

ESTUDIO DEL COMPORTAMIENTO DE COMPRA DE LOS ASOCIADOS Y CLIENTES DEL SERVICIO DE SUPERMERCADO DE UNA COOPERATIVA, APLICANDO REGLAS DE ASOCIACIÓN*

Olga Lucía Fernández Romero **
Enlace ORCID: <http://orcid.org/0000-0002-1921-2025>

Luis Armando Portilla Granados **
Enlace ORCID: <http://orcid.org/0000-0002-2559-8251>

José Orlando Maldonado Bautista ****
Enlace ORCID: <http://orcid.org/0000-0002-2281-8546>

Fecha de Recepción: Diciembre 2 de 2016
Fecha de Aprobación: Marzo 20 de 2016

Resumen:

La generación creciente de importantes volúmenes de datos, como resultado de la actividad de la empresa, ha obligado de forma generalizada al uso de recursos informáticos que permitan su tratamiento y disponibilidad, creando una cultura de confianza y uso adecuado de la información. Lo que no es tan frecuente encontrar es empresas que hagan uso del conocimiento que se puede generar de esos datos a partir de uso de herramientas de minería de datos. Las reglas de asociación se basan en métodos y técnicas de la minería de datos que permiten estimar o predecir comportamientos sobre datos recolectados como resultado del quehacer operativo de la empresa, con las que se espera conocer, a partir de la aplicación de modelos formales, el comportamiento de compra de los clientes de un supermercado de mediana superficie y proponer, a partir de este conocimiento, estrategias y acciones a seguir para de un lado fortalecer los hallazgos favorables encontrados, y de otro, mejorar las posibles falencias identificadas.

Palabras Claves: Gestión, Organización, Comportamiento de compra, Minería de datos, Reglas de Asociación.

* El presente artículo se ha obtenido como resultado del trabajo de grado de la autora en la especialización en Gerencia Estratégica de Mercadeo, de la Universidad Nacional Abierta y a Distancia-UNAD.

** Especialista en Gerencia Estratégica de Mercadeo. Gerente COOMULTUP LTDA

*** Profesor Titular - Universidad de Pamplona, adscrito a la Facultad de Ingeniería y Arquitectura en el programa de Ingeniería de Sistemas, pregrado en matemáticas y computación, con maestrías en ciencias computacionales y gestión de la calidad de educación superior, investigador del grupo de investigación Ciencias Computacionales – CICOM. Correo electrónico: laportia@unipamplona.edu.co

**** Profesor Asistente - Universidad de Pamplona, adscrito a la Facultad de Ingeniería y Arquitectura en el programa de Ingeniería de Sistemas, pregrado en matemáticas y computación, con maestría en ciencias computacionales y doctorado en ingeniería informática, investigador del grupo de investigación Ciencias Computacionales – CICOM. Correo electrónico: ormaldonado@unipamplona.edu.co

STUDY OF THE BEHAVIOR OF PURCHASE OF THE ASSOCIATES AND CLIENTS OF THE SUPERMARKET SERVICE OF A COOPERATIVE, APPLYING ASSOCIATION RULES

Abstract

The growing generation of large volumes of data as a result of the activity of the company, have widely forced to use computer resources to treatment and availability, creating a culture of trust and proper use of the information. What is not so common to find companies that use is knowledge that can be generated from these data using data mining tools. The association rules are based on methods and techniques of data mining that allow to estimate or predict behaviors on data collected as a result of the operational work of the company, with which it is expected to know, from the application of formal models, the Purchase behavior of the clients of a supermarket of medium surface and to propose, from this knowledge, strategies and actions to be followed to reinforce the favorable findings on the one hand and, on the other hand, to improve the possible identified shortcomings.

Keywords: Management, Organization Behavior purchase, Data Mining, Association Rules

ESTUDO DO COMPORTAMENTO DE COMPRA DE PARCEIROS E CLIENTES SERVIÇO DE SUPERMERCADO UMA COOPERATIVA, REGRAS DE ASSOCIAÇÃO APLICANDO

Resumo

A geração crescente de grandes volumes de dados, como resultado da actividade da empresa, foram amplamente forçado a usar recursos do computador ao tratamento e disponibilidade, criando uma cultura de confiança e uso adequado das informações. O que não é tão comum encontrar empresas que usam é o conhecimento que pode ser gerado a partir destes dados, utilizando ferramentas de mineração de dados. regras de associação baseiam-se em métodos e técnicas de mineração de dados que estimar ou prever comportamentos sobre os dados recolhidos, como resultado do trabalho operacional da empresa, que se espera venha a conhecer, a partir da aplicação de modelos formais, o o comportamento de compra dos clientes de um supermercado hipermercados e propor a partir deste conhecimento, estratégias e medidas de ação para um lado para reforçar os resultados favoráveis, e outros, melhorar as possíveis deficiências identificadas.

Palavras-chave: gestão, organização comportamento de compra, de mineração de dados, regras de associação

1. INTRODUCCIÓN:

El valor de la información en las organizaciones es indiscutible, el uso de las herramientas manejadoras de bases de datos ya hacen parte de casi la totalidad de empresas, lo que sí es discutible es el uso que se le está dando a la información y al aprovechamiento de los recursos tecnológicos que aporten información válida para la toma de decisiones.

Hoy día es indiscutible en cualquier tipo empresa el valor que tiene la información para lograr ventajas competitivas, implementando procesos y herramientas para la recopilación y tratamiento de sus datos, lo que no es común, y que se está haciendo cada vez más necesario, es la incorporación de métodos formales que permitan estudiar características y relaciones existentes entre los datos, y como interpretar estas relaciones para adquirir conocimiento sobre comportamientos que puedan servir de apoyo a los procesos de toma de decisiones en la organización.

El presente proyecto pretende aportar en esta búsqueda de conocimiento a partir de la naturaleza de cada empresa o institución y de los datos almacenados producto de sus transacciones, pretende utilizar técnicas de la minería de datos para encontrar patrones de predicción y comportamiento, que aporten conocimiento a partir de las relaciones que se puedan encontrar entre los datos.

Los resultados del proyecto de investigación van a traer, entre muchas otros aportes, la posibilidad de mejorar el inventario para atender efectivamente las necesidades de los usuarios, y poder disponer de las mercancías estratégicamente que además de brindarle comodidad al usuario, se le influencia para aumentar su compra cotidiana con nuevos productos relacionados con sus preferencias de compra, que se verán reflejados en el aumento de las ventas y en las satisfacción del cliente con el logro de su preferencia.

2. MARCO TEÓRICO:

2.1. Reglas de Asociación

Las reglas de asociación es una técnica de minería de datos sobre grandes volúmenes de información que permiten conocer cómo se relacionan los datos, que adquiere gran relevancia por su amplia variedad de campos de aplicación como ciencia, la educación, los deportes, la industria, los negocios, la medicina, etc.

