AtoZ
novas práticas em informação e conhecimento

ISSN: 2237-826X

Análisis de Canasta de mercado en supermercados mediante mapas auto-organizados

Market basket analysis in supermarkets using self-organized maps

Joaquín Ignacio Cordero Lustig[1], Alfredo Bolt[2], Mauricio A. Valle[3]

[1] *Universidad Finis Terrae, Providencia, Chile. ORCID: https://orcid.org/0000-0003-2809-3133*

[2] *Universidad Finis Terrae, Providencia, Chile. ORCID: https://orcid.org/0000-0003-1901-9804*

[2] *Universidad Finis Terrae, Providencia, Chile. ORCID: https://orcid.org/0000-0003-1362-2776*

Autor para correspondência/*Mail to:* Alfredo Bolt, abolt@uft.cl

Recebido/*Submitted*: 15 de junho de 2021; **Aceito/*Approved*:** 29 de setembro de 2021

Resumo

Introucción: las canastas de mercado o canastas de compra, son todos los productos que son comprados por un cliente en un determinado momento. El análisis de estas canastas nos permite conocer las preferencias de nuestros clientes, lo que puede ser utilizado para diversos fines: operativos, publicitarios, estratégicos y logísticos. Lo mejor de todo: nos permite "predecir" sus futuras preferencias. Se presenta el caso de estudio de una importante cadena de supermercados de la zona poniente de la capital de Chile que necesita obtener información clave acerca de las canastas de compra de sus clientes para tomar decisiones. **Método:** se realizó un preprocesamiento de datos para transformar los datos originales en canastas de compra. Se utilizó un algoritmo de clústering de canastas de compra mediante redes neuronales artificiales de la clase mapas auto-organizados - self organizing maps (SOM). La utilización del algoritmo incluyó la búsqueda de los mejores hiperparámetros: tamaño de la grilla y tasa de aprendizaje. **Resultados:** el resultado del mejor SOM encontrado identifica finalmente seis clústeres de canastas de compra centradas alrededor de algún producto predominante e reconoce los productos más relacionados a ellos. **Conclusiones:** se han hecho recomendaciones sobre canastas de compra frecuentes a la cadena de supermercados que ha proporcionado los datos utilizados en la investigación.

Palavras-chave: SOM; Análisis de canasta de compra; Python, Minería de Datos; Redes Neuronales.

Abstract

Introduction: *market baskets, or shopping baskets, are all products that are purchased by a customer at a certain point in time. The analysis of these baskets allows us to know the preferences of our customers, which can be used for various operational, advertising, strategic and logistical purposes. Best of all: it allows us to "predict" their future preferences. We present the case study of an important supermarket chain in the western part of the capital of Chile that needs to obtain key information about their customer market baskets to make decisions.* ***Method:*** *data preprocessing was performed in order to transform the original data into market baskets. We clustered market baskets using artificial neural networks of the self-organizing maps (SOM) class. The use of the algorithm included the search for the best hyperparameters: grid size and learning rate.* ***Results:*** *the result of the best SOM found identifies six clusters of market baskets, each based in one predominant product, and identifies the products most related to them.* ***Conclusions:*** *recommendations on frequent shopping baskets have been made to the supermarket chain that has provided the data used in the research.*

Keywords: *Digital environments; Attitudes; Personal learning environments; PLE.*

INTRODUCCIÓN

El contexto de la presente investigación se refiere al análisis del comportamiento de los clientes de supermercados al momento de comprar productos. Esto porque determinados productos, habitualmente, se complementan con productos predeterminados; de esta manera, analizando una gran cantidad de datos relacionados a las boletas de compra de los clientes, se puede llegar a obtener patrones de compra de productos. De ahí el nombre "Market Basket Analysis" que trata del análisis de las canastas de compra de los clientes en supermercados.

Toda esta información transaccional (como la fecha, id de transacción, productos comprados por cada cliente) representa un gran volumen de datos. Cada compra en el sistema está asociada a una transacción en particular, la cual nos ayuda a identificar los productos adquiridos dentro de esa canasta. De esta manera, podemos encontrar patrones en los productos que se adquieren en conjunto. Estos patrones y relaciones pueden ser utilizados para conformar promociones y/o que eventualmente se puedan tomar decisiones con respecto a las compras de los clientes y establecer relaciones entre los productos: se podría determinar cuáles productos en las vitrinas podrían ir acompañados con sus complementos.

Debido a esta gran cantidad de información almacenada y a las complejas relaciones que cada producto de esta contiene, proponemos utilizar técnicas de inteligencia artificial (en concreto, redes neuronales de tipo mapas auto-organizados) para poder analizar, trabajar y descubrir estas relaciones.

Problemática

En esta investigación se trató el problema de análisis de canastas de compra en un caso de estudio concreto: una cadena de supermercado de la zona poniente de la capital de Chile, utilizando el enfoque previamente mencionado en relación con una red neuronal artificial o ANN (sigla en inglés).Se demostró, así, que la forma de resolver el problema con este método es posible, potente y ágil cuando se trata de una gran cantidad de datos que se quieran clasificar y para determinar el comportamiento de las características que definen a las neuronas dentro de la red.

Los datos utilizados corresponden a los registros de compras de 3 meses de una cadena de supermercados ubicada en la zona poniente de Santiago, capital de Chile. La necesidad de realizar este análisis e investigación nace de encontrar promociones o bundles "no típicas" para, de esta manera, potenciar las ventas.

