Mapas Auto-Organizados aplicados a la segmentación de clientes en entornos empresariales

Roberth Figueroa-Diaz[1], Sairy Chamba[1], Rene Guaman-Quinche[1], Mario Cueva-Hurtado[1]

[1] Carrera de Ingeniería en Sistemas, Universidad Nacional de Loja

roberth.figueroa@unl.edu.ec, sfchambaj@unl.edu.ec, rguaman@unl.edu.ec, mecueva@unl.edu.ec

Resumen. Este artículo se ha enfocado en la aplicación de técnicas de Minería de Datos para segmentación de clientes, sobre datos reales de una empresa dedicada a la comercialización de productos tecnológicos de la región sur del Ecuador. Se aplicó la metodología CRISP-DM para el proceso de Minería de Datos y en base al modelo Recencia, Frecuencia, Monto (RFM), se aplicaron algoritmos de clustering: k-means y Mapas Auto-Organizados (SOM en inglés Self-Organizing Map). Para validar el resultado de los algoritmos de clustering y seleccionar el que proporcione grupos de mejor calidad, se usó la técnica de evaluación en cascada, para ello se aplicó un algoritmo de clasificación, colocando como etiqueta de clase a los grupos encontrados por los algoritmos de clustering y luego se midió la precisión de predicción con respecto a estos grupos. El algoritmo SOM fue el que proporcionó los mejores resultados. El proceso mencionado se llevó a cabo utilizando el lenguaje R a través de R-Studio.

Palabras Claves: Clustering, K-means, RFM, CRIS-DM, Segmentación de Clientes, SOM, lenguaje R, Minería de datos.

1 Introducción

Las empresas generan y almacenan diariamente gran cantidad de información [1], pero los datos tal cual se almacenan no suelen proporcionar beneficios directos, su valor real reside en la información que podemos extraer de ellos, es decir información que ayude a tomar decisiones o a mejorar la comprensión de los fenómenos que la rodean [2] [3]. En este contexto es que la Minería de Datos permite extraer información sensible que reside de manera implícita en los datos[4] [5].La segmentación de clientes se utiliza como una herramienta de diferenciación de marketing, que permite a las organizaciones entender a sus clientes y construir estrategias diferenciadas [6].

El modelo RFM [7] basado en la recencia de compra, la cual hace referencia al tiempo transcurrido desde la última compra realizada por el cliente, la frecuencia de compra y el valor monetario gastado en compras, se ha utilizado durante años para el marketing directo y ha demostrado ser muy efectivo al usarse sobre las bases de datos transaccionales [8]. Las variables RFM se encuentran directamente relacionadas a la compra e influyen en las posibilidades futuras de compra de los clientes, por lo tanto

permite determinar el nivel de lealtad de los clientes [9].

En el presente trabajo se aplican técnicas de minería de datos para segmentar clientes en base al modelo RFM, específicamente se hace uso de algoritmos de *clustering* o agrupación. La técnica de *clustering* consiste en agrupar un conjunto de objetos físicos o abstractos, en donde los objetos pertenecientes a un grupo son similares entre sí y son diferentes con respecto al resto de grupos [10] [11]. Los algoritmos que se utilizan en este trabajo son k-means y SOM.

El algoritmo K-means es uno de los algoritmos de *clustering* más conocidos, se basa en el agrupamiento particional y consiste en asignar cada elemento al grupo con el centroide más cercano, el cual está representado por el valor de la media de los puntos de un grupo [12] [13], la función objetivo de k-means es minimizar la suma del error cuadrático, además este algoritmo requiere conocer con anticipación el número de grupos a formar y es muy sensible a la selección de los centroides iniciales y al ruido[14].

Los Mapas auto-organizados o SOM, son un modelo de red neuronal no supervisado que convirtiendo un espacio de entrada de altas dimensiones en un mapa más pequeño, normalmente dos dimensiones, preservando las propiedades topológicas de los vectores de entrada por medio de un concepto de vecindad, con disposición de las neuronas en rejillas de forma rectangular, triangular o hexagonal [15] [16]. Su función esencial es, descubrir la estructura subyacente de los datos introducidos en el mapa, durante el entrenamiento la neurona que presente menor diferencia entre el vector de peso y el vector de datos será la neurona ganadora y sus vecinas verán modificadas sus vectores de peso.

