ANÁLISIS DE REGRESIÓN DE UNA RED DE TRANSACCIONES DE UNA TIENDA DE ABARROTES

REGRESSION ANALYSIS OF A GROCERY STORE TRANSACTION NETWORK

Karen Arlet Izquierdo Izquierdo
Tecnológico Nacional de México / IT de Celaya
karenarletizq@gmail.com

Salvador Hernández González
Tecnológico Nacional de México / IT de Celaya
salvador.hernandez@itcelaya.edu.mx

José Alfredo Jiménez García
Tecnológico Nacional de México / IT de Celaya
alfredo.jimenez@itcelaya.edu.mx

Recepción: 12/noviembre/2019 **Aceptación:** 20/febrero/2020

Resumen

Las tiendas de abarrotes que utilizan software de punto de venta en su negocio almacenan grandes cantidades de datos en forma de transacciones. La inteligencia de negocio transformará los datos en información útil para conocer el comportamiento del consumidor. En este estudio se modelaron transacciones como un sistema para conocer los componentes del mismo. Esto mediante la recopilación de los datos de un mes de ventas de una tienda de abarrotes en Centro, Tabasco. Con los cuales se creó una red de transacciones, aplicando el enfoque de redes complejas. Usando los cálculos obtenidos de los grados de nodos de la red, se identificaron productos con mayores ventas y las relaciones entre los productos del catálogo que la empresa ofrece. Por medio de un análisis de regresión se analizó el grado para predecir el comportamiento del consumidor.

Palabras Clave: Análisis de regresión, cesta de supermercado, grado, redes complejas.

Abstract

Grocery stores that use point-of-sale software in their businesses store large amounts of data in the form of transactions. Business intelligence will transform data into useful information to know consumer behavior. In this study transactions were modeled as a system to know the components of it. This by collecting data from a month of sales of a grocery store in Centro, Tabasco. Whereby the transaction network was created, applying the complex networks approach. Using the calculations obtained from the degrees of network nodes, products with higher sales were identified and the relationships between the products of the catalog that the company offers. Through a regression analysis, the degree to predict consumer behavior was analyzed.

Keywords**:** *Complex networks, grade, market basket, regression analysis.*

1. Introducción

Las pequeñas empresas se están enfrentando al reto de hacer uso eficiente a la tecnología con la que cuenta. El emplear software de punto de venta como base para el control de sus procesos implica el usar los datos almacenados para la mejora de toma de decisiones en sus procesos. [Moraleda, 2004] La minería de datos es una metodología para encontrar patrones y relaciones ocultas entre las variables contenidas en los datos guardados en bases de datos [Arpitha & Kumar, 2018]. En este trabajo se utilizará la técnica de minería de datos llamada análisis de la cesta de supermercado. Proponiendo la creación de una red compleja de un sistema de transacciones del mes de octubre de 2018 de una tienda de Abarrotes.

Destacando las bondades de las redes complejas, como son los valores de sus métricas tales como el grado. Y la representación visual de las asociaciones significativas mostradas en la red de ventas. Para esto se utilizará el software Gephi, es el software líder de visualización y exploración para todo tipo de gráficos y redes. Es de código abierto y gratuito [Gephi.org, 2017]. Permitirá representar la red, para posteriormente con las métricas de grado realizar un análisis de regresión para predecir el comportamiento de compra de los consumidores.

Cesta de supermercado

Una cesta de supermercado también conocido como aprendizaje de reglas de asociación o análisis de afinidad, es una técnica de minería de datos [Raeder, 2011]. Es la resultante de combinar distintas mercancías en diversas cantidades. Este concepto tiene gran importancia en la teoría de la conducta del consumidor y en concreto, en el estudio de las preferencias del mismo [Kaur & Kang, 2016].

En la actualidad, con el desarrollo de bases de datos, el tamaño de los conjuntos analizados ha aumentado a millones y hasta miles de millones de registros de elementos [Hair, Bush, & Ortinau, 2018].

Red compleja

Las redes complejas están compuestas de muchas partes llamadas nodos y unidas mediante relaciones los cuales reciben el nombre de enlaces y generalmente formando múltiples agrupaciones o islas [Espinosa, 2012].

