Trabajo Práctico 1

Nicolás Muschitiello
Roberto Vélez
Jairo Jiménez

July 24, 2015

1 Introducción

En el presente informe, se presenta el resultado de trabajar con el algoritmo apriori [Agrawal and Srikant, 1994] para generar reglas de asociación interesantes sobre una base de datos de ventas de una empresa.

Se estructura en tres secciones: en la primera, se representan los resultados técnicos, que incluyen el detalle del preprocesamiento de los datos, software y algoritmo utilizado, justificación de la elección de los parámetros del mismo y explicación del criterio utilizado para la selección de los resultados no técnicos. En la segunda, aparecen los resultados no técnicos, que consisten en la selección de reglas interesantes, siguiendo distintos enfoques. Por último, se esboza una breve conclusión del trabajo realizado.

2 Resultados técnicos

2.1 Preprocesamiento

En este apartado, describimos cómo preprocesamos las tablas previo a introducirlas en el algoritmo de generación de reglas de asociación:

- Comenzamos aplanando la base usando MS Access, con lo que se obtuvieron las variables presentadas en la tabla 1.
- Removimos de la base aplanada todos aquellos renglones que no tenían transacciones asociadas (entre ellos, 1 Categoría y 16 SubCategorías)
- Eliminamos 363 transacciones con cantidades negativas, ya que representan devoluciones.
- Detectamos la existencia productos con precio 0. Se decidió no hacer nada al respecto, aunque distorsionan las mediciones del gasto de los clientes y la ponderación de reglas por su impacto en las ventas
- Separamos el campo DescAdic, utilizando el siguiente procedimiento:
    1. Realización de un query para sacar la categoría y subcategoría por producto con las descripciones generales y adicionales. Estas últimas son las que vamos a separar en una tabla nueva que contendrá un registro por cada atributo con su valor asociado con la subcategoría. Ver Anexo 1 6

Table 1: Tabla de variables obtenidas

Denominación	**Descripción**	**Tabla de Origen**
Cat_Desc	Descrición de la Categoría a la que pertenece el Producto.	TP_Categoría
cli_CodPos	Código postal del Cliente.	TP_Clientes
cli_Loc	Localidad del