El presente artículo se encuentra organizado de la siguiente manera: En la sección II se describe el método aplicado para llevar a cabo el proceso de segmentación de clientes, en la sección III se presenta el análisis de los resultados obtenidos y en la sección IV se presentan las conclusiones obtenidas.

2 Metodología

El proceso de minería de datos se llevó a cabo usando la metodología CRISP-DM del inglés *Cross Industry Standard Process for Data Mining* [17], la cual contiene una serie de fases que se siguieron para obtener el resultado final.

2.1 Comprensión del Negocio

La empresa en la que se llevó a cabo el análisis es una empresa dedicada a la comercialización de productos y servicios tecnológicos de la región sur del Ecuador, los departamentos involucrados en el análisis realizado fueron Marketing y Ventas. En las áreas de Marketing y Ventas de la empresa en estudio, se presenta un inconveniente a la hora de elaborar estrategias de retención de clientes, la empresa es consciente de que

posee distintos tipos de clientes pero al momento no puede identificarlos para llegar a ellos de forma efectiva. A pesar de que posee una gran cantidad de datos acerca de sus clientes, resulta difícil manejar todos esos datos sin las técnicas, herramientas y el procedimiento adecuado. La necesidad que posee la empresa en este ámbito es de establecer grupos de clientes, que le permita identificar la lealtad de cada uno de ellos.

Para dar solución a este problema, debe realizarse un proceso de creación de grupos de clientes en base a su comportamiento de compra y la identificación del nivel de lealtad de los clientes hacia la empresa [18].

2.2 Comprensión de los Datos

Esta etapa abarcó varias actividades: recopilación, descripción, exploración y verificación de la calidad de los datos. Los datos proporcionados corresponden a la información transaccional de la empresa, desde el año 2010 hasta el año 2014. Se cuenta con datos de: clientes, tipo de cliente, institución, factura, detalle de factura, productos, grupo de productos y marcas de productos. Para el análisis RFM serán necesarios los datos de Clientes y Facturas, el tamaño inicial de este conjunto de datos se muestra en la Tabla 1.

Tabla 1. Tamaño Inicial del Conjunto de Datos.

Datos	Número de registros
Clientes	44800
Transacciones	136278

La base de datos posee una información transaccional muy amplia, con limitados datos sociodemográficos de los clientes de la empresa.

2.3 Preparación de los datos

En esta etapa se seleccionó el conjunto de datos sobre el cual se trabajó para el análisis, tomando en cuenta que la segmentación se realizará en base a las variables RFM, se seleccionaron los datos de la tabla factura y se procedió a eliminar las facturas anuladas, repetidas y aquellas que han sido creadas para pruebas. Asimismo, de la tabla clientes se seleccionó a los clientes finales y se trabajó únicamente con los atributos: id, código, nombre, ciudad e institución, estos dos últimos se utilizarán en la etapa de comprobación de los grupos creados.

Luego de la selección y limpieza de datos quedó el conjunto de datos final, cuyo tamaño se describe en la Tabla 2.

Tabla 2. Tamaño Final del Conjunto de Datos.

Datos	**Número de registros**
Clientes	30647
Transacciones	77218

Para realizar el análisis RFM, se construyeron los atributos Recencia, Frecuencia y Monto, tomando en cuenta las experiencias realizadas en [19] [12] [20]. La información recolectada durante el periodo 2010 al 2014 fue la base para la generación de los atributos. La Recencia se construyó mediante la diferencia entre la fecha actual y la fecha de la última compra realizada por cada cliente. Para construir el atributo Frecuencia se contabilizó el número de transacciones que cada cliente ha realizado y el Monto se construyó calculando el total de dinero gastado por cada cliente en todas sus compras.

Luego se definieron las escalas para los atributos RFM, las cuales se describen en la Tabla 3., para ello, se utilizó el método conocido como *hard-coding* [21], el cual sugiere elegir los intervalos y el peso de las variables tomando en cuenta el análisis de los datos almacenados, el conocimiento y experiencia de las personas dentro de la empresa, de acuerdo a esto, las variables RFM se evaluarán bajo el mismo peso.

Tabla 3. Escala de los atributos para los datos de la empresa.