La problemática en analizar el comportamiento de compra de los clientes de una cadena de supermercados es la cantidad de información que está asociada a las compras realizadas por los clientes. Y, preferencialmente, de un periodo no acotado de tiempo para tener una muestra aun mayor y, por ende, con más datos. La información por lo general está compuesta por miles de filas, lo que complica que una persona pueda a simple vista filtrar, revisar y analizar los datos en cortos periodos de tiempo y, en base a esto, determinar el comportamiento de compra de los clientes que se encuentra implícito en los datos. Para identificar el comportamiento entre los productos, se utiliza la técnica clave de análisis de canasta de mercado (MBA) que utiliza las grandes firmas de supermercados, consistiendo en descubrir las relaciones de los productos. Para esto, busca las combinaciones de productos que fueron adquiridos frecuentemente en la misma compra para identificar el comportamiento. Una vez determinado el comportamiento de los clientes, este se vuelve información clave y ventaja contra la competencia, ya que se logra saber qué canastas de compra están adquiriendo los clientes que frecuentan la cadena de supermercados que proporcionó los datos. De esa forma, se puede saber qué productos debe tener siempre en stock, ya que la ausencia de uno puede implicar en la no compra del complemento, también se puede utilizar para reducir el tiempo que los clientes pasan en el supermercado utilizando la infraestructura, lo que supone un costo para la cadena.

Para darle uso a los grandes volúmenes de datos que poseen los supermercados, producidos por cada venta que se realiza en las múltiples cajas, nace la oportunidad de utilizar los datos de las ventas realizadas para descubrir las complejas relaciones que existen entre los productos adquiridos por los clientes. Esto tiene relación con que compra cada producto y en qué cantidad, para llegar a promociones en base a las relaciones encontradas.

Con los datos transaccionales del supermercado y el resultado del análisis, se determinará el comportamiento de los clientes al momento de comprar, tomando en consideración las relaciones de los productos.

ESTADO DEL ARTE

Una canasta de compra es el conjunto de productos que un cliente compra durante una sola transacción. El análisis de canastas de mercado (i.e., *market basket análysis*) es un término genérico utilizado para describir las metodologías que estudian la composición de las canastas de compra adquiridas durante una sola compra.

Reglas de Asociación

Una de las principales aplicaciones del análisis de canastas de compra es el método de reglas de asociación. Esta es una de las principales técnicas para detectar y extraer información útil de datos de transacciones a gran escala (Hahsler & Karpienko, 2017). Existen varios trabajos en donde se hace referencia al estado del arte (Dunham, Xiao, Gruenwald, & Hossain, 2000); (Zhao & Bhowmick, 2003), los cuales son documentos que exponen y explican las Reglas de Asociación, sus factores claves y los diferentes tipos de algoritmos existentes para llevar a cabo su procedimiento.

Las reglas de asociación pueden responder algunas preguntas, entre ellas: qué tipo de productos tienden a ser comprados en conjunto por los clientes. Las reglas con una alta confianza (confidence) y un fuerte soporte (*support*) pueden denominarse reglas fuertes. Por ejemplo, en un estudio (Tsai & Chen, 2010) seleccionaron variables para la rotación de clientes de una compañía de servicio multimedia a pedido según las reglas de asociación para determinar el comportamiento de potenciales clientes y poder clasificarlos como VIP y no VIP según su nivel monetario, cantidad de veces que renueva servicios a pedido, entre otros atributos. También, en otro estudio (Shim, Choi, & Suh, 2012) propusieron estrategias de CRM basadas en reglas de asociación y patrones secuenciales para analizar los datos de transacciones de centros comerciales en línea de pequeño tamaño. Para obtener información desde las bases de datos, se utilizan comúnmente las reglas de asociación para buscar patrones existentes entre miles de transacciones (Agrawal, 1993b).

Mapas auto-organizados:

Las redes neuronales artificiales (RNA o en inglés ANN), pueden ser el paradigma que más representa el aprendizaje basado en pesos. Sus fundamentos teóricos están basados en el comportamiento del sistema nervioso de los animales, es decir, un sistema de neuronas interconectado de manera colaborativa y, de esta manera, puede producir un estímulo de salida o un resultado representativo.

Los SOM o mapas auto-organizados (Agrawal, 1993a) son un tipo de red neuronal artificial (RNA) que utiliza el método de entrenamiento no supervisado, que se usa para reducir las dimensiones de fuentes de datos n-dimensionales. Los resultados, por lo general, están conformados por dos dimensiones. Este mapa auto-organizado se encarga de agrupar y clusterizar los datos recibidos como entrada para, posteriormente, poder clasificar cada clúster encontrado y, así, sus componentes.

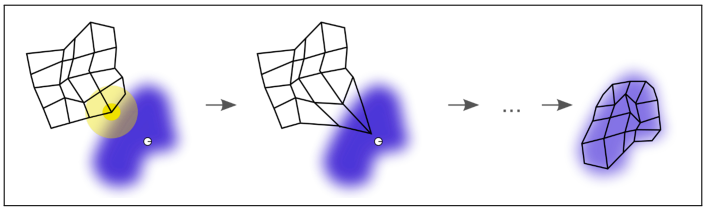


Figura 1. Funcionamiento de un SOM.

La Figura 1 ilustra el funcionamiento de un SOM. La mancha azul es la distribución de los datos de entrenamiento y el pequeño disco blanco es el dato de entrenamiento actual extraído de esa distribución. Al principio (izquierda) los nodos SOM se colocan arbitrariamente en el espacio de datos. Se selecciona el nodo (resaltado en amarillo) más cercano al dato de entrenamiento extraído. Se mueve hacia el dato de entrenamiento, al igual que sus vecinos en la cuadrícula (que se desplazan en menor medida). Después de muchas iteraciones, la cuadrícula tiende a aproximarse a la distribución de datos (derecha).