Menzies define las reglas de asociación en función a su uso como: las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos (Menzies & Hu, 2003).

La idea de utilizar reglas de asociación para aplicar a conjuntos de elementos en grandes bases de datos fue de Rakesh Agrawal, Tomasz Imielinski y Arun Swami del Centro de investigación Almaden de IBM en California. El artículo en el que se trató este tema por primera vez tenía el nombre de "*Mining Association Rules between Sets of Items in Large Databases*" (Agrawal, Imielinski, & Swami, 1993).

Otros trabajos de finales de los 80 y principios de los 90, como la creación de la KDD, proceso de extracción de conocimiento, en inglés *Knowledge Discovery in Database* (Frawley, Piatetsky-Shapiro, & Matheus, 1991) y la propuesta de creación de el Algoritmo A priori para la aplicación de las reglas de asociación (Agrawal & Srikant, 1994), conforman el punto de partida a diversos estudios en diversas áreas y contextos de aplicación sobre los que se va a centrar el presente proyecto de investigación.

El procedimiento para la aplicación de reglas de asociación se da en dos momentos, uno la aplicación de un método eficiente para encontrar reglas de asociación (todas las reglas posibles), y dos, la utilización de métricas de evaluación de la calidad de las reglas de asociación (depurarlas – limitar el número de reglas), métricas entre las que se van a trabajar Soporte (*support*), Confianza (*confidence*) y Lift (*lift*).

2.2. Algoritmos

Las reglas de asociación cuentan con una serie de algoritmos diseñados para la generación de reglas y su depuración entre las que se destacan:

- **A priori:** Se puede decir que es el algoritmo clásico utilizado para encontrar Reglas de asociación en un conjunto de datos. Se denomina a priori debido a que se basa en el conocimiento previo de los conjuntos frecuentes, con lo cual se reduce el espacio de búsqueda y de tal forma que se aumenta la eficiencia (Agrawal & Srikant, 1994).
- **DHP:** El algoritmo DHP (*Direct Hasing and Pruning*), es una variación del algoritmo A priori. Propuesto por (Park, Chen, & Yu, 1997), los autores afirman que su algoritmo supone tres mejoras del anterior: la generación eficiente de grandes *itemsets*, reducción efectiva en el número de transacciones requeridas sobre la base de datos y la opción de reducir el número de búsquedas, sobre la base de datos. En general el algoritmo utiliza una técnica de *hashing* para filtrar

conjuntos de elementos innecesarios para la generación del siguiente conjunto de *itemsets* candidatos.

- **Partition.** Propuesto en (Savasere, Omiecinski, & Navathe, 1995), sus autores afirman que es fundamentalmente diferente de todos los algoritmos anteriores en que lee la base de **datos** a lo sumo dos veces para generar todas las reglas de asociación significativas. En contraste con los algoritmos anteriores, donde la base de datos no sólo es escaneada varias veces, sino que el número de escaneos no puede ser determinado por adelantado.
- **FP-Growth.** Fue introducido por (Han, Pei, Yin, & Mao, 2004), se basa en una representación de árbol de prefijos de una base de datos de transacciones llamada *Frecuente Pattern Tree*, o *FP-Tree*. El éxito de dicho algoritmo se basa en un esquema de eliminación recursiva, que intenta disminuir el tamaño de la base de datos de consulta.
- **Eclat.** Este tipo de algoritmos fue presentado por (Zaki, Parthasarathy, Ogihara, & Li, 1997). Su fortaleza consiste en la reducción de operaciones de entrada y salida, mediante la ejecución de un algoritmo de agrupamiento para aproximarse al conjunto de ítem frecuentes maximales y luego generar los ítems frecuentes presentes en cada clúster generado.

3. METODOLOGÍA:

3.1. Escenario de aplicación

El proyecto se realiza en un supermercado de mediana superficie, que hace parte de los servicios de una empresa cooperativa que ha venido reflejado en los últimos años un crecimiento en inversión y liderazgo en el sector comercial con productos de consumo, vestuario y electrodomésticos, a nivel local y regional, crecimiento que obliga a la administración a implementar estrategias de mercado que le permita proyectarse al nivel de los supermercados de gran superficie, donde se utilizan estrategias de mercado en el conocimiento del comportamiento de compra de sus clientes.

Actualmente el supermercado va incorporando productos en la medida que la demanda los requiere, con la aparición de nuevas líneas o con promociones y oportunidades de la época o por temporadas; y la estantería está organizada a partir de visiones de lo que se ha aprendido con la experiencia y del aporte informal de proveedores a través del personal de impulsores, que reflejan la priorización

de los productos y no la que debe primar como es las necesidades y comodidad del cliente.

3.2. Herramientas - RapidMiner:

Para la búsqueda de reglas de asociación se cuenta con un amplio grupo de sistemas de minería de datos, algunos licenciados como Clementine, Data Surveyor, Intelligent Miner, Darwin, Enterprice Miner, SGI MineSet, sin embargo existe gran variedad de herramientas de uso libre, de múltiples aplicaciones en minería de datos, calidad y fáciles de usar, entre las que se destacan Orange, RapidMiner, WEKA, JHepWork y KNIME. Para la aplicación del proyecto se apoya en la Herramienta RapidMiner.

RapidMiner es una herramienta de Minería de Datos ampliamente usada y probada a nivel internacional en aplicaciones empresariales, de gobierno y academia, implementa más de 500 técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de prueba de modelos, visualización de datos, etc. (Rapidminer, 2016)

Las reglas de asociación cuentan con una serie de algoritmos diseñados para la generación de reglas y su depuración entre las que se destacan: A priori, DHP, Partition, FP-Growth, Eclat.

RapidMiner implementa el Algoritmo FP-Growth, este operador calcula de manera eficiente todos los conjuntos de elementos frecuentes de la matriz determinada, utilizando la estructura de datos de FP-tree. Es obligatorio que todos los atributos de la base de entrada deben ser binominal.

3.3. Fases de investigación

El proyecto se basa en el desarrollo de fases formuladas en el proceso de extracción de conocimiento a partir de datos (KDD) (Hernandez, Ramírez, & Ferri, 2008), así:

3.3.1 Preparar los datos (selección, limpieza y transformación):

Determinación de fuentes de datos, identificación de datos relevantes al proyecto, depuración y adaptación de datos al modelo de aplicación.