Escala	**Nombre**	**Recencia (días)**	**Frecuencia**	**Monto**
5	MUY ALTO	[0-193]	[7,+]	[500,+]
4	ALTO	[194-442]	[5-6]	[92-500]
3	MEDIO	[443-823]	[3,4]	[33- 92]
2	BAJO	[824-1278]	[2]	[16- 33]
1	MUY BAJO	[1279,+]	[1]	[0,16]

2.4 Modelado

Para crear los grupos de clientes se aplicaron los algoritmos de clustering: k-means y SOM. Primero, se aplicó el algoritmo k-means, para obtener la agrupación sobre los atributos RFM. Este algoritmo requiere conocer el número de grupos con anticipación y por lo tanto, para determinar este parámetro se aplicaron dos métodos de evaluación interna [22]: curva de distorsión [23] e índice de la silueta [24].

En la Fig. 1, se muestra el resultado de la curva de distorsión, el eje X representa el número de clúster y el eje Y representa la suma de cuadrados para los clústeres (within-cluster sum of squares(WCSS)) según [25]. La solución del clúster apropiado se define en el momento en que ocurre una reducción dramática de la suma de cuadrados en clúster. Esto produce un "codo" en la trama y en este caso puede observarse este codo en el número de 5 clústeres.

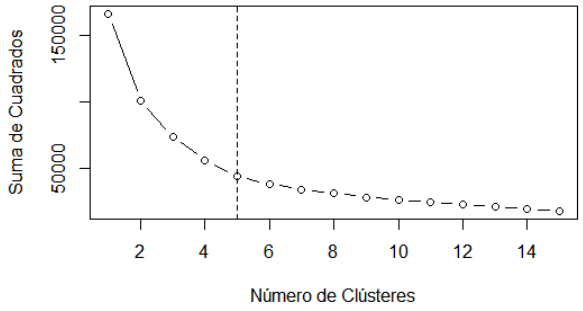

Fig. 1. Curva de distorsión

Para validar la interpretación de este gráfico, también se obtuvo el número de grupos mediante el índice de la silueta, obteniéndose el índice más alto igualmente en 5 grupos. Considerando que k-means es bastante sensible a la inicialización de los centroides iniciales, el valor de la semilla inicial es otro parámetro que se tomó en cuenta y después de probar con algunos valores, se estableció en 25 semillas iniciales, es decir 25 pruebas con distintos centros de clústeres iniciales, de las que se escoge la que entrega el mejor resultado. Con los parámetros establecidos, se procedió a crear grupos de clientes en base a los atributos RFM normalizados, aplicando el algoritmo k-means y los resultados obtenidos se muestran en la Tabla 4. Para asignar la etiqueta de lealtad de cada grupo, se calculó la distancia al punto cero con los centroides de cada grupo, un mayor valor significa mayor lealtad, mientras que un menor valor indica menor lealtad [26].

Luego de haber encontrado los grupos mediante k-means, se procedió a crear grupos sobre los mismos atributos RFM, pero esta vez aplicando el algoritmo SOM. Primeramente necesitamos conocer, el tamaño del mapa, para ello se utilizó la fórmula $5x\sqrt{N}$ [27], donde N es el número de muestras, en este caso N equivale a 30647 registros de clientes, dando un total de 900 neuronas, por lo tanto el tamaño del mapa se estableció en 30x30, lo que corresponde a un modelo de mapa de 2 dimensiones.

El número de iteraciones se estableció en 100, este parámetro se fijó mediante un proceso de prueba y error el cual se describe en [28] y al realizar varias pruebas se determinó que es un valor aceptable ya que como se observa en la siguiente figura a partir de la iteración 60 el promedio no presentar mayor variación para el experimento, considerando la estabilidad de los grupos y la representación gráfica del proceso de entrenamiento que se observa en la Fig. 2.

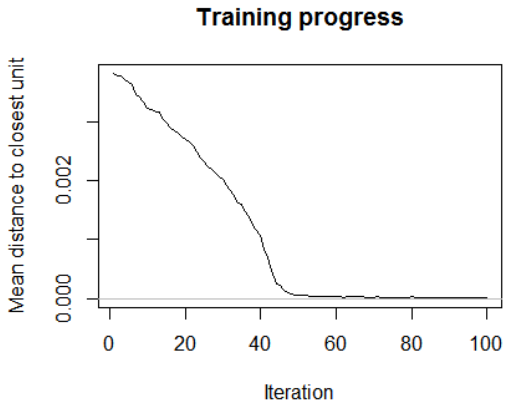

Fig. 2. Progreso de entrenamiento del Mapa auto-organizado (SOM)

Los valores de la tasa de aprendizaje se fijaron en 0.05 hasta 0.01. Con los parámetros establecidos, se aplicó el algoritmo SOM sobre el conjunto de datos y el resultado obtenido fueron los clientes agrupados en 107 nodos o neuronas, esto se interpreta como 107 grupos de clientes, esta representación puede verse en la Fig. 3.