SOM utiliza diversos hiperparámetros para crear distintos modelos que se adapten a los datos y los agrupen de diversas maneras. Los principales hiperparámetros son las dimensiones de la grilla (filas x columnas) y la tasa de entrenamiento. Las dimensiones de la grilla tienen un efecto significativo en la cantidad y tamaño de los clústeres encontrados. Los SOM con grillas grandes producen un gran número de clústeres pequeños pero "compactos"(los datos asignados a cada clúster son bastante similares), mientras que los SOM con grillas pequeñas producen menos clústeres pero más generalizados.

Por otro lado, la tasa de aprendizaje determinará la velocidad de convergencia hacia una solución, pero muchas veces una alta tasa de aprendizaje puede evitar que encontremos algunas buenas soluciones y una tasa de aprendizaje muy baja puede provocar que no podamos salir de algunos óptimos locales. Durante el desarrollo de este trabajo exploramos diversos valores para ambos hiperparámetros en la búsqueda del mejor SOM para nuestro problema.

Los mapas auto-organizados ya han sido utilizados para analizar las canastas de compra. La investigación realizada por Decker e Monien (2003) reporta sobre la utilización de mapas auto-organizados para el análisis de canastas de compra en una cadena de retail alemana. Este estudio logra establecer claras relaciones de co-compra entre productos utilizando SOM. Sin embargo, creemos que el estudio tiene un contexto de aplicación muy limitado: los datos utilizados corresponden a apenas 2.000 canastas de compras de 25 productos y estos pertenecen a una tienda de artículos de higiene personal/propósito general, no a un supermercado.

En este artículo, al igual que en el estudio mencionado (Decker & Monien, 2003), aplicamos SOM a datos de canastas de compras, pero con un mayor volumen de datos y productos, un mejor análisis de sensibilidad de parámetros y utilizamos datos reales de una cadena de supermercados de la ciudad de Santiago, Chile.

METODOLOGÍA

La cadena de supermercados nos proporcionó una base de datos con compras de clientes compuesta por 146.621 filas. Cada fila hace referencia a una compra de producto individual con sus datos correspondientes.

La base de datos proporcionada contiene 9 columnas, las cuales son: Identificador de cliente, fecha de la transacción, día de la semana (1 a 7), día del mes (1 a 31) el año, categoría, subcategoría, nombre del producto y precio. La Tabla 1 muestra un extracto de datos incluidos en esta base de datos.

ID Cliente	Fecha Transacción	Dia	Mes	Año	Categoría	Sub-categoría	Producto	Precio
429103	3/09/2011	7	9	2021	Leche Abarrotes	Leche polvo descremada	Leche polvo descremada x 800g	2.990

Tabela 1. Ejemplo fila Base de Datos.

Se puede apreciar en la Tabla 1 que el campo "Día" representa el día de la semana y tiene un valor entre 1 y 7, donde 1 es lunes y 7 domingo. Además, cada fila representa la venta de un producto individual, pero no hay información sobre canastas de compras: es decir, productos que se vendieron en conjunto.

Con el fin de identificar las canastas de compra en esta Base de Datos, realizamos un preprocesamiento de éstos con la herramienta RapidMiner Studio (Team, 2019), la cual permite el desarrollo de procesos de análisis y minería de datos mediante el encadenamiento de operadores los cuales ofrecen parámetros configurables para cada uno, entradas y salidas. Gracias a esta útil herramienta se desarrolló un flujo de trabajo para realizar el preprocesamiento y filtrado de datos relacionados con cada compra realizada.

Este flujo es presentado en la Figura 2 y se encuentra disponible en el enlace[1].

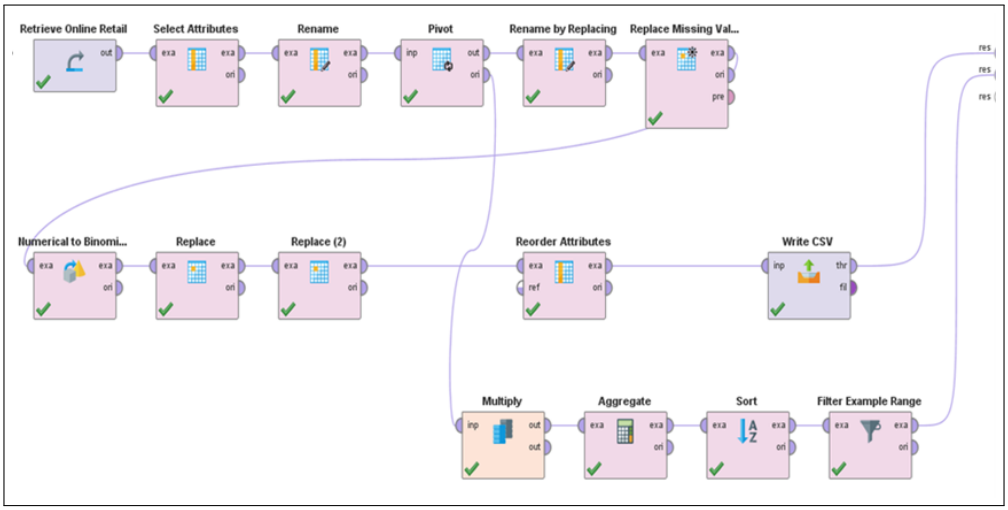


Figura 2. Flujo de preprocesamiento de datos en RapidMiner Studio.

En este flujo asumimos que los productos comprados en el mismo día por el mismo cliente corresponden a una misma canasta de compra. Posteriormente, utilizamos un script en Python[2] para remover y renombrar columnas.

En concreto, la combinación del preprocesamiento en RapidMiner y del script en Python nos permite tomar los datos de ventas de productos individuales mostrados en la Tabla 1 y agruparlos en canastas de compra según su Fecha de Transacción e ID de cliente.