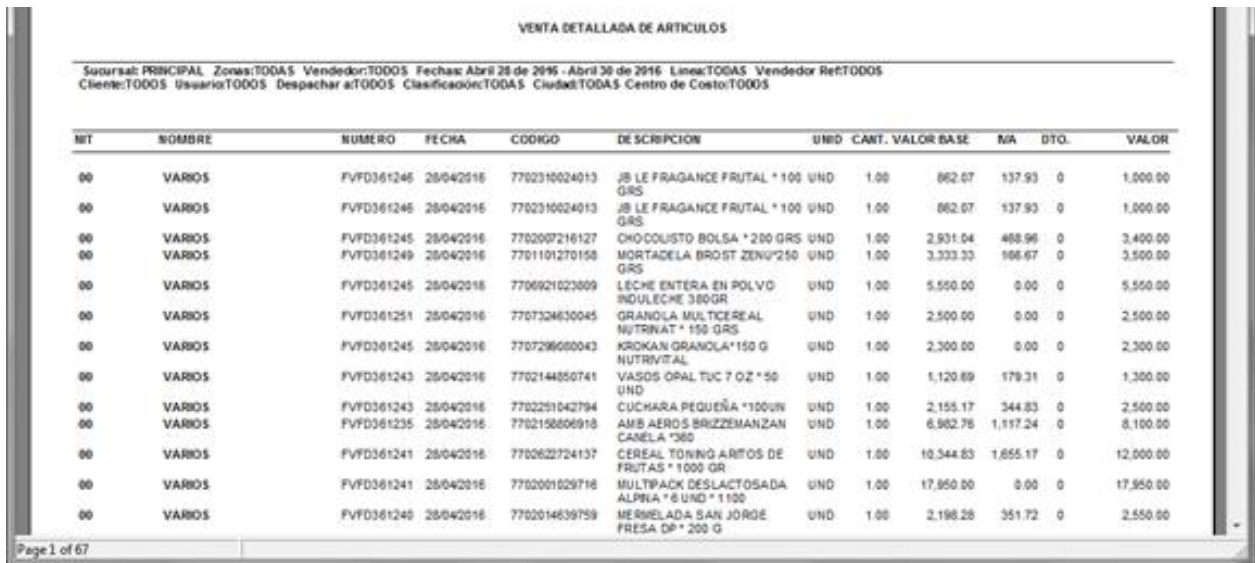
- La actividad inicia con un reconocimiento del sistema de base de **datos** con que cuenta la empresa y de forma específica las relaciones involucradas en las operaciones del supermercado.
- Seguida de la revisión y clasificación de los campos relevantes de la base de datos que se requieran y se ajusten al modelo de implementación definido.

- Y termina con la generación de tablas los datos relevantes y depurados (matriz binaria en formato Excel).
- Sobre el Sistema de Información Corporativo**

La cooperativa soporta todas sus operaciones financieras y administrativas sobre el sistema Contable Integrado Visual TNS 2015.

Para el objeto de aplicación del proyecto sólo es de interés el módulo de facturación, compras e inventarios, que registra las diferentes transacciones de Entrada - Salida de artículos y la emisión de los documentos correspondientes de acuerdo a los formatos de la empresa. Maneja varios precios sugeridos de venta y posibilita establecer políticas de descuentos dependiendo de los plazos de pago. Genera en forma integrada los comprobantes de contabilidad por ventas, costos de ventas, compras y notas de inventario. Calculo de costos con inventario permanente, de los que se utilizó el informe Ventas/Ventas Detalladas Vendedor/Cliente por Artículo, que permite generar resultados por rangos de fechas en formatos de tipo texto y Excel, agrupados por cliente, cada grupo con subtotales acumulados por rango de consulta, como se puede apreciar en las figuras 1 y 2.

Figura 1. Reporte TNS tipo texto (Fuente: TNS - Cooperativa)



VENTA DETALLADA DE ARTICULOS											
Sucursal: PRINCIPAL Zonas: TODAS Vendedor: TODOS Fecha: Abril 28 de 2016 - Abril 30 de 2016 Linea: TODAS Vendedor Ref: TODOS Cliente: TODOS Usuario: TODOS Despachar a: TODOS Clasificación: TODAS Ciudad: TODAS Centro de Costo: TODOS											
MIT	NOMBRE	NUMERO	FECHA	COODGO	DESCRIPCION	UNID	CANT.	VALOR BASE	IVA	DTOL.	VALOR
00	VARIOS	FVFD361246	28/04/2016	7702310024013	JS LE FRAGRANCE FRUTAL * 100 GRS	UND	1.00	862.67	137.93	0	1,000.00
00	VARIOS	FVFD361246	28/04/2016	7702310024013	JS LE FRAGRANCE FRUTAL * 100 GRS	UND	1.00	862.67	137.93	0	1,000.00
00	VARIOS	FVFD361245	28/04/2016	7702007216127	CHOCOLISTO BOLSA * 200 GRS	UND	1.00	2,931.04	488.96	0	3,400.00
00	VARIOS	FVFD361249	28/04/2016	7701101270158	MORTADELA BROST ZENU*250 GRS	UND	1.00	3,333.33	166.67	0	3,500.00
00	VARIOS	FVFD361245	28/04/2016	7706921523809	LECHE ENTERA EN POLVO INDULCEHE 380GR	UND	1.00	5,550.00	0.00	0	5,550.00
00	VARIOS	FVFD361251	28/04/2016	7707324630045	GRANOLA MULTICEREAL NUTRINAT * 150 GRS	UND	1.00	2,500.00	0.00	0	2,500.00
00	VARIOS	FVFD361245	28/04/2016	7707296080043	KROKAN GRANOLA*150 G NUTRIVITAL	UND	1.00	2,300.00	0.00	0	2,300.00
00	VARIOS	FVFD361243	28/04/2016	7702144850741	VASOS OPAL TUC 7 OZ * 50 UND	UND	1.00	1,120.89	179.31	0	1,300.00
00	VARIOS	FVFD361243	28/04/2016	7702251042794	CUCHARA PEQUEÑA *100UN	UND	1.00	2,155.17	344.83	0	2,500.00
00	VARIOS	FVFD361235	28/04/2016	7702158606918	AMB AEROS BRIZZEMANZAN CANELA *360	UND	1.00	6,962.76	1,117.24	0	8,100.00
00	VARIOS	FVFD361241	28/04/2016	7702622724137	CEREAL TONING ARTOS DE FRUTAS * 1000 GR	UND	1.00	10,344.83	1,655.17	0	12,000.00
00	VARIOS	FVFD361241	28/04/2016	7702001529718	MULTIPACK DESLACTOSADA ALPIA 18 UND * 1100	UND	1.00	17,950.00	0.00	0	17,950.00
00	VARIOS	FVFD361240	28/04/2016	7702014639759	MESENLADA SAN JORGE FRESA DP* 200 G	UND	1.00	2,198.28	351.72	0	2,550.00

Page 1 of 67

Figura 2. Reporte TNS tipo Excel (Fuente: TNS - Cooperativa)

ESTUDIO DEL COMPORTAMIENTO DE COMPRA DE LOS ASOCIADOS Y CLIENTES DEL SERVICIO DE SUPERMERCADO DE UNA COOPERATIVA,
 APLICANDO REGLAS DE ASOCIACIÓN