El número de nodos, o N, representa el número de componentes en el sistema. A menudo llamaremos N al tamaño de la red. El número de enlaces o también conocidos por links, representa el número total de interacciones entre los nodos. Los enlaces rara vez se etiquetan, ya que se pueden identificar a través de los nodos que se conectan [Barabási A. L., 2015]. Los enlaces de una red pueden ser dirigidos o no dirigidos [Easley & Kleinberg, 2010]. Un enlace se dirige si se ejecuta en una sola dirección, en cambio los enlaces no dirigidos corren en ambas direcciones [Newman, 2003].

Grado

El grado del nodo de una red se denota como k_i. Por ejemplo, para la red no dirigida de la figura 1 se tiene $k_{1=2}$ $k_{2=3}$ $k_{3=2}$ $k_{4=1}$ [Barabási A. L., 2015].

En una red no dirigida el número total de enlaces, L, puede expresarse como la suma de los grados de los nodos [Aguirre, 2011]. La distribución de grado pk proporciona la probabilidad de que, al seleccionar un nodo al azar en la red, tenga el grado k [Barabási & Bonabea, 2003].

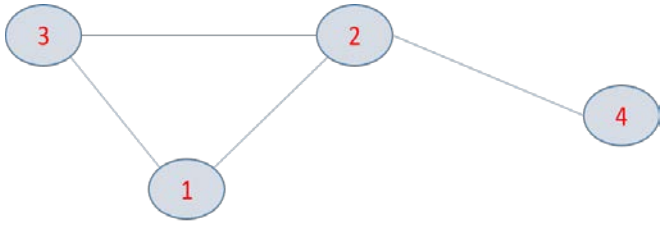

Figura 1 Ejemplo de grado de nodos.

Regresión lineal

Se tienen dos conjuntos de datos numéricos llamados X y Y, ambos con variabilidad aleatoria, esto es, no se puede predecir el valor exacto siguiente, pero se reconoce un patrón de cambio. Se sospecha que existe alguna relación entre dos grupos de números y se tienen n observaciones hechas por pares xj e yi. Cuando se observa el valor de X, xj, aparece observado el valor de Y como yi, a X se le llama variable independiente y a Y, variable dependiente, lo que significa que se cree que los valores de Y se pueden explicar o modelar en función de los valores de X. Lo primero que se recomienda en estos casos es dibujar una gráfica cartesiana que represente los pares de xj, yi como puntos [Hernández Ripalda, Tapia Esquivias, & Hernández Gonzalez, 2019].

2. Métodos

El método utilizado consta De cuatro etapas las cuales se muestran en figura 2:

- Etapa 1: Análisis de tickets: Para la elaboración de la red se utilizaron los tickets de ventas registrados en el software de punto de venta de la tienda de abarrotes, ubicada en Centro, Tabasco. En el mes de octubre se tienen registradas 986 transacciones. Las cuales fueron analizadas para encontrar las relaciones existentes entre productos.
- Etapa 2: Creación de red y componente gigante: Con las relaciones entre productos encontradas se construyó la red de transacciones empleando el software Gephi. Cabe mencionar que los nodos de la red son definidos por los productos que se vendieron en los tickets, dando un total de 360 nodos. Los enlaces presentan una conexión no dirigida, el peso de estos se considera como el número de veces que se vendieron los mismos productos en tickets

diferentes. Posteriormente, con la red de transacciones obtenidas de Gephi se hizo la depuración de nodos que presentaran grado menor que 1. Debido a que para el análisis de cesta de supermercado se necesitan conocer asociaciones significativas en productos. El resultado de eliminar los nodos con grado 0 se conoce como componente gigante.

- Etapa 3: Análisis de regresión de k: Con los resultados obtenidos en Gephi de métricas de grado, se realizaron intervalos de grados y se calculó su probabilidad y su logaritmo natural. El resultado obtenido de log nos permitió hacer el análisis de regresión. El tamaño del nodo es proporcional a k.
- Etapa 4: Predicción: La ecuación de regresión obtenida nos permitió predecir las compras futuras de los productos ofertados.

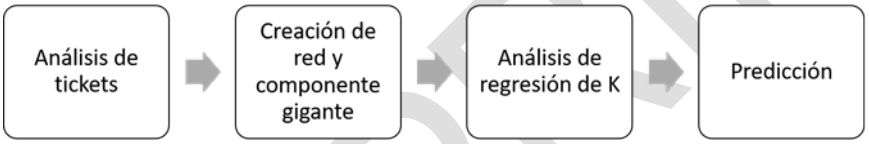

Figura 2 Etapas del método utilizado.