A estos dos atributos, se suman como columnas todos los productos comprados en todas las canastas para confeccionar el vector de productos posibles. Finalmente, agrupamos todas las compras que ha hecho un mismo cliente (ID Cliente) dentro de un mismo día (Fecha Transacción) en una misma "Canasta de Compra" y marcamos con un 1 en el vector correspondiente todos los productos que fueron comprados por ese cliente en ese día o 0 si el producto no fue comprado por ese cliente en ese día.

Utilizando este supuesto pudimos transformar los datos presentados en la Tabla 1 en vectores de canastas de compra como se muestra a continuación en la Tabla 2.

ID Transación	Fecha	Atún	Fideos	Arroz grado 2	Yogur	Aceite
001	07/06/2011	0	1	1	0	1

Tabela 2. Vector de canasta de compra.

En la Tabla 2, cada fila corresponde a una transacción o "canasta de compra" (generada a partir de la canasta de compra de un cliente en particular), en donde un valor de 0 o 1 indica la ausencia o presencia de un producto

[1] https://drive.google.com/file/d/1qG0x4c4ozlIPyswDq$_{B4}$hEyEdNBYOAL/view?usp = sharing
[2] https://github.com/pitfox94/SOM

en esa canasta. En el ejemplo de la Tabla 2, podemos ver una canasta de compra ID 001 hecha en la fecha 07/06/2011 en donde se compraron Fideos, Arroz grado 2 y Aceite, pero no se compraron Atún ni Yogur. Debemos notar que estos vectores de compra son largos, ya que tienen una "coordenada" por cada producto disponible en el catálogo.

Luego de pre-procesar los datos de la cadena de supermercados proporcionada en la base de datos, se observaron 48.358 transacciones o canastas de compra en 3 meses: 17.542 transacciones de compras realizadas por los clientes en el 1° mes, 15.089 en el 2° y 15.517 en el 3°.

DESARROLLO

Una de las características que distinguen una red Kohonen es que las relaciones topológicas (i.e., de vecindad) de las canastas de compra, utilizadas como parámetro de entrada, se reflejan en la disposición de las correspondientes celdas o unidades en la red neuronal artificial (Ritter & Schulten, 1986). El objetivo del análisis es agrupar canastas "similares", es decir, canastas que tengan características en común en subconjuntos disjuntos llamados clúster.

Si la propiedad que hace que la topología del mapa de Kohonen se conserve, los agrupa según el parámetro de entrada. Es decir, los subconjuntos de datos, que son vecinos cercanos, deben asignarse en la red o grilla del SOM conservando una relación cercana. Cualquier grupo de los vectores utilizados también debería aparecer en la red de menor dimensionalidad, en caso de que el producto sea comprado en su mayoría individualmente, como la "Cola".

Para investigar las capacidades de agrupamiento de un mapa de Kohonen (SOM), utilizamos los datos y parámetros de la siguiente manera: el total de transacciones de la muestra está compuesto por un total de 48.359 vectores transaccionales en donde cada vector representa una canasta de compra de manera binaria y, a su vez, cada vector tiene un largo de 189 productos. Se deben inicializar valores en la red o grilla del SOM con vectores binarios aleatorios con el mismo formato para poder hacer el proceso de comparación. Los números aleatorios son generados por la misma semilla gracias a la librería "numpy" para tener resultados determinísticos.

La precisión del resultado del entrenamiento del SOM en comparación con la muestra de datos reales está por sobre el 90%. Es decir, las relaciones halladas en las canastas de compras obtenidas del mapa entrenado tienen un 90% de precisión al momento de comparar cada una de las canastas obtenidas con canastas aleatoriamente seleccionadas de los datos reales.

A lo largo de la investigación se realizaron pruebas en Python antes de comenzar con los datos proporcionados por la cadena de supermercados, en un principio se desarrollaron y realizaron pruebas con los 48.359 vectores proporcionados para identificar diversos errores dentro del código y corregirlos, reportar avances inmediatos y realizar depuración (i.e., debugging) para entender la lógica y funcionamiento del algoritmo en desarrollo.

Estos datos luego se utilizaron para entrenar SOMs. Para este propósito, fueron empleados los siguientes parámetros: tamaño de la grilla (filas x columnas) tasa de aprendizaje e iteraciones (épocas). Las celdas obtenidas del entrenamiento del SOM no son adecuadas directamente para identificar grupos de datos según los parámetros de entrada. A continuación, se describirá un método llamado matriz-U que permite obtener una imagen más adecuada de la distribución vectorial.

La matriz-U contiene, por lo tanto, una aproximación geométrica de la distribución del vector en la red de Kohonen. Para obtener una visualización digital de cómo es esta distribución se propone utilizar una matriz-U de dos dimensiones, donde las celdas que contengan las distancias más pequeñas con sus vecinas serán graficadas en negro, mientras que las celdas más distantes serán graficadas en blanco. De esta manera, mientras más oscuro sea el conjunto de neuronas o celdas significa que tienen características similares entre ellas. Los lugares blancos o que tienden a volverse más claros, indican una distancia mayor entre las celdas y, por consecuencia, un cambio de características con respecto a las neuronas o celdas que las rodean. Por ejemplo, si se tiene una celda que contiene los siguientes valores de características (e.g., 2.0, 1.0, 1.5, 0.7) y las distancias euclidianas a las cuatro celdas vecinas son (e.g., 7.0, 12.5, 11.5, 5.0) luego la celda en la matriz-U tiene un valor de 36 antes de promediar y luego 9 después de aplicar el promedio. Un valor muy cercano a cero en una celda de la matriz-U indica que la celda está muy cerca de sus vecinos y, por lo tanto, las celdas vecinas conforman un grupo de características similares. La matriz-U implementada se grafica gracias a la librería "Matplotlib".