Olga Lucía Fernández Romero - Luis Armando Portilla Granados – José Orlando Maldonado Bautista

VENTA DETALLADA										
3										
4	Sucursal:	PRINCIPAL								Fecha:
5	Linea:	Todas								Ven:
6	Ciudad:	Todas	Usuario:							Vem Ref:
7										
8	NUMERO	FECHA	CODIGO	DESCRIPCION	UNID	CANT.	VALOR BASE	IVA	DTO.	VALOR
9	00	VIARIOS								
10	FVFD316070	13/01/2015	7707197477228	ARROZ GELVEZ * 3 KLS	UND	1,00	7.300,00	0,00	0,00	7.300
11	FVFD316069	13/01/2015	7702085012079	PASTA CABELLO DE ANGEL DORIA * 250 GRS	UND	1,00	1.190,48	59,52	0,00	1.250
12	FVFD316066	13/01/2015	7702166020153	DETERGENTE AS LIMON * 500 GRS	UND	1,00	2.068,96	331,04	0,00	2.400
13	FVFD316063	13/01/2015	7702166006058	JB REY AZUL * 3 UND * 300GR	UND	1,00	3.362,07	537,93	0,00	3.900
14	FVFD316061	13/01/2015	7702120007657	PAPEL HIG SUAVE GOLD * 4 UND	UND	1,00	5.000,00	800,00	0,00	5.800
15	FVFD316063	13/01/2015	7702626204659	VANISH PODER O2 BLANCO TOTAL * 210 GRS	UND	1,00	8.491,38	1.358,62	0,00	9.850
16	FVFD316065	13/01/2015	7702020053013	PASTA CONAZONI TORTELLINI CON CHAMPIÑ	UND	1,00	5.646,55	903,45	0,00	6.550
17	FVFD316063	13/01/2015	7702026168841	PH FAMILIA ACOLCHADO MEGA ROLLO * 12 UN	UND	1,00	13.793,10	2.206,90	0,00	16.000
18	FVFD316069	13/01/2015	7702084800288	HARINA PAN * 500 GR	UND	1,00	1.333,33	66,67	0,00	1.400
19	FVFD316065	13/01/2015	7702175108385	BICARBONATO EL REY * 50 GR	UND	1,00	603,45	96,55	0,00	700
20	FVFD316059	13/01/2015	7702175111248	COLOR BOLSA EL REY * 55 G	UND	1,00	775,86	124,14	0,00	900
21	FVFD316061	13/01/2015	10012012	MANI CON PASAS DELEITES ANGELITA	UND	1,00	1.000,00	0,00	0,00	1.000
22	FVFD316061	13/01/2015	10012012	MANI CON PASAS DELEITES ANGELITA	UND	1,00	1.000,00	0,00	0,00	1.000
23	FVFD316066	13/01/2015	10012012	MANI CON PASAS DELEITES ANGELITA	UND	1,00	1.000,00	0,00	0,00	1.000
24	FVFD316060	13/01/2015	7702166049024	JB AS BARRA DETERGENTE * 350 G	UND	1,00	1.120,69	179,31	0,00	1.300
25	FVFD316066	13/01/2015	7702166049024	JB AS BARRA DETERGENTE * 350 G	UND	1,00	1.120,69	179,31	0,00	1.300

Sobre los Datos

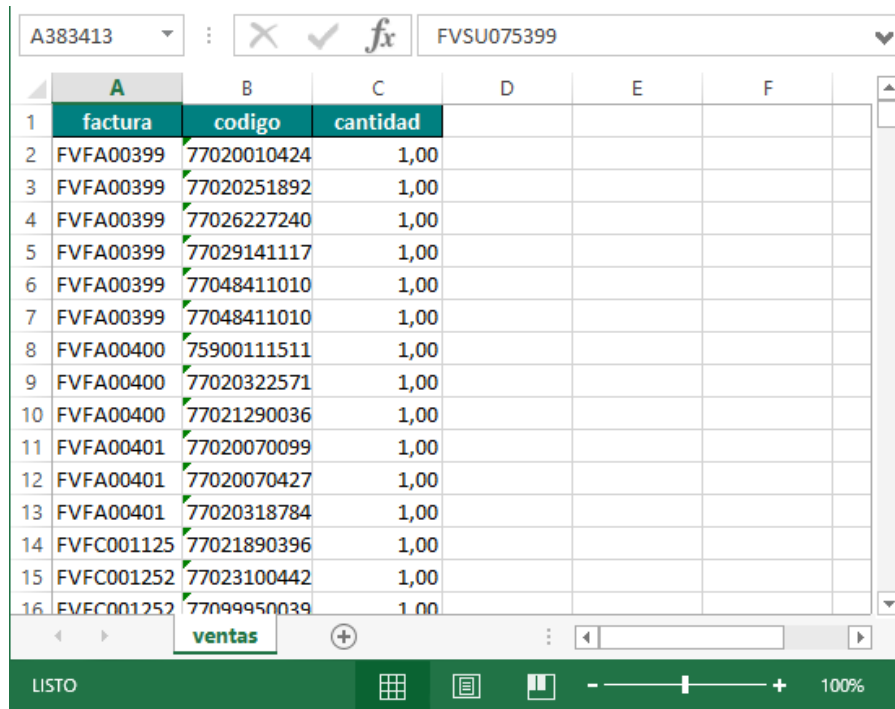
Los datos se obtienen en formato Excel, con consultas para las ventas realizadas durante el periodo 2015 y sobre los que se decide tomar sólo el mes de Septiembre como muestra, teniendo en cuenta las siguientes consideraciones:

- El volumen de datos de facturas generadas durante 2015 resulta muy alto, 383.412 registros de facturas
- Y segundo, por el objeto de la investigación, que es parte del supuesto de que lo facturado en un mes sirve para evidenciar el comportamiento de compra de los clientes, ya que se compraría lo necesario para un mes determinado y en los demás meses se espera que se compre más o menos lo mismo.

Los datos se depuraron de la siguiente forma, teniendo en cuenta que lo que se requiere es la información de los artículos comprados por factura, a fin de generar una matriz binaria de facturas X artículos, que represente los artículos que están presentes en cada factura, sin importar el número de artículos comprados.

- Del mes de Septiembre de 2015 se obtiene un archivo Excel con 33.030 registros de operaciones con los atributos: NUMERO, FECHA, CODIGO, DESCRIPCION, UNID, CANT., VALOR BASE, IVA, DTO., VALOR, COD. AREA, NOM. AREA, VENDEDOR, NIT VENDEDOR, DESPACHAR A, NIT DESPACHAR A, OBSERVACIONES
- En primer lugar el archivo se limpia de las filas que contienen subtotalet por grupo que genera la consulta a fin de dejar una tabla de filas columnas sin espacios.
- Seguidamente sólo se conserva los campos NUMERO (factura) que identifica las facturas, CODIGO (código) que identifica el artículo y UNID (cantidad) que tiene la información del número de unidades por cada artículo facturado. Figura 3.