3. Resultados

La red de transacciones obtenida al emplear el método descrito anteriormente se muestra en figura 3, está compuesta por 360 nodos y 1313 aristas.

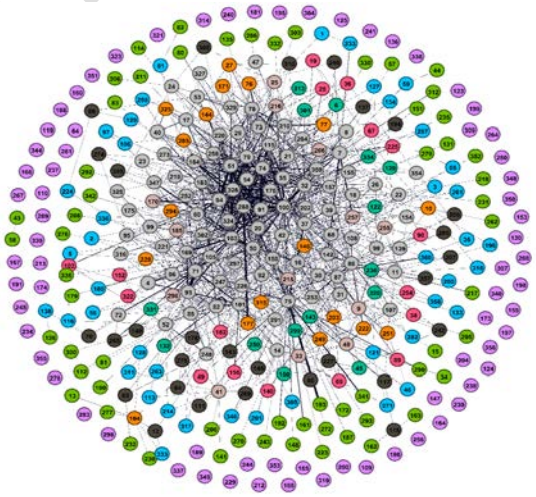

Figura 3 Red de transacciones sin eliminación de nodos.

La distribución utilizada en el software fue Fruchterman-Reingold, la cual permite una mejor visualización de las relaciones existentes entre productos. La figura 4 muestra la red de transacciones con componente gigante, la cual no tiene nodos con grado 0. Está compuesta por 276 nodos y 1243 aristas.

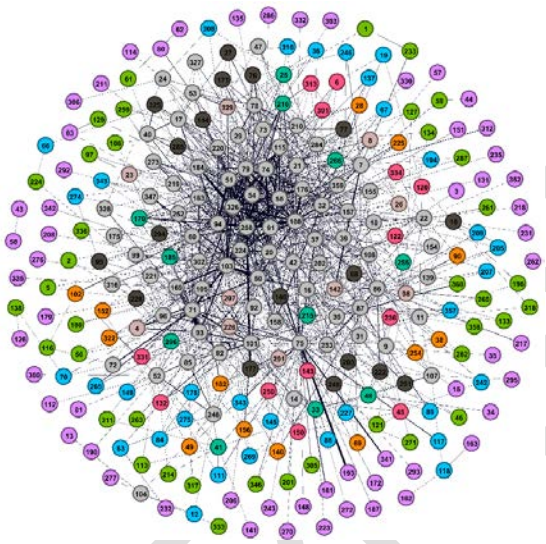

Figura 4 Red de transacciones aplicando componente gigante.

Grado

Al ingresar los nodos de productos con su ID y el peso de los enlaces en el programa Gephi; obtuvimos el grado de cada nodo. La tabla 1 muestra los resultados de los 5 nodos con mayor grado.

Tabla 1 Nodos con mayor grado.

Id	Label	Grado
100	Coca cola 3 L	77
288	Queso rojo a granel	69
103	Coca Cola Mini 250 ml	65
32	Azúcar 500 g	60
51	Carne de res a granel	57

Posteriormente con los valores de grado de cada nodo obtuvimos la probabilidad y probabilidad acumulada, valor al cual le calculamos su logaritmo (tabla 2). Cabe mencionar que se tomó en cuenta solo nodos que tienen grado igual o mayor a 1.

Tabla 2 Logaritmo de k y de probabilidad.

Grado	Probabilidad	Log k	Log probabilidad
1	0.210144928	0	-0.677481089
2	0.141304348	0.30103	-0.849844475
3	0.119565217	0.47712125	-0.922395142
4	0.047101449	0.60205999	-1.32696573
5	0.032608696	0.69897	-1.486666573

Los cálculos del análisis de regresión (tabla 3) se realizaron con el software Minitab [Minitab, 2013].

Tabla 3 Análisis de regresión.