Luego de realizar pruebas con 48.359 vectores, una grilla de 20 filas y 20 columnas resultaban en una gran cantidad de neuronas vacías y con los mismos valores vectoriales. Por otra parte, una grilla compuesta por 8 filas y 8 columnas resultaba en una grilla con una agrupación de canastas muy generalizada, lo que dificultaba la detección de distintos grupos de canastas. Por lo cual, una grilla de 10 filas y 12 columnas permitía la visualización de 3 grupos de canastas. Estos resultados pueden deducirse de las matrices-U presentadas para cada configuración de tamaño de grilla exhibidas en la Tabla 3.

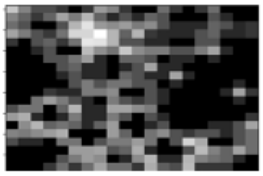
Nombre	Matriz-U 1	Matriz-U 2	Matriz-U 3
Matriz			
Filas	8	20	10
Columnas	8	20	12
Tasa	0,8	0,8	0,8
Iteraciones	20000	20000	20000

Tabla 3. Matrices-U para distintas dimensiones de Grilla.

Las Matrices-U, mostradas en la Tabla 3, están conformadas por la variación de los parámetros utilizados por el programa. Las dimensiones (filas x columnas) de la grilla influyen en la posible distancia que pueda haber entre cada nodo: una distancia muy grande puede generar información redundante dentro de la red y muchos clústeres poco representativos. En este caso, podemos diferenciar 3 clústeres principales. Esto se puede apreciar en la Matriz-U 2 en la Tabla 3: una distancia pequeña crea nodos muy generalizados que tienen muchos productos dentro de sus componentes y una red en la cual resulta difícil identificar clústeres de manera gráfica. Dado lo anterior, se optó por utilizar una dimensión de 10 x 12 en la cual se pueden apreciar 3 diferentes clústeres de manera clara.

La Tabla **??** tiene como finalidad representar el impacto que tiene la variación de la tasa de aprendizaje en el entrenamiento de SOM, como se puede apreciar en la Matriz-U 1 y 2 una tasa muy pequeña tiene como resultados grupos que no se diferencian tanto entre si tanto como por la distancia entre los clústeres.

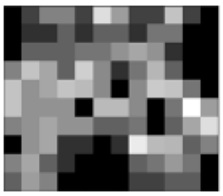
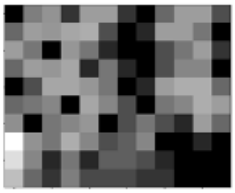
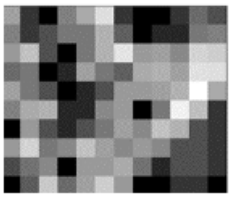
Nombre	Matriz-U 4	Matriz-U 5	Matriz-U 3
Matriz			
Filas	10	10	10
Columnas	12	12	12
Tasa	0,3	0,5	0,8
Iteraciones	20.000	20.000	20.000

Tabla 4. Matrices-U para distintas dimensiones de Grilla

La Matriz-U 3 de la Tabla 4 posee la ventaja de que sus clústeres encontrados son claramente visualizados gráficamente en la matriz resultante, lo que ayuda a determinar los productos que gobiernan esos clústeres y así acotar la información recibida, para que esta luego pueda ser analizada.

Para definir las dimensiones de la grilla del SOM se debe evitar perder relaciones de productos al tener una grilla muy pequeña o muchas neuronas vacías producto de una grilla muy grande. Luego de las pruebas, se determinó implementar una red de 10 filas y 12 columnas, la tasa de aprendizaje se fijó en 0.8 y la cantidad iteraciones que, a mayor cantidad de datos, más iteraciones deben ser, la cantidad de iteraciones utilizadas fue de 20.000.

Tomando en consideración lo anterior, se encontraron preliminarmente 3 agrupaciones o clúster de productos que comparten características en común, los cuales son descriptos en la Figura 5 como C1, C2 y C3. Estos grupos fueron conformados solamente tomando en consideración la distancia de las celdas con sus vecinas en la matriz-U, sin embargo, no tenemos información sobre qué tipo de productos están incluidos en estos clústeres,

por lo que usar la matriz-U no será suficiente para identificar los clústeres finales y será utilizada solamente para identificar los mejores parámetros para este modelo SOM: tamaño de la grilla y tasa de aprendizaje.