Figura 3. Depuración de columnas (Fuente: los autores)



	A	B	C	D	E	F
1	factura	codigo	cantidad			
2	FVFA00399	77020010424	1,00			
3	FVFA00399	77020251892	1,00			
4	FVFA00399	77026227240	1,00			
5	FVFA00399	77029141117	1,00			
6	FVFA00399	77048411010	1,00			
7	FVFA00399	77048411010	1,00			
8	FVFA00400	75900111511	1,00			
9	FVFA00400	77020322571	1,00			
10	FVFA00400	77021290036	1,00			
11	FVFA00401	77020070099	1,00			
12	FVFA00401	77020070427	1,00			
13	FVFA00401	77020318784	1,00			
14	FVFC001125	77021890396	1,00			
15	FVFC001252	77023100442	1,00			
16	FVEFC001252	77099950039	1,00			

- Una vez depuradas las columnas en la columna “cantidad” se reemplaza cualquier número de unidades por unos (1), que es donde se encuentran las unidades de cada artículo facturados, como lo que se requiere construir es una matriz binaria, sólo con ceros (no se compró ese producto) y unos (se compró ese producto), como ya se anotó, para el estudio no es importante saber el número de unidades compradas en una factura, sino conocer si está (1) o no está (0) en una factura.
- En cuanto a filas se identifica la facturación de un servicio de préstamo rotativo en efectivo por caja que se factura y que no aplica para el estudio, por lo que se retiran las facturas asociadas a este producto (EFECTIVO y COMISION).
- Finalmente la transformación de los datos a Matriz, se genera la matriz binaria Facturas X Artículos, que es requerida para su aplicación de los algoritmos de reglas de asociación con la herramienta RapidMiner. Esta matriz se genera con la funcionalidad de Excel “Tablas dinámicas”, sabiendo que esto se puede hacer con operaciones de sistemas manejadores de bases de datos, que no son objeto del presente proyecto. La tabla dinámica se consigue a partir de 3 campos (facturas, códigos, cantidad) con lo que se obtiene una matriz que mantenga las filas con las facturas, que trasponga códigos (artículos) como columnas y la cantidad la intersección entre facturas y códigos (que sería los unos donde hay artículos en una factura y ceros donde no). Por efectos de espacio se cambia los códigos de las facturas por f1, f2, f3, ... y los artículos por a1, a2, a3, ..., quedando con 5.342 facturas (filas) y 5.520 artículos (columnas) que contiene 28.558 unos o artículos incluidos en facturas o en las canastas, que se muestra en la figura 4.

Figura 4. Matriz binaria (Fuente: los autores)

Olga Lucía Fernández Romero - Luis Armando Portilla Granados – José Orlando Maldonado Bautista

	A	B	C	D	E	F	G	H	I	J
1	facturas	a1	a2	a3	a4	a5	a6	a7	a8	a9
2	f1	0	0	0	0	0	0	0	0	0
3	f2	0	0	0	0	0	0	0	0	0
4	f3	0	0	0	0	0	0	0	0	0
5	f4	0	0	0	0	0	0	0	0	0
6	f5	0	0	0	0	0	0	0	0	0
7	f6	0	0	0	0	0	0	0	0	0
8	f7	0	0	0	0	0	0	0	0	0
9	f8	0	0	0	0	0	0	0	0	0
10	f9	0	0	0	0	0	0	0	0	0
11	f10	0	0	0	0	0	0	0	0	0
12	f11	0	0	0	0	0	0	0	0	0
13	f12	0	0	0	0	0	0	0	0	0
14	f13	0	0	0	0	0	0	0	0	0
15	f14	0	0	0	0	0	0	0	0	0
16	f15	0	0	0	0	0	0	0	0	0
17	f16	0	0	0	0	0	0	0	0	0
18	f17	0	0	0	0	0	0	0	0	0

- Haciendo sumatorias por filas se puede calcular el número de artículos diferentes incluidos en cada factura y sobre el que se realiza una nueva depuración, teniendo en cuenta que lo que se busca es conocer la relación entre los artículos comprados, pues es claro que las facturas que compraron un solo artículo no pueden generar regla alguna, y buscando mejorar estas relaciones se decide incluir en el estudio sólo las facturas con 10 o más artículos comprados, quedando con una matriz de 732 facturas X 4.398 artículos, donde se aprecia que la mayoría de las facturas generadas (4.610) en el mes de Septiembre de 2015 tienen 9 o menos productos, de los que 1.937 corresponden a facturas con un solo artículo.

3.3.2 Aplicación de técnicas y herramientas de minería de datos

Modelamiento y aplicación de métodos de minería de datos para extraer patrones previamente desconocidos, potencialmente útiles y comprensibles, contenidos en los datos base de investigación.

Como ya se referenció el proyecto se apoya en la herramienta RapidMiner versión 7.1.000 (Figura 5) y en sus algoritmos para la generación de reglas de asociación (FP-Growth y Create Association Rules), que se dispone de los siguientes 4 pasos (ver además figura 6):

Figura 5. Entorno de trabajo Rapid Miner (Fuente: autores)