	GL	SC Ajust.	MC Ajust.	Valor F	Valor P
	1	9.253	9.25294	205.52	0
$log\ k$	1	9.253	9.25294	205.52	0
Error	40	1.801	0.04502		
Total	41	11.054			
		S	**R-cuad.**	**R-cuad. (ajustado)**	**R-cuad. (pred)**
		0.212182	83.71%	83.30%	82.43%

Al realizar un análisis de regresión a los valores de grado se obtuvieron las ecuaciones de regresión para $Log\ probabilidad$ y $Probabilidad$, ecuaciones 1 y 2, respectivamente.

$$Log\ probabilidad = -0.609 - 1.0819\ log\ k \quad (1)$$

$$Probabilidad = 0.24603\ k - 1.0819 \quad (2)$$

4. Discusión

La aplicación al análisis de la cesta de la compra consiste en aislar las comunidades de nodos altamente conectados existentes en la red de productos. Esto permitirá identificar fuertes relaciones entre los productos y, por lo tanto, correlaciones significativas en el comportamiento de compra del cliente. Además, las comunidades pueden ser grandes, y representan relaciones mucho más expresamente y con menos redundancia que las reglas de asociación ordinarias. La red de productos presentada nos permite identificar nodos (productos) importantes

que la empresa debe tener en su stock, crear promociones con los productos con los cuales son comprados en la misma cesta, diseñar y ordenar los estantes de manera que los productos puedan ser visualizados y relacionados por los clientes para efectuar compras y con esto lograr un incremento en las ventas de la tienda. El diámetro de la red nos demuestra que es una red de mundo pequeño, al ser de tamaño 7, esto nos dice que el mayor número de artículos que existen en un ticket son 7. Una red de mundo es aquella en donde la distancia entre dos nodos elegidos aleatoriamente en una red es corta [Watts & Strogatz, 1998]. La importancia de un nodo se define de acuerdo al grado con el que cuenta. La presencia de un enlace no implica necesariamente una fuerte relación entre productos. La red cuenta con enlaces de grado 1 lo que significa que esos productos se compraron juntos solo una vez en el mes lo cual no representan relaciones sólidas. Algunos productos se compran con poca frecuencia con la mayoría de sus nodos vecinos, y frecuentemente con solo unos pocos. Con este análisis podemos definir que los productos como coca cola de 3 L, queso rojo a granel, coca cola mini 250 ml, azúcar 500 g, carne de res a granel, chile güero a granel, tortilla 500 g, cebolla blanca a granel, cilantro a granel, y carburo a granel son los 10 productos más vendidos en la tienda de abarrotes. Los cuales tienen que siempre estar presentes en el inventario para que puedan ser productos que sirvan de puente para que los consumidores adquieran otros que la tienda oferta. Estos productos son hub en la red, ya que permiten que un producto se conecte con otro. La figura 5 muestra un producto hub en la red el cual es coca cola 3 L.

La red de productos permite con facilidad detectar la alta conexión que tienen estos con productos con otros.

El grado promedio de la red de transacciones es de tamaño 9, lo que nos muestra que un producto tendrá como promedio 9 conexiones con otros productos del catálogo. Lo que nos dice que en promedio al comprar una coca cola 3 L. El consumidor podrá adquirir cebolla blanca granel, queso rojo a granel, azúcar 500 g, huevo blanco por pieza, leche lala light 1 L, tomate a granel, bebida Del Valle Frut cítricos 1.5 L, carburo a granel y ajo a granel. Aunque por ser un nodo con grado k alto podrá adquirirse con más de 9 productos del catálogo.

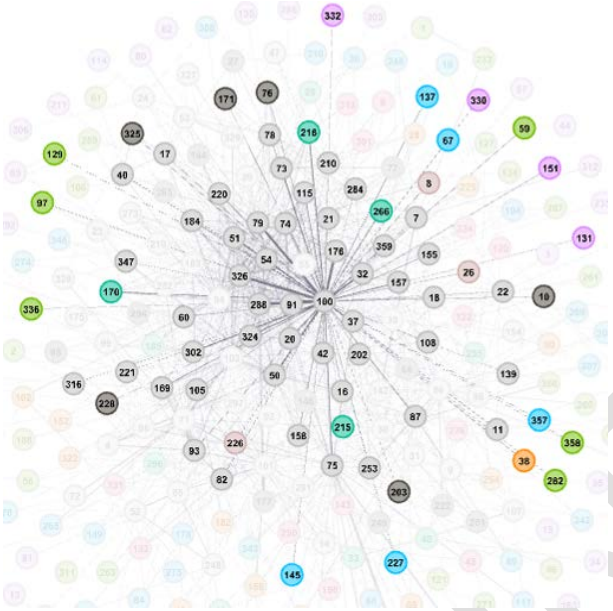

Figura 5 Relaciones del nodo 100 (Coca cola 3 L) con otros productos.