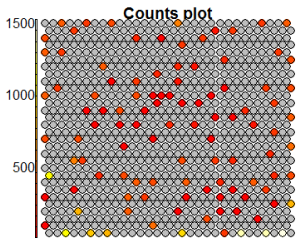

Fig. 3. Representación de la agrupación mediante SOM, en un mapa de dimensiones 30x30

Considerando que para la empresa sería complicado manejar 107 grupos de clientes, se aplicó un algoritmo jerárquico, usando el método de Ward descrito en [29], sobre la agrupación realizada por SOM, con la finalidad de reducir el número de grupos. De esta manera, los 107 nodos fueron agrupados en 5 grupos.

En la Fig. 4., se observan los cinco grupos creados sobre el mapa SOM, grupo 1 (azul), grupo 2 (naranja), grupo 3 (verde), grupo 4 (rojo) y grupo 5 (morado).

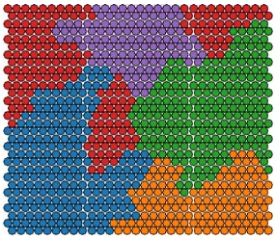

Fig. 4. División de grupos en el mapa (SOM)

Se obtuvieron los centros de cada grupo y se calculó la distancia al punto cero, al igual que se hizo para los resultados de k-means, en la Tabla 4 se muestra el resultado de los dos algoritmos aplicados: SOM y k-means.

Grupo		**1**	**2**	**3**	**4**	**5**
RECENCIA	SOM	3.99	1.41	3.96	2.22	4.51
	K-MEANS	3.84	1.42	4.14	2.13	4.53
FRECUENCIA	SOM	1.22	1.38	4.14	1.61	1.95
	K-MEANS	1.17	1.14	4.57	2.14	2.02
MONTO	SOM	1.77	1.95	4.04	4.11	4.46
	K-MEANS	1.67	1.83	4.45	4.12	4.03
DISTANCIA AL PUNTO CERO	SOM	4.53	2.78	7.01	4.94	6.64
	K-MEANS	4.35	2.59	7.61	5.07	6.40
LEALTAD	SOM	Bajo	Muy Bajo	Muy Alto	Medio	Alto
	K-MEANS	Bajo	Muy Bajo	Muy Alto	Medio	Alto
NÚMERO DE INSTANCIAS	SOM	7506	8141	4470	6512	4018
	K-MEANS	7065	7673	3086	7138	5685

Tabla 4. Resultados de 5 grupos creados por los algoritmos K.means y SOM.

3 Análisis de Resultados

3.1 Evaluación

Se han obtenido dos agrupaciones distintas, mediante k-means y mediante SOM, para seleccionar cuál de estos dos algoritmos ha proporcionado el mejor resultado, se hizo uso de la técnica de evaluación en cascada [30], que consiste en aplicar un algoritmo de clasificación sobre el conjunto de datos, para ello primeramente se debe establecer como etiqueta de clase el grupo creado por el algoritmo de segmentación y si es posible agregar para cada cliente, otros atributos además de las variables RFM [9], en este caso se agregaron los atributos: Ciudad e Institución.

Se dividió el conjunto de datos en 67% para entrenamiento y 33% para pruebas y se aplicó el algoritmo de clasificación supervisada LEM2 (Learning from Examples Module v2) [12]. Se siguió este procedimiento con los resultados obtenidos por k-means y SOM. Las reglas generadas sobre los resultados de k-means y SOM tienen el siguiente formato:

> *IF Recencia IS Medio and Monto IS Medio and Frecuencia IS Muy Bajo THEN IS Medio.*

Lo cual significa que, si un cliente tiene una Recencia de compra Media (443-823 días), un Monto en compras Medio (33- 92 dólares) y una Frecuencia de compra Muy Baja (1 compra), entonces pertenece al grupo de lealtad Medio. El total de reglas creadas sobre los resultados de k-means y SOM es de 67 y 74 respectivamente.

