase los resultados son analizados en la empresa objeto de estudio.

Después de revisar los resultados y la valoración, el modelo es juzgado. Si los resultados son consistentes con el objetivo primario de los negocios, y parecen satisfacer las necesidades del negocio, entonces se utilizarán en un entorno real. Una nueva fase se definirá de otra manera y el proceso se repetirá. En la Tabla 9 se muestra un ejemplo con 10 clientes de la empresa del informe final, el cual potencia el desarrollo de la planeación agregada en la empresa objeto de estudio.

Tabla 9 Resultado de la aplicación en la empresa objeto del estudio

Clúster de clasificación general (VC)	**ID_Clientes**	***Cluster* de la variable**					
		R	F	M	V	C	T
1	1	1	1	1	1	1	1
	2	1	2	1	1	2	4
	3	1	1	2	3	1	1
	8	2	1	1	1	2	1
	9	1	2	1	1	3	2
2	4	2	2	3	2	2	4
	6	2	1	2	2	2	3
	7	3	2	2	3	1	1

	10	1	3	3	2	2	3
3	5	2	3	3	1	3	4

Fuente: Elaboración de los autores

CONCLUSIONES

Los resultados de los conglomerados obtenidos de la variable VC dan una panorámica de los grupos de clientes con características similares, los cuales se ponderan en un conjunto de seis variables que tributan a VC. De esta forma quedan determinados los mejores clientes de la empresa.

Los conglomerados individuales de cada variable permiten tributar a la planeación agregada y así poder tomar decisiones en conjunto con distintas empresas y optimizar el proceso de venta para cada una de ellas desde una visión general.

Para la planeación agregada de la institución, se puede afirmar que:

Los conglomerados de clientes resultado de la variable C, permiten compartir recursos, específicamente transporte, ya que estos están agrupados según la distancia a los almacenes de la institución objeto de estudio.

Los conglomerados de clientes, resultado de la variable T, permiten compartir recursos (espacio de almacenamiento, transporte, entre otros) para cubrir sus demandas, ya que estas se realizan en un mismo horizonte temporal. En tal sentido la planeación agregada para los clientes 1, 3, 7 y 8 considera el primer trimestre del año; los clientes 2, 4 y 5 el cuarto; los clientes 6 y 10 el tercero; y el cliente 9 el segundo trimestre.

REFERENCIAS

Acosta, J., Janeth, D., Cuesta, L., Umaña, S. y Coronado, J. (2022). Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test. *Procedia Computer Science*, 198, 512-517. doi: 10.1016/j.procs.2021.12.278

Adeniyi, D. A., Wei, Z. y Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification

method. *Applied Computing and Informatics*, 12, 90-108. doi: 10.1016/j.aci.2014.10.001

Asllani, A. y Halstead, D. (2015). A Multi-Objective Optimization Approach Using the RFM Model in Direct Marketing. *Academy of Marketing Studies Journal*, *19*(3), 49-60. Recuperado de https://www.semanticscholar.org/paper/A-Multi-Objective-Optimization-Approach-Using-the-Asllani-Halstead/4c0253aa14d4e561b821c87d519549eeaf454567

Chen, Y. L., Kuo, M. H., Wu, S. Y. y Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data *Electronic Commerce Research and Applications*, *8*(5), 241-251. Recuperado de https://doi.org/10.1016/j.elerap.2009.03.002

Geler, Z., Kurbalija, V., Ivanović, M. y Radovanović, M. (2020). Weighted kNN and constrained elastic distances for time-series classification. *Expert Systems with Applications*, 162. doi: 10.1016/j.eswa.2020.113829

Han, J. y Kamber, M. (2006). *Data mining: concepts and techniques* (2nd. ed.). San Francisco, Estados Unidos: Elsevier.

Harinadha, K. (2022). Soft computing fuzzy set through knn−ML to identify islanding state of integrated electrical grid at different operational events. *International Journal of Electrical Power & Energy Systems*, 136, 107615. doi: 10.1016/j.ijepes.2021.107615

Hosseini, S. M., Maleki, A. y Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259-5264. Recuperado de https://www.sciencedirect.com/science/article/abs/pii/S0957417409011166

Huber, S., Wiemer, H., Schneider, D. y Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*, 79, 403-408. doi: 10.1016/j.procir.2019.02.106

Imandoust, S. B. y Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal for Research in Applied Science*, *3*(5), 605-610. Recuperado de https://www.academia.edu/4607757/Application_of_K-Nearest_Neighbor_kNN_Approach_for_Predicting_Economic_Events_Theoretical_Background

Jasmir, J., Nurmaini, S. y Tutuko, B. (2021). Fine-Grained Algorithm for Improving KNN Computational Performance on Clinical Trials Text Classification. *Big Data and Cognitive Computing*, *5*(4), 60. doi: 10.3390/bdcc5040060.