El coeficiente de correlación tiene un valor de 83.71%. Las gráficas (figuras 6 y 7) dispersión muestran la existencia de una correlación entre el grado del nodo k y la probabilidad pi de ser seleccionado aleatoriamente. La figura 6 nos muestra gráficamente la probabilidad de grado (k) vs el grado del nodo en escala decimal. La figura 7 nos muestra gráficamente la probabilidad de grado (k) vs el grado del nodo escala logarítmica.

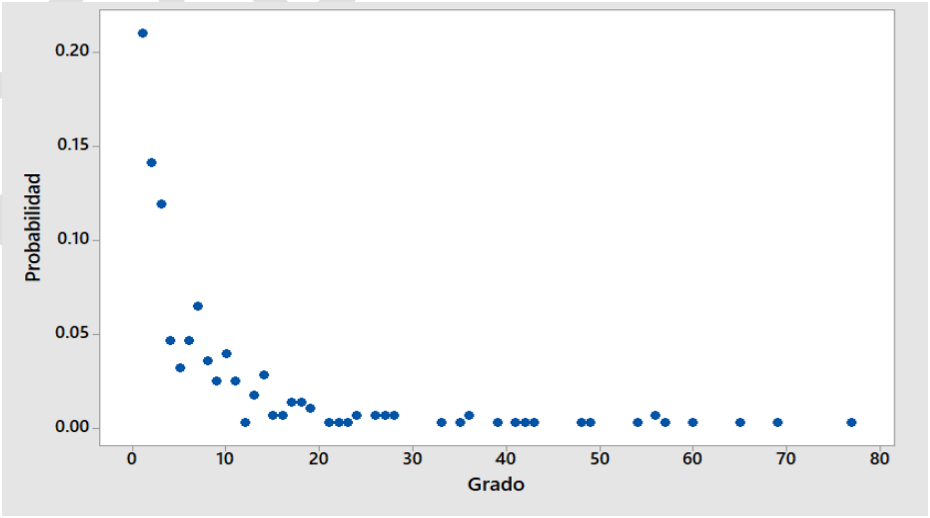

Figura 6 Probabilidad de grado (k) vs grado del nodo escala decimal.

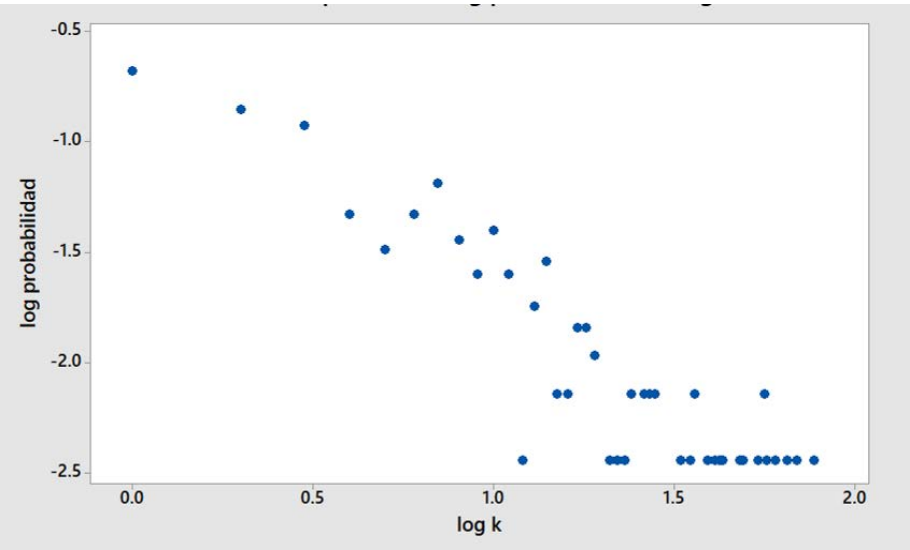

Figura 7 Probabilidad de grado (k) vs grado del nodo escala logarítmica.