Una vez obtenidas las reglas de clasificación se midió la capacidad de éstas para predecir el grupo al que pertenecen los clientes que no fueron parte del entrenamiento, es decir el 33% que se dejó para pruebas. Este experimento se repitió 10 veces, con la finalidad de que en cada iteración se seleccionen casos aleatorios para los conjuntos de entrenamiento y pruebas, al final de estas 10 iteraciones se calcula un promedio del nivel de precisión.

En la Tabla 5, se muestra el resultado de la precisión de predicción para los grupos de k-means y SOM. Como se puede observar, la mayor precisión de predicción se presenta con las reglas generadas para predecir los grupos creados por SOM.

Tabla 5. Comparación de Resultados para los algoritmos K-means y SOM.

Experimento	Método	Nº Grupos	Precisión
1	K-MEANS	5	0.91629
	SOM	5	0.92845
2	K-MEANS	5	0.92401
	SOM	5	0.93602
3	K-MEANS	5	0.93684
	SOM	5	0.94277
4	K-MEANS	5	0.94151
	SOM	5	0.95007
5	K-MEANS	5	0.95273
	SOM	5	0.95860
6	K-MEANS	5	0.96032
	SOM	5	0.96402
7	K-MEANS	5	0.97398
	SOM	5	0.97102
8	K-MEANS	5	0.98150
	SOM	5	0.98561
9	K-MEANS	5	0.98905
	SOM	5	0.99013
10	K-MEANS	5	0.99979
	SOM	5	0.99998

3.2 Interpretación

Es importante interpretar los resultados de los grupos creados por el algoritmo seleccionado, en este caso SOM.

En la fig. 5, se muestra una gráfica de componentes principales de los grupos creados. Como se indica, el grupo Muy Alto posee una lealtad Muy Alta, cuyo número de instancias es de 4470, el grupo Alto posee una lealtad alta, cuyo número de instancias es de 4018, el grupo Medio posee una lealtad media, cuyo número de instancias es de 6512,

el grupo Bajo posee una lealtad baja, cuyo número de instancias es de 7506 y el grupo Muy Bajo posee una lealtad muy baja, cuyo número de instancias es de 8141.

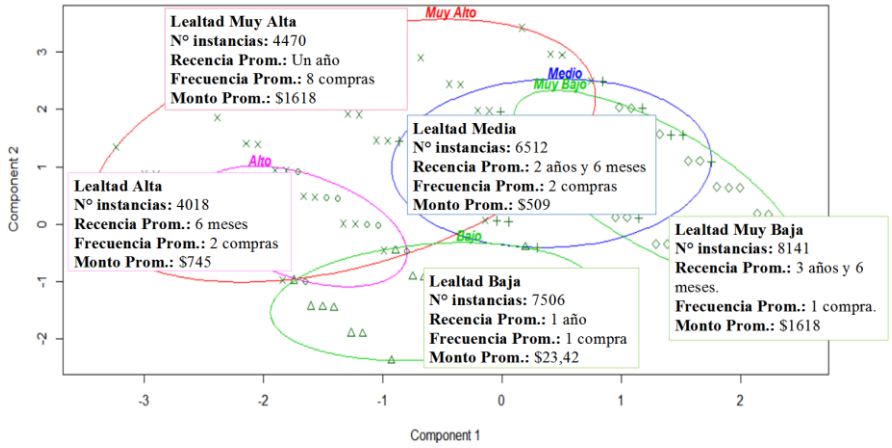

Fig. 5. Gráfica de los grupos creados por SOM

En la Fig. 6 se puede observar la distribución de las variables RFM para cada grupo de clientes.

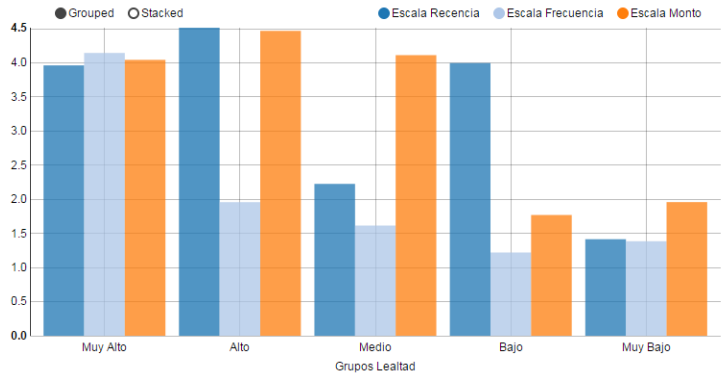

Fig. 6. Distribución de los atributos RFM para los grupos de lealtad de clientes creados por SOM

La distribución de las variables RFM ha sido interpretada para encontrar las características de los clientes dentro de cada grupo de lealtad, A continuación se describen las características RFM de cada grupo de clientes.