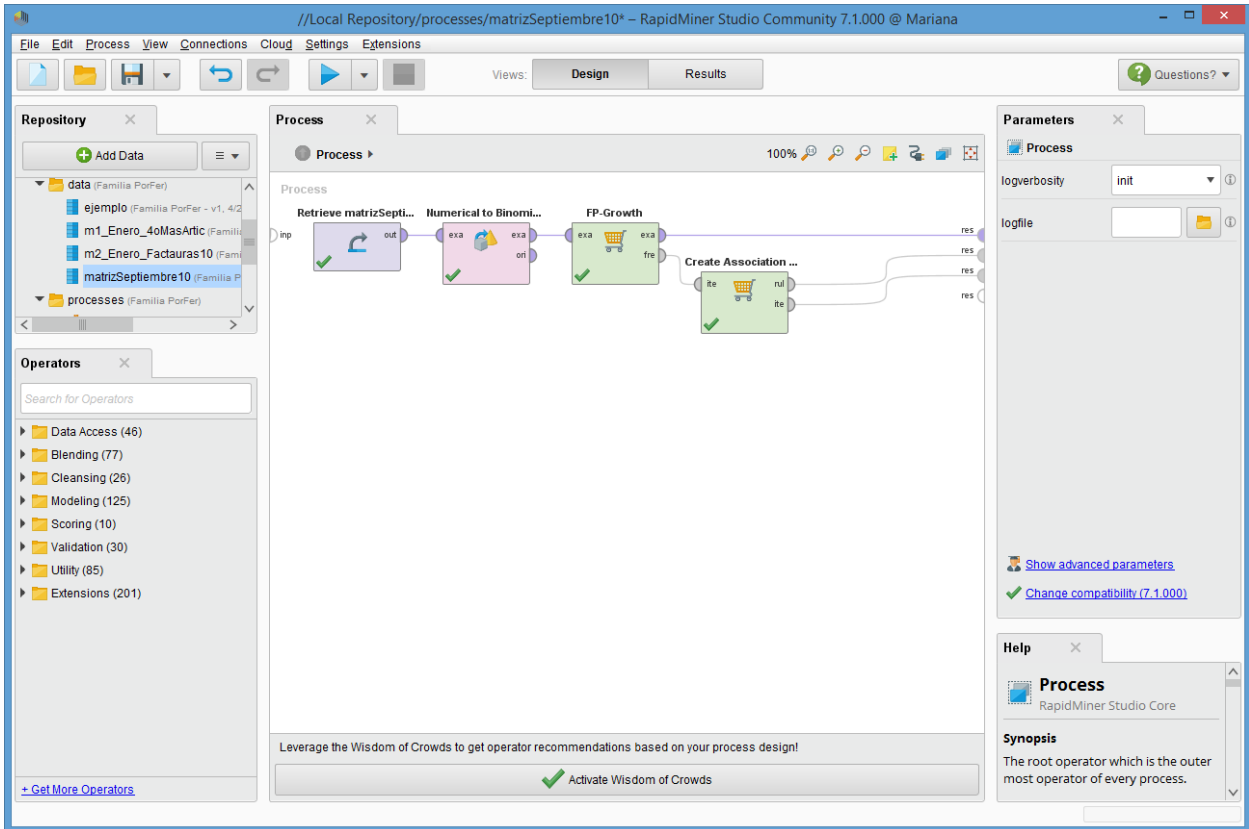
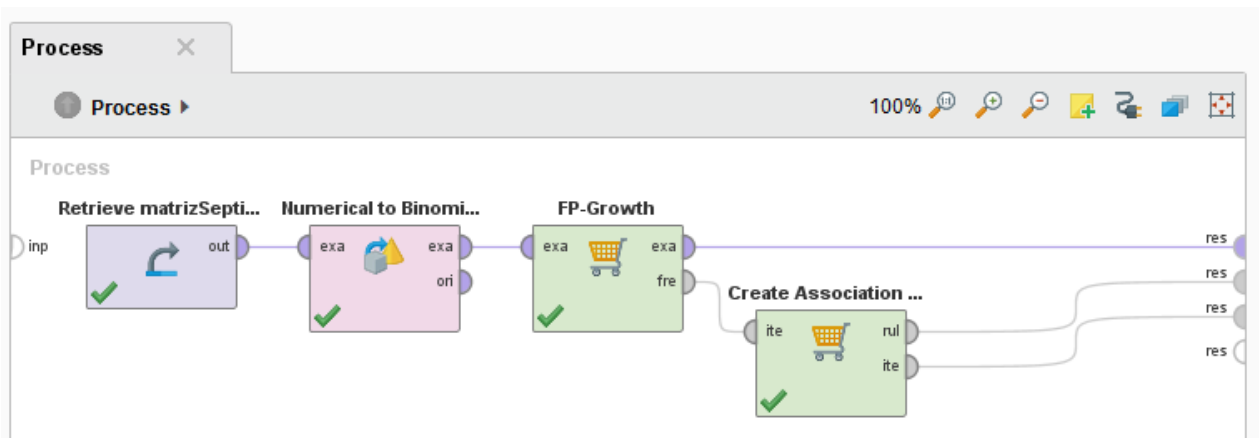


Figura 6. Process RapidMiner (Fuente: autores)



Paso 1: Cargar la matriz de datos (Retrieve). Este operador lee un objeto desde el repositorio de datos, el resultado es el observado en la figura 7.

Figura 7. Data (Resultado matriz binaria) (Fuente: autores)

Olga Lucía Fernández Romero - Luis Armando Portilla Granados – José Orlando Maldonado Bautista

Row No.	facturas	a10	a100	a1001	a1002	a1003	a1004	a1005	a1006	a1007	a10
1	f1006	0	0	0	0	0	0	0	0	0	0
2	f1015	0	0	0	0	0	0	0	0	0	0
3	f1027	0	0	0	0	0	0	0	0	0	0
4	f103	0	0	0	0	0	0	0	0	0	0
5	f1032	0	0	0	0	0	0	0	0	0	0
6	f104	0	0	0	0	0	0	0	0	0	0
7	f1060	0	0	0	0	0	0	0	0	0	0
8	f1083	0	0	0	0	0	0	0	0	0	0
9	f1086	0	0	0	0	0	0	0	0	0	0
10	f109	0	0	0	0	0	0	0	0	0	0
11	f1091	0	0	0	0	0	0	0	0	0	0
12	f1102	0	0	0	0	0	0	0	0	0	0
13	f1115	0	0	0	0	0	0	0	0	0	0
14	f1126	0	0	0	0	0	0	0	0	0	0
15	f1128	0	0	0	0	0	0	0	0	0	0
16	f1136	0	0	0	0	0	1	0	0	0	0
17	f114	0	0	0	0	0	0	0	0	0	1
18	f1152	0	0	0	0	0	0	0	0	0	0
19	f1156	0	0	0	0	0	0	0	0	0	0

Paso 2: Preprocesamiento de datos (*Numerical to Binominal*). Este operador cambia el tipo de los atributos numéricos seleccionados a un tipo binominal. También asigna todos los valores de estos atributos a los valores correspondientes de forma binominal. Mapea todos los valores numéricos a 'false' si se encuentran en el rango especificado (generalmente: igual 0.0) y al 'true' de otra manera, el resultado se muestra en la figura 8.

Figura 8. Resultado *Numerical to Binominal* (matriz binominal) (Fuente: autores)

Row No.	facturas	a10	a100	a1001	a1002	a1003	a1004	a1005	a1006	a1007	a10
1	f1006	false	false	false	false	false	false	false	false	false	false
2	f1015	false	false	false	false	false	false	false	false	false	false
3	f1027	false	false	false	false	false	false	false	false	false	false
4	f103	false	false	false	false	false	false	false	false	false	false
5	f1032	false	false	false	false	false	false	false	false	false	false
6	f104	false	false	false	false	false	false	false	false	false	false
7	f1060	false	false	false	false	false	false	false	false	false	false
8	f1083	false	false	false	false	false	false	false	false	false	false
9	f1086	false	false	false	false	false	false	false	false	false	false
10	f109	false	false	false	false	false	false	false	false	false	false
11	f1091	false	false	false	false	false	false	false	false	false	false
12	f1102	false	false	false	false	false	false	false	false	false	false
13	f1115	false	false	false	false	false	false	false	false	false	false
14	f1126	false	false	false	false	false	false	false	false	false	false
15	f1128	false	false	false	false	false	false	false	false	false	false
16	f1136	false	false	false	false	false	true	false	false	false	false
17	f114	false	false	false	false	false	false	false	false	false	true
18	f1152	false	false	false	false	false	false	false	false	false	false
19	f1156	false	false	false	false	false	false	false	false	false	false
20	f1157	false	false	false	false	false	false	false	false	false	false

Paso 3: Conjuntos de ítem (artículos) frecuentes (FP-Growth). Este operador calcula de manera eficiente todos los conjuntos de elementos frecuentes de la matriz determinada. Es obligatorio que todos los atributos de la matriz de entrada deben ser binominal (valores 'true' para los 1 y 'false' para los 0), el resultado se muestra en la figura 9.