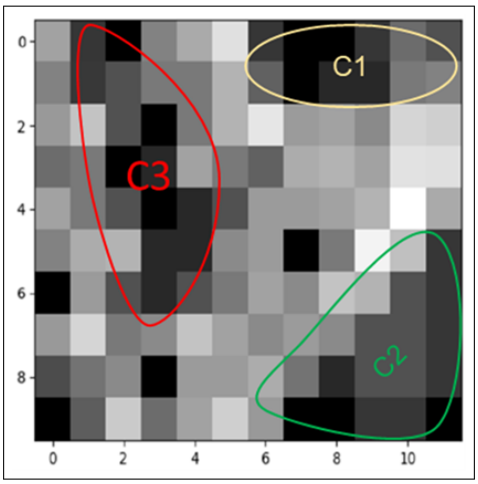
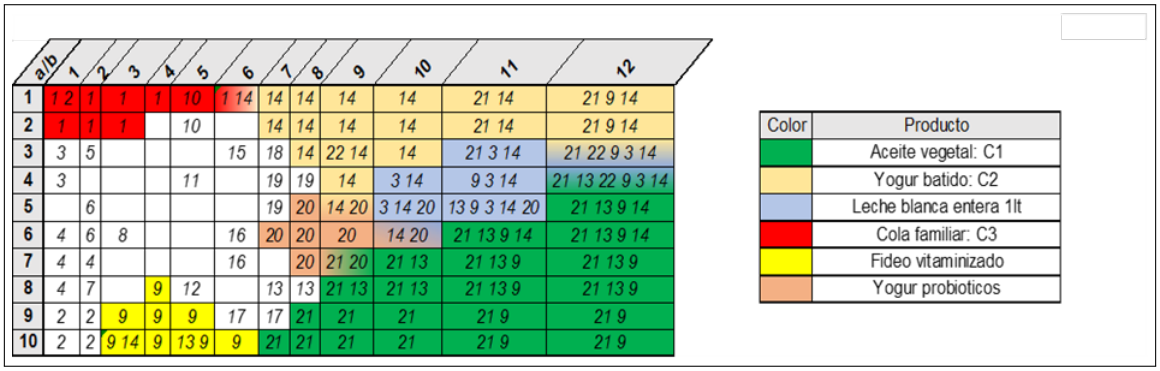


Figura 5. Clústeres de productos encontrados en la Matriz-U 3.

Para identificar los clústeres concretos que podemos encontrar mediante este SOM, debemos mirar al interior de las celdas resultantes y ver qué productos se encuentran en ellas. La Figura 6 muestra esta representación, donde en cada celda de la grilla se identifican los productos con asociaciones significativas a ellas (ver Tabla 3).



a/b	1	2	3	4	5	6	7	8	9	10	11	12
1	1 2	1	1	1	10	1 14	14	14	14	14	21 14	21 9 14
2	1	1	1		10		14	14	14	14	21 14	21 9 14
3	3	5				15	18	14	22 14	14	21 3 14	21 22 9 3 14
4	3				11		19	19	14	3 14	9 3 14	21 13 22 9 3 14
5		6					19	20	14 20	3 14 20	13 9 3 14 20	21 13 9 14
6	4	6	8			16	20	20	20	14 20	21 13 9 14	21 13 9 14
7	4	4				16		20	21 20	21 13	21 13 9	21 13 9
8	4	7		9	12		13	13	21 13	21 13	21 13 9	21 13 9
9	2	2	9	9	9	17	17	21	21	21	21 9	21 9
10	2	2	9 14	9	13 9	9	21	21	21	21	21 9	21 9

Color	Producto
	Aceite vegetal: C1
	Yogur batido: C2
	Leche blanca entera 1lt
	Cola familiar: C3
	Fideo vitaminizado
	Yogur probioticos

Figura 6. Mapa SOM 10 x 12.

Podemos identificar 6 clústeres distintos en base a la presencia de productos en celdas contiguas (a diferencia de los 3 clústeres identificados anteriormente solo con la matriz-U). Los 6 clústeres fueron coloreados por los autores para demarcar la presencia de ciertos productos emblemáticos. Notamos que sólo 22 de los 189 productos tuvieron asociaciones significativas a las celdas del SOM resultante. Esto se debe a que varios de los 189 productos son comprados con muy poca frecuencia y luego del entrenamiento del SOM no se encontró una relación considerable.

Los valores de las celdas del mapa de la Figura 6 representan el resultado del entrenamiento en cada una de las neuronas de la grilla del SOM, el cual consiste en determinar si éstos han sido adquiridos en la misma canasta de compra una cantidad de veces, lo que sugiere un comportamiento según la base de datos utilizada para entrenar el SOM.

Las celdas vacías o en blanco, representan una significativa diferencia entre las características de cada una de las celdas, lo que manifiesta la correcta agrupación de vectores transaccionales similares. Las celdas coloreadas pertenecen a un mismo clúster.

En la Figura 6 se pueden visualizar todos los índices de productos de la Tabla 3 y además se pueden identificar los tres clústeres encontrados en la Matriz-U 3 de la Figura 3, los cuales corresponden a Aceite vegetal, Yogur batido y Cola familiar, respectivamente.

RESULTADOS Y DISCUSIÓN

A continuación, analizamos algunos de los clústeres encontrados anteriormente. Con respecto al clúster C2, podemos notar que el elemento predominante en este clúster es el aceite vegetal. La Figura 7 muestra un diagrama de composición de frecuencia de la presencia de productos en canastas de este clúster: un 16.5% de

ID Producto	Nombre Producto
1	Bebida cola familiar
2	Bebida Sabores familiar
3	Leche blanca entera natural 1 lt.
4	Trutro pollo
5	Leche sabor 1 lt
6	Vino tinto caja
7	Pack bebidas
8	Bebida Cola Light
9	Fideo vitaminizado
10	Yogur con cereales
11	Yogur Bolsa
12	Quesillo Envasado
13	Arroz grado 2
14	Yogur batido
15	Leche Sabor
16	Yogur batido con fruta
17	Queso Crema
18	Leche polvo entera
19	Queso laminado granel
20	Yogur probiótico
21	Aceite vegetal
22	Cloro tradicional

Tabela 3. Tabla Id Productos.

las canastas de este clúster sólo contiene aceite vegetal, un 43.4% además contienen fideos vitaminizados y un 23.4%, adicionalmente, incluyen arroz grado 2 (es decir, aceite, fideos y arroz). Sin embargo, hay canastas que no contienen ('incluyen', para no repetir) aceite, como es el caso del 13.8% de canastas de este clúster, que solo contienen fideos o el 18% de las canastas de este clúster que solo contienen yogur batido. Esto tiene completo sentido ya que el aceite es un producto complementario a los fideos o arroz, ya que es un insumo necesario para cocinarlos.