- Los clientes de lealtad **Muy Alta**, son clientes que han realizado su última compra en promedio hace 1 año atrás, además han realizado un promedio de 8 compras en un periodo de 5 años y han invertido un alto monto en sus compras, en promedio 1618 dólares. Estos son clientes muy valiosos para la empresa.
- Los clientes de lealtad **Alta**, son los clientes con la actividad de compra más reciente, en promedio 6 meses atrás, tienen una frecuencia de compra baja, 2 compras promedio en un periodo de 5 años y han invertido un Monto alto en sus compras, en promedio 745 dólares. Estos son clientes que gastan mucho dinero en sus compras, considerando que tienen una frecuencia de compra baja, podrían considerarse como clientes potencialmente valiosos.
- Los clientes de lealtad **Media**, son clientes que realizaron su última compra hace algún tiempo atrás, en promedio 2 años y seis meses, tienen un número aproximado de 2 compras, en un periodo de 5 años, también tienen un Monto alto en compras, 509 dólares en promedio. Son clientes que la empresa está en riesgo de perder.
- Los clientes de lealtad **Baja,** son clientes que han realizado su última compra hace aproximadamente 1 año atrás, pero el promedio de veces que han comprado es de 1 y el monto promedio gastado es de 23.42 dólares, que indica que han invertido poco dinero en sus compras. Los clientes de este grupo podrían ser considerados como los clientes nuevos.
- Los clientes de lealtad **Muy Baja,** han realizado su última compra hace mucho tiempo, en promedio 3 años y 6 meses atrás, también tienen una frecuencia promedio de una sola compra en un periodo de cinco años y un monto bajo que indica que han invertido poco dinero en sus compras, en promedio 33.75 dólares. Los clientes de este grupo se podrían considerarse como clientes que la empresa prácticamente ha perdido.

4 Conclusiones

Los algoritmos de segmentación aplicados: k-means y SOM, proporcionaron resultados bastante precisos al momento de tomar los grupos como clase para generar reglas y hacer predicciones. Esto significa que la estructura del agrupamiento realizado por k-means y SOM es bastante buena, pero como debemos elegir un modelo, se ha hecho una selección muy estricta de la precisión de predicción y en este caso el valor más alto, se obtuvo al predecir los grupos generados por el algoritmo SOM.

De acuerdo a los altos valores obtenidos en la evaluación realizada sobre los grupos, se pudo constatar que al medir la lealtad y comportamiento de los clientes en base al modelo RFM se obtienen resultados bastante confiables, ya que este modelo se basa en el análisis de transacciones reales.

Los grupos obtenidos mediante la aplicación de técnicas de Minería de Datos sobre las variables RFM de los clientes de la empresa en estudio, revelaron los niveles de lealtad: Muy Alto, Alto, Medio, Bajo y Muy Bajo, estos resultados le permitirán a la empresa elaborar estrategias de retención hacia sus clientes, en lugar de pagar un alto costo por la atracción de nuevos clientes.

El grupo más representativo de clientes corresponde al nivel de lealtad Muy Baja, abarca el 26.5% del total de clientes, es decir que este porcentaje de clientes han realizado en promedio una sola compra dentro de la empresa, hace aproximadamente 3 años y 6 meses en promedio, también su monto promedio de gasto en compras es de 33.75 dólares.

El grupo con el menor número de clientes corresponde al nivel de lealtad Alto, con un número de 4018 clientes que equivale al 13.11%, este porcentaje de clientes ha realizado un promedio de 2 compras en un periodo de cinco años, su última compra es en promedio hace seis meses y el monto promedio gastado en todas sus compras es de 745 dólares.

Dentro de los trabajos futuros se puede mencionar la evaluación de los datos preprocesados con técnicas de segmentación basadas en algoritmos probabilístico como Expectación-Maximización (EM), algoritmos de clustering jerárquico como Cobweb y la utilización de diversas técnicas de clasificación basadas en redes neuronales como Perceptron Multicapa (MLP) o árboles de decisión.

Así mismo como aplicación de apoyo a la toma de decisiones se puede integrar los resultados obtenidos para evaluación y prueba sobre la segmentación de clientes en tiempo real, mismo que podrá ir mejorando con la incorporación de nuevos algoritmos de clustering.