Figura 9. Resultado *FrequentItemSets* (Fuente: autores)

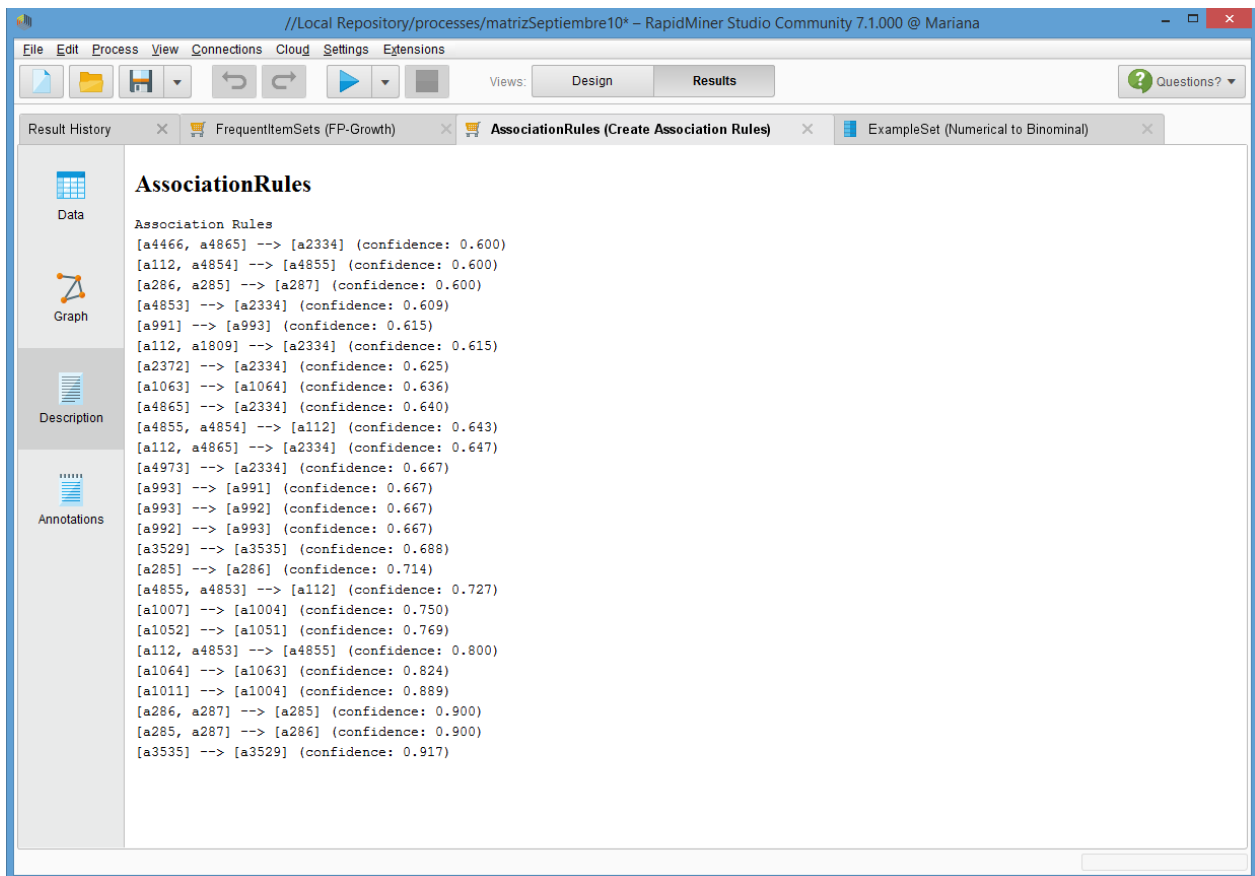
Size	Support	Item 1	Item 2	Item 3
2	0.014	a286	a287	
2	0.014	a4854	a4853	
2	0.019	a1106	a1105	
2	0.019	a1063	a1064	
2	0.014	a285	a287	
2	0.014	a1051	a1052	
2	0.012	a1005	a1007	
2	0.011	a1069	a1070	
2	0.015	a3529	a3535	
2	0.011	a991	a993	
2	0.011	a993	a992	
3	0.016	a2334	a112	a4466
3	0.015	a2334	a112	a4865
3	0.011	a2334	a112	a1809
3	0.012	a2334	a4466	a4865
3	0.012	a2334	a4466	a105
3	0.012	a112	a4855	a4854
3	0.011	a112	a4855	a4853
3	0.011	a4466	a4865	a105
3	0.012	a286	a285	a287

Paso 4: Crear reglas de asociación (*Create Association Rules*). Este operador genera un conjunto de reglas de asociación de los conjuntos de elementos frecuentes recibidos. Esta actividad termina como se muestra en la figura 10, en donde se muestran los datos finales. La figura 11 muestra la descripción de los resultados finales obtenidos.

Figura 10. Resultado *AssociationRules / Data* (Fuente: autores)

No.	Premises	Conclusion	Supp...	Confidence	LaPlace	Gain	p-s	Lift	Convict
8	a1063	a1064	0.019	0.636	0.989	-0.041	0.018	27.401	2.686
9	a4865	a2334	0.044	0.640	0.977	-0.093	0.031	3.346	2.247
10	a4855, a4854	a112	0.012	0.643	0.993	-0.026	0.010	4.902	2.433
11	a112, a4865	a2334	0.015	0.647	0.992	-0.031	0.011	3.383	2.291
12	a4973	a2334	0.011	0.667	0.995	-0.022	0.008	3.486	2.426
13	a993	a991	0.011	0.667	0.995	-0.022	0.011	37.538	2.947
14	a993	a992	0.011	0.667	0.995	-0.022	0.011	40.667	2.951
15	a992	a993	0.011	0.667	0.995	-0.022	0.011	40.667	2.951
16	a3529	a3535	0.015	0.688	0.993	-0.029	0.015	41.938	3.148
17	a285	a286	0.020	0.714	0.992	-0.037	0.019	18.030	3.361
18	a4855, a4853	a112	0.011	0.727	0.996	-0.019	0.009	5.545	3.186
19	a1007	a1004	0.016	0.750	0.995	-0.027	0.015	14.838	3.798
20	a1052	a1051	0.014	0.769	0.996	-0.022	0.013	26.813	4.209
21	a112, a4853	a4855	0.011	0.800	0.997	-0.016	0.010	15.411	4.740
22	a1064	a1063	0.019	0.824	0.996	-0.027	0.018	27.401	5.496
23	a1011	a1004	0.011	0.889	0.999	-0.014	0.010	17.586	8.545
24	a286, a287	a285	0.012	0.900	0.999	-0.015	0.012	31.371	9.713
25	a285, a287	a286	0.012	0.900	0.999	-0.015	0.012	22.717	9.604
26	a3535	a3529	0.015	0.917	0.999	-0.018	0.015	41.938	11.738