Kandeil, D. A., Saad, A. A. y Youssef, S. M. (2014). A Two-Phase Clustering Analysis for B2B Customer Segmentation. Ponencia presentada en *Conferencia Internacional 2014 sobre Redes Inteligentes y Sistemas Colaborativos*. Salerno, Italia.

Khajvand, M., Zolfaghar, K., Ashoori, S., y Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*(1), 57-63. Recuperado de https://www.sciencedirect.com/science/article/pii/S1877050910003868

Kord, P. T. y Tavoli, R. (2015). A Review of Different Data Mining Techniques in Customer Segmentation. *Journal of Advances in Computer Research, 6*(3), 51-63. Recuperado de http://jacr.iausari.ac.ir/article_643199.html

Krajewski, L., Malhotra, M. y Ritzman, L. (2018). *Operations Management Processes and Supply Chains* (12ma. ed.). New York, United States of America: Prentice Hall.

Miguéis, V. L., Poel, D. V. den, Camanho, A. S., y Cunha, J. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, *39*(12), 11250-11256. doi:10.1016/j.eswa.2012.03.073

Moghaddam, S. Q., Abdolvand, N. y Harandi, S. R. (2017). A RFMV Model and Customer Segmentation Based on Variety of Products. *Journal of Information Systems and Telecommunication*, 5(3), 155-161. Recuperado de https://www.academia.edu/38332240/A_RFMV_Model_and_Customer_Segmentation_Based_on_Variety_of_Products

Moro, S., Laureano, R. y Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An Application of the Crisp-Dm Methodology. *2011 Proceedings of European Simulation and Modelling Conference-ESM*, 117-121.

Noori, B. (2015). An Analysis of Mobile Banking User Behavior Using Customer Segmentation. *International Journal of Global Business*, *8*(2), 55-64. Recuperado de http://www.proquest.com/docview/1776310680

Pooja, R. (2017). A Review of various KNN Techniques. *International Journal for Research in Applied Science*, *5*(8), 1174-1179.

Saaty, T. (2001). *The seven pillars of the analytic hierarchy process*. USA: University of Pittsburgh.

Sarmadi, H. y Karamodin, A. (2020). A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects. *Mechanical Systems and Signal Processing*, 140. doi: 10.1016/j.ymssp.2019.106495.

Schröer, C., Kruse, F. y Marx, J. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, 526-534. doi: 10.1016/j.procs.2021.01.199.

Shah, K., Patel, H., Sanghvi, D. y Shah M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augment Hum Res*, *5*(12). doi: 10.1007/s41133-020-00032-0.

Silva, J., Varela, N., Borrero, L. A. y Rojas, R. H. (2019). Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm. *Procedia Computer Science*, 151, 1207-1212. Recuperado de https://www.sciencedirect.com/journal/procedia-computer-science

Tsai, C. F., Hu, Y. H. y Lu, Y. H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, *32*(1), 65-76. Recuperado de https://www.semanticscholar.org/paper/Customer-segmentation-issues-and-strategies-for-an-Tsai-Hu/11ccdd200271110c98af8756fcab856b74398e3b

Yikun, W., Zhibin, P. y Dong, J. (2022). A new two-layer nearest neighbor selection method for kNN classifier. *Knowledge-Based Systems*, 235, 0950-7051, doi: 10.1016/j.knosys.2021.107604.Z.

Declaración de conflicto de interés y conflictos éticos

Los autores declaramos que el presente manuscrito es original y no ha sido enviado a otra revista. Los autores somos responsables del contenido recogido en el artículo, y en él no existen: ni plagios, ni conflictos de interés, ni éticos.

Declaración de contribuciones de los autores

Carlos Jesús Madariaga Fernández. Gestión de la información para favorecer la actualización del artículo. Concepción preliminar (idea generadora) y diseño del artículo. Fundamentos teóricos, proyecto y desarrollo de la metodología. Análisis de los resultados. Elaboración de conclusiones.

Yosvani Orlando Lao León. Revisión teórica general del artículo. Perfeccionamiento del resumen y conclusiones. Profundización de los resultados y sus inferencias logísticas. Revisión de las referencias bibliográficas.

Dagnier Antonio Curra Sosa. Procesamiento de datos, redacción de las metainferencias.

Rafael Lorenzo Martín. Revisión técnica y uso de términos (tesauro especializado) del artículo. Coherencia y lógica investigativa, redacción y revisión de la estructura y relaciones del artículo. Gestión del colchón editorial a publicar.