La gráfica de dispersión (figura 8) presenta una disminución para los valores de grado $k > 15$, que nos dice que es muy poco probable que un nodo se relacione con otro teniendo un grado mayor que 15. Con base a la ecuación 2, se construyeron nuevos intervalos de clase, para predecir la probabilidad de los grados k.

Para la predicción del modelo se trabaja con las métricas de grado obtenidas del componente gigante. La tabla 4 nos muestra la diferencia de los datos reales vs la predicción del modelo obtenido.

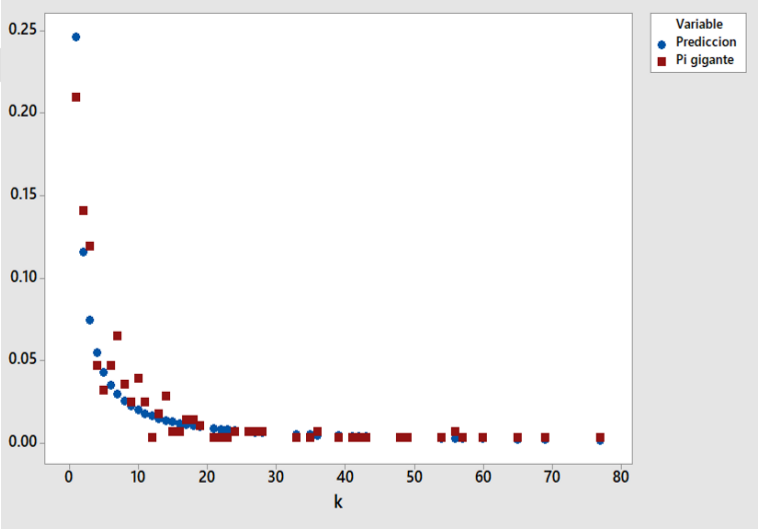

Figura 8 Curva de regresión ajustada del grado del nodo y predicción.

Tabla 4 Datos reales vs Predicción del modelo.

k	Probabilidad	Predicción
1	0.210144928	0.24603
5	0.032608696	0.043129301
22	0.003623188	0.008682024
39	0.003623188	0.00467321
54	0.003623188	0.003286331
77	0.003623188	0.002238689

La función exponencial predice la probabilidad de que un producto (nodo en la red) seleccionado al azar pertenezca la clase k. Al realizar el componente gigante la red de transacciones redujo su número de nodos de 360 a 276 (76.6% del total de nodos). De acuerdo al modelo de regresión obtenido hay una probabilidad de 0.043129301 de seleccionar un artículo que tenga grado $k \leq 5$. La figura 9 representa el nodo 72 (chicle max air 2 pzas.) y su relación con nodos.

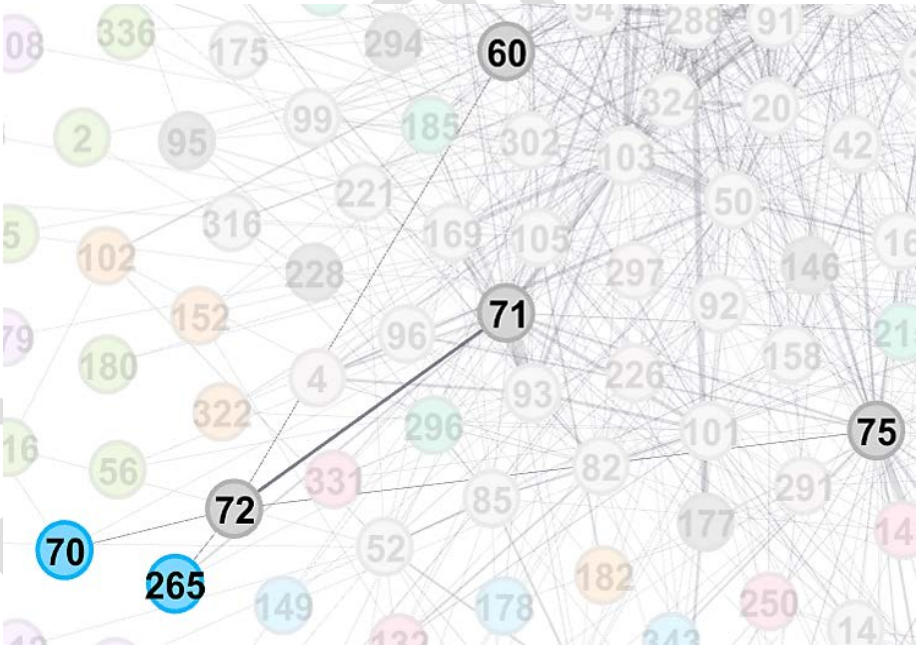

Figura 9 Nodo 72 (Chicle max air 2 pzas) con grado $k = 5$.