Figura 11. Resultado *AssociationRules / Description* (Fuente: autores)



3.3.3 Evaluación e interpretación de resultados

Identificación de patrones válidos para el objeto de la investigación, interpretación y formalización de los resultados obtenidos que se formulan en el aparte de conclusiones.

3.3.4 Difusión y uso de modelos

Visualización de los resultados y formulación de acciones a seguir en búsqueda del objetivo de investigación.

- A nivel de conclusiones se presenta conclusiones generales enfocadas a la construcción de la propuesta del plan de acción.
- Las acciones están referidas a las estrategias a seguir por parte de la cooperativa para el aprovechamiento de la información resultante del proyecto.
- Las recomendaciones desde el punto de vista de consideraciones de mejora para la dirección y desde el punto de vista de la proyección de nuevos estudios y/o la reutilización y seguimiento del presente proyecto.

4. RESULTADOS Y DISCUSIÓN:

Durante la depuración de los datos se pudo encontrar información estadística importante, que aunque no fueron generadas por las reglas de asociación, si es de importancia para la empresa y que se sintetizan a continuación:

- Se pudo establecer que del total de las facturas para el caso del mes de Septiembre de 2015 el 36,25% (1.937 facturas) corresponden a facturas en las que se compró 1 artículo, y que el 17,82% (952 facturas) corresponden a facturas con 2 artículos comprados, igualmente el 10,2 (547 facturas) corresponde a facturas que registran compras con 3 artículos, consolidando el 64,27% (3.436 facturas) del total (5.342) facturan entre 1 y 3 artículos.
- Para los datos revisados, es decir facturas con 10 o más artículos, se tiene que corresponde al 13,70% (732 facturas).

A partir de los resultados surgen las siguientes recomendaciones:

- Que el personal a cargo de los diferentes procesos conozca y aplique las herramientas y metodología que aporta el proyecto, en las que se puede por ejemplo aplicar para diferentes meses y realizar comparaciones, igualmente un estudio específico de líneas de productos, marcas, proveedores, de clientes, de temporadas, etc...
- Tomar los resultados obtenidos en las reglas de asociación como referente para el momento de la compra de mercancías y para su distribución en estantería.
- A manera experimental realizar el mismo proceso con otras herramientas y métodos para evaluar sus rendimientos.
- Continuar el estudio con la identificación de recursos computacionales eficientes para la manipulación de grandes cantidades de datos.

5. CONCLUSIONES:

- Las herramientas y metodología elegidas y aplicadas para la generación de reglas de asociación rindieron los resultados esperados, con lo que se valida el éxito de las mismas.
- El requerimiento computacional para la ejecución de este tipo de herramientas es muy alto.
- Se pudo establecer que el soporte determina el número de FrequentItemSets o conjunto de ítem frecuentes.
- Igualmente se pudo establecer que a partir de los conjuntos de ítem frecuentes se genera las reglas de asociación y que a mayor Confianza menor es el número de reglas obtenidas.
- Teniendo en cuenta los datos obtenidos en la fase de depuración, se encontró que la mayoría de facturas son de entre 1 y 3 artículos (64,27%) y que las coincidencias de artículos en facturas es baja (Soporte) por la gran variedad de artículos existentes, que de sólo 732 facturas estudiadas correspondieron a 4.398 artículos diferentes, de un total del mes de Septiembre de 2015 de 5.342 facturas que incluyen 5.520 productos diferentes.

- Teniendo en cuenta que las reglas de asociación pueden existir entre cualquiera de los atributos, las reglas resultantes son muchas, cada una de las cuales puede tener una conclusión diferente, lo que conlleva a un tiempo de procesamiento más largo y con mayor costo computacional.
- Con las mediciones de Confianza y Lift, se pudo constatar que la totalidad de las reglas generadas son fuertes y cuentan con un alto grado de credibilidad, presentando una confianza entre el 63 y 91%, o dicho de otra forma existe una probabilidad que las reglas se cumplan entre el 63% al 91%, y que el Lift > 1, entre 5 y 41, que indica que las reglas no son producto del azar y que las relaciones encontradas se presentan con mayor frecuencia de lo esperado.
- Se evidencia un agrupamiento por variedad de los mismos artículos, por ejemplo entre sabores de compotas, yox, yogurt, jugos o pulpas.
- Igualmente se evidencia de forma clara el agrupamiento de artículos relacionados por tipo, por ejemplo la presencia en las reglas de artículos como harinas, panela, granos y pastas.
- Predomina la inclusión de artículos como la panela y la harina pan en las compras de los clientes ya que están presente en varias reglas de asociación.
- Finalmente se puede concluir que a partir de la validación exitosa de las herramientas y metodología escogidas, el presente proyecto se puede convertir en una guía de aplicación y uso cotidiano de apoyo a la toma de decisiones con su implementación en nuevas bases de datos o en nuevos escenarios de información.

6. REFERENCIAS:

- Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proc.Int. Conf. Very Large Database (VLDB'94)*, (págs. 487-499). Santiago, Chile.
- Agrawal, R., Imieliski, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (págs. 207-216). Washington, D.C., USA: ACM.
- Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1991). *Knowledge Discovery in Databases*. Cambridge: MIT Press.

- Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery*, 53–87.
- Hernandez, J., Ramírez, M. J., & Ferri, C. (2008). *Introducción a la minería de datos*. Madrid: Pearson / Prentice Hall.
- Menzies, T., & Hu, Y. (2003). Data Mining for Very Busy People. *IEEE Computer*, 18-25.
- Park, J., Chen, M.-S., & Yu, P. S. (1997). Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Trans. on Knowl. and Data Eng.*, 813-825.
- Rapidminer. (diciembre de 1 de 2016). *Rapidminer*.
Obtenido de <https://rapidminer.com/us/>
- Savasere, A., Omiecinski, E., & Navathe, S. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. *In Proceedings of the 21st International Conference On Very Large Data Bases*, (págs. 432-444). Zurich, Switzerland.
- Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. *KDD'97 Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (págs. 283-286). Newport Beach, CA: AAAI Press.