Referencias

[1] L. C. Molina, “Data mining: torturando a los datos hasta que confiesen,” *Fuoc*, pp. 1–11, 2002.

[2] S. P. Jiménez, J. J. Puldón, and R. A. E. Andrade, “Modelo clustering para el análisis en la ejecución de procesos de negocio,” *Investig. Operacional*, vol. 33, no. 3, pp. 210–221, 2012.

[3] Y. josefina M. Aular and R. T. Pereira, “Minería de datos como soporte a la toma de decisiones empresariales,” *Opción*, vol. 23, no. 52, pp. 1–6, 2007.

[4] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, “A survey of machine learning for big data processing,” vol. 2016, no. 1. Springer International Publishing, 2016.

[5] C. W. Tsai, C. F. Lai, M. C. Chiang, and L. T. Yang, "Data mining for internet of things: A survey," vol. 16, no. 1, pp. 77–97, 2014.

[6] P. Kotler and K. L. Keller, *Marketing Management*, vol. 22, no. 4. 2009.

[7] J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression," *J. Bus. Res.*, vol. 60, no. 6, pp. 656–662, 2007.

[8] S. A. Robert C. Blattberg, Byung-Do Kim, "Database Marketing: Analyzing and Managing Customers," in *Springer Science & Business Media*, 2008, pp. 607–633.

[9] D. Birant, "Data Mining Using RFM Analysis," *Knowledge-Oriented Appl. Data Min.*, no. iii, pp. 91–108, 2011.

[10] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5259–5264, 2010.

[11] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 54, no. Second Edition. 2006.

[12] C. H. Cheng and Y. S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 4176–4184, 2009.

[13] J. B. MacQueen, "Kmeans Some Methods for classification and Analysis of Multivariate Observations," *5th Berkeley Symp. Math. Stat. Probab. 1967*, vol. 1, no. 233, pp. 281–297, 1967.

[14] R. S. Wu and P. H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," *Electron. Commer. Res. Appl.*, vol. 10, no. 3, pp. 331–341, 2011.

[15] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[16] C. Chang and S. Chen, "A comparative analysis on artificial neural network-based two-stage clustering," *Cogent Eng.*, vol. 2, 2015.

[17] C. Pete, C. Julian, K. Randy, K. Thomas, R. Thomas, S. Colin, and R. Wirth, "CRISP-DM 1.0: Step-by-step data minning guide," *Cris. Consort.*, p. 76, 2000.

[18] S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," vol. 5, no. 1, pp. 1–16, 2016.

[19] D. Birant, "Data Mining Using RFM Analysis," *Knowledge-Oriented Appl. Data Min.*, no. iii, pp. 91–108, 2011.

[20] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Mark. Cust. Strateg. Manag.*, vol. 19, no. 3, pp. 197–208, 2012.

[21] R. G. Drozdenko and P. D. Drake, *Optimal Database Marketing: Strategy Development and Data Mining*. SAGE Publications, 2002.

[22] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of Internal Clustering Validation Measures," 2010.

[23] C. Sugar and J. Gareth, "Finding the number of clusters in a data set : An information theoretic approach," *J. Am. Stat. Assoc.*, vol. 98, pp. 750–763, 2003.

[24] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus External cluster validation indexes," vol. 5, no. 1, 2011.

[25] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.

[26] Q. Razieh, M. Baqeri-Dehnavi, B. Minaei-Bidgoli, and G. Amooee, "Developing a model for measuring customer ' s loyalty and value with RFM technique and clustering algorithms," *J. Math. Comput. Sci.*, vol. 4, no. 2, pp. 172–181, 2012.

[27] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.

[28] R. J. Kuo, L. M. Ho, and C. M. Hu, "Cluster analysis in industrial market segmentation through artificial neural network," *Comput. Ind. Eng.*, vol. 42, no. 2–4, pp. 391–399, 2002.

[29] L. Ferreira and D. B. Hitchcock, "A Comparison of Hierarchical Methods for Clustering Functional Data," *Commun. Stat. - Simul. Comput.*, vol. 38, no. 9, pp. 1925–1949, 2009.

[30] L. Candillier, I. Tellier, F. Torre, and O. Bousquet, "Cascade evaluation of clustering algorithms," *17th Eur. Conf. Mach. Learn.*, vol. LNAI 4212, pp. 574–581, 2006.