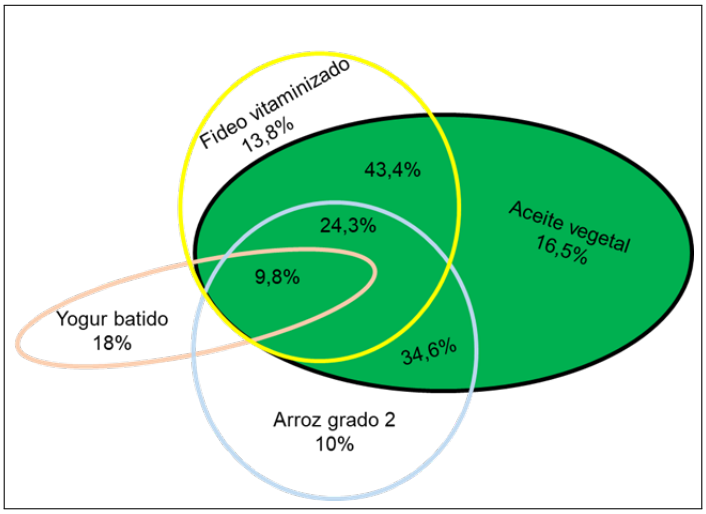


Figura 7. Relaciones aceite vegetal.

Con respecto al clúster C1 (ver Figura 8), el elemento predominante y más comúnmente comprado en este clúster es el yogur batido, el cual se compra en aislación en un 18.2% de las canastas de compra de este clúster. Podemos observar que de las canastas de compra de este clúster que incluyen yogur batido, casi un cuarto (24.1%) también incluyen (engloban, para no repetir) aceite vegetal. Esto puede ser explicado por el hecho de que el aceite es un producto de primera necesidad y, aparentemente, aparece en gran cantidad de canastas básicas, las cuales también incluyen (comprehenden, para no repetir) el yogur.

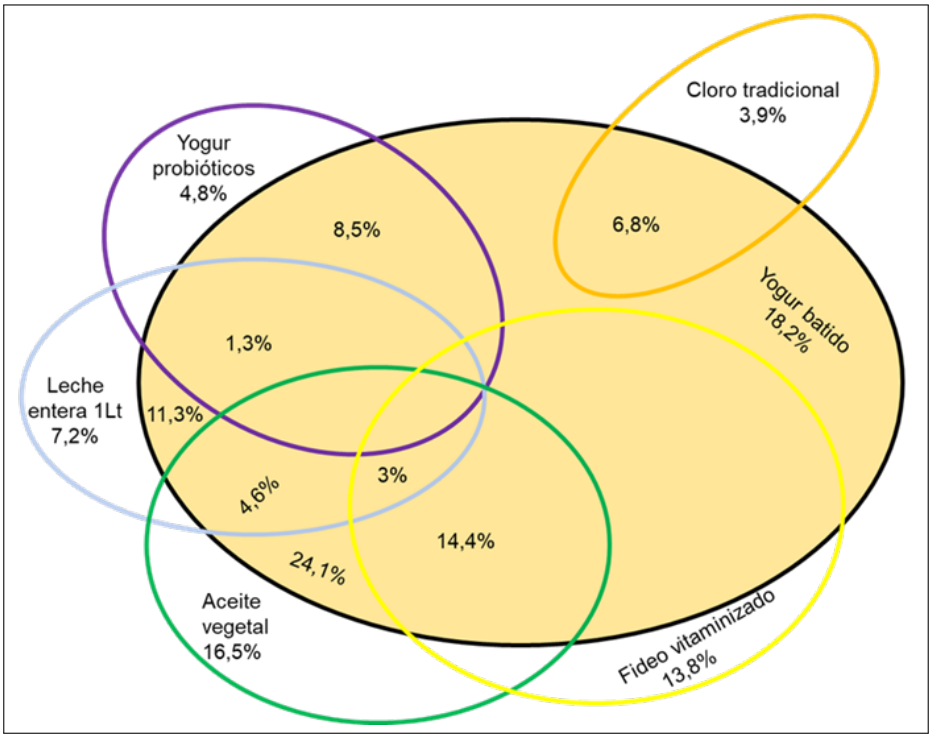


Figura 8. Relaciones yogur batido.

Con respecto al clúster C3 (ver Figura 9), el elemento predominante y más comúnmente comprado es la bebida cola familiar, el cual está presente en aislamiento en 11.6% de las canastas de compra del clúster. Las relaciones halladas para el producto "Cola familiar" y señaladas en la Figura 9 son bastante intuitivas y sencillas, ya que el hecho de que su complemento ideal sean las bebidas de sabores familiar, esto se puede apreciar en promociones y packs de bebidas de cola y sabores en los supermercados actuales en Chile.

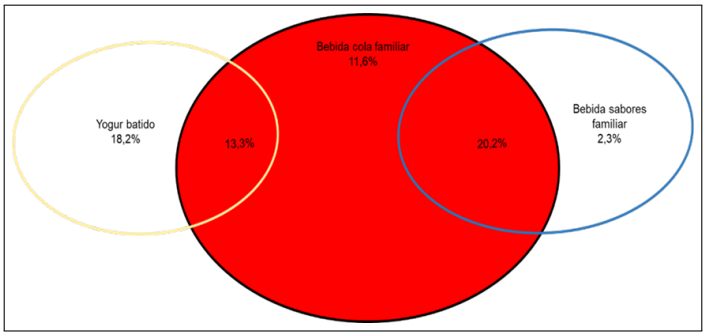


Figura 9. Relaciones cola familiar.