Si un cliente compra “chicle max air 2pzas” existen $k = 5$ alternativas en su cesta de supermercado, “cheetos bolitas sabor queso y chile 100 g”, “chicle bubbaloo fresa”, “chicle clorets 2pzas”, “paquete de navajitas gillete” y “chile morrón a granel”.

En conclusión, el análisis de regresión a la métrica de grado nos muestra que el uso de redes complejas en el análisis de un sistema de transacciones brinda un panorama de los productos más vendidos y su relación con otros productos del catálogo. Información relevante para la toma de decisiones de la gerencia de la tienda de abarrotes. La cual le permitirá detectar oportunidades de ventas tales como promociones y diseño de estantes, mejorar su gestión de inventario y conocer las preferencias de sus consumidores. Además, permite detectar que productos no tienen un número significativo de ventas y diseñar estrategias para incrementar las ventas.

5. Conclusiones

En conclusión, el análisis de regresión a la métrica de grado nos muestra que el uso de redes complejas en el análisis de un sistema de transacciones brinda un panorama de los productos más vendidos y su relación con otros productos del catálogo. Información relevante para la toma de decisiones de la gerencia de la tienda de abarrotes. La cual le permitirá detectar oportunidades de ventas tales como promociones y diseño de estantes, mejorar su gestión de inventario y conocer las preferencias de sus consumidores. Además, permite detectar que productos no tienen un número significativo de ventas y diseñar estrategias para incrementar las ventas.

6. Bibliografía y Referencias

[1] Aguirre, J. L. (5 de Marzo de 2011). Introducción al análisis de redes sociales: http://www.pensamientocomplejo.org/docs/files/J.%20Aguirre.%20Introducc%C3%B3n%20al%20An%C3%A1lisis%20de%20Redes%20Sociales.pdf.

[2] Arpitha, P., & Kumar, P. (2018). Market Basket Analysis for Data mining: Concepts and techniques. 8(4), 309-312.

[3] Barabási, A. L. (2015). Network Science. Recuperado el 10 de Noviembre de 2018: http://networksciencebook.com/chapter/2#networks-graphs.

[4] Barabási, A. L., & Bonabea, E. (2003). Scale-free networks. 288(5), 60-69.

[5] Easley, D., & Kleinberg, J. (2010). Networks, crowds, and markets. Academia.Edu: https://s3.amazonaws.com/academia.edu.documents/34741809/networks_crowds_and_markets.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1552294580&Signature=WVJusxmDYwN4WU2wbyqvE4VRZQ%3D&response-content-disposition=inline%3B%20filename%3DNetworks_Crowds_an.

[6] Espinosa, M. A. (2012). Redes complejas. Teoría y práctica. Tlatemoania, 1(11), 2-14.

[7] Gephi.org. (2017). Gephi: The Open Graph Viz Platform: https://gephi.org/.

[8] Hair, J. F., Bush, R. P., & Ortinau, D. J. (2018). Market Basket Analysis for data mining: Concepts and techniques. 8(4), 309-312.

[9] Hernández Ripalda, M. D., Tapia Esquivias, M., & Hernández Gonzalez, S. (2019). Estadística inferencial 2. Aplicaciones para ingeniería. México: Patria.

[10] Kaur, M., & Kang, S. (2016). Market basket analysis: Identify the changing trends of market data usin association rule mining. Procedia computer science, 1(85), 78-85.

[11] Minitab. (2013). Minitab 17 Software estadistico . Pennsylvania, Estados Unidos.

[12] Moraleda, A. (2004). La innovación , clave para la competitividad empresarial. Universia Business Review, 1(1), 128-136.

[13] Newman, M. E. (2003). The Structure and Function of Complex Networks. SIAM review, 45(2), 167-256.

[14] Raeder, T. (2011). Market basket analysis with networks. Social network analysis and mining, 1(2).

[15] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. Nature, 39(6684), 440.