RECOMENDACIONES

En base a los resultados obtenidos, se pueden realizar las siguientes recomendaciones a la cadena de supermercados:

- El aceite vegetal, con los fideos vitaminizados y el arroz, deben ser colocados cada uno cerca del otro en las estanterías, ya que es altamente probable que estos productos se adquieran en una sola compra, por lo que también se recomienda conformar packs de promociones que conformen estos 3 productos en su conjunto.
- El yogur batido, se ve complementado con el aceite vegetal y a su vez con los fideos, sin embargo, el yogur requiere de refrigeración, a diferencia de los fideos o aceite, por lo que no podrían ponerse cerca en góndola. Recomendamos promociones o recordatorios de compras en este caso.
- Para la bebida cola familiar, se recomienda conformar packs que conformen bebida cola familiar en el doble de la cantidad de las bebidas sabores familiares y también localizar estos productos uno cerca del otro, para la inmediata adquisición de los complementos

Discussión

Al ser las redes neuronales artificiales y el análisis de datos, en general, un método en auge, a lo largo del tiempo se han ido desarrollando nuevas técnicas, tecnologías y librerías para llevar a cabo de manera eficaz para obtener, aun, mejores resultados, con más confianza y facilitando la visualización con cada actualización disponible. Por esto, la presente investigación podría seguir extendiéndose con el uso de estas nuevas tecnologías, ya que, al igual que el análisis de datos, el lenguaje de programación Python cada vez es más utilizado tanto como por compañías como por instituciones académicas, por la versatilidad de su programación para realizar diferentes tipos de tareas. Tanto así que se podrían automatizar la creación de SOM's con datos obtenidos desde API's para tener un mapa en tiempo real y tomar decisiones con información fresca.

CONCLUSIÓN

Los resultados en las tablas previamente señaladas sirven para visualizar de manera didáctica todas las interrelaciones halladas gracias al resultado del SOM. Estas interrelaciones corresponden significativamente a los productos más comprados por los clientes en las canastas provistas por la base de datos de la cadena de supermercados. Esta información resulta sumamente importante y útil al momento de tomar decisiones ya que, brinda una clara idea de cuáles productos se están comprando juntos, debido a que las relaciones encontradas son el resultado de la coincidencia de compra de los clientes que frecuentan la cadena de supermercados.

La integración de variables de la "canasta"que describen el comportamiento de compra de acuerdo con el tipo de productos comprados, constituye una mejora real en el modelo predictivo de comportamiento de compra. Sin embargo, si el número de productos es grande, esto lleva a la búsqueda de un indicador analítico que permita la caracterización de perfiles de personas estrechamente relacionadas al comportamiento, a menudo llamado "patrones".

Más allá de este uso de facilitar la comprensión global de un mercado, el comportamiento es coherente en términos de la alta frecuencia de compra de varios productos. El interés potencial por un uso predictivo se verificó en una aplicación empírica, como la del aceite vegetal con fideos y/o yogur con leche.

La ventaja de SOM por sobre otros métodos para enfrentar este tipo de problemas se debe a que posee una alta cantidad de herramientas, aplicaciones y librerías relacionadas, por lo que su implementación resulta más eficiente que algunos otros métodos como redes neuronales de gas o NGN (siglas en inglés).

Trabajo Futuro

La investigación se podría seguir extendiendo debido a que se puede lograr determinar el cambio del comportamiento de la canasta de compras en diferentes fechas del año, para identificar la influencia de la temporada en la decisión de los clientes al momento de consumir determinados productos.

REFERÊNCIAS

Agrawal, R. I. (1993a). Kohonen, t. *Proceedings of the IEEE*(78), 1464–1480. doi: 10.1109/5.58325

Agrawal, R. I. (1993b). Mining association rules between sets of items in large databases. *ACM SIGMOD Rules*, *22*(1), 207–216. doi: 10.1145/170036.170072

Decker, R., & Monien, K. (2003). Market basket analysis with neural gas networks and self-organizing maps. *Journal of Targeting, Measurement and Analysis for Marketing*, *11*, 373–386. doi: 10.1057/palgrave.jt.5740092

Dunham, M. H., Xiao, Y., Gruenwald, L., & Hossain, Z. (2000). *A survey of association rules.*

Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. *Journal of Business Economics*, *87*(3), 317–335. doi: 10.1007/s11573-016-0822-8

Ritter, R., & Schulten, K. (1986). On the stationary state of kohonen's self-organizing sensory mapping. *Biological cybernetics*, *54*(2), 99–106. doi: 10.1007/BF00320480

Shim, B., Choi, K., & Suh, Y. (2012). Crm strategies for a small-sized online shopping mall based on association rules and sequential patterns. *Expert Systems with Applications*, *39*(9), 7736–7742. doi: 10.1016/j.eswa.2012.01.080

Team, R. C. (2019). *Rapidminer studio: programa para el an*$_l$*isisyminer*$_a$*dedatos* (*v*. 135). *Boston. Recuperadode*

Tsai, C. F., & Chen, M. Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications*, *37*(3), 2006–2015. doi: 10.1016/j.eswa.2009.06.076

Zhao, Q., & Bhowmick, S. S. (2003). *Association rule mining: A survey* (Vol. 135). Singapore: Nanyang Technological University. Retrieved from https://personal.ntu.edu.sg/assourav/Unpublished/UP-ARMSurvey.pdf

Lusting, J. I. C., Bolt, A., & Valle, M. A. (2021). Análisis de Canasta de mercado en supermercados mediante mapas auto-organizados. *AtoZ: novas práticas em informação e conhecimento, 10*(3), 1 – 11. Recuperado de: http://dx.doi.org/10.5380/atoz.v10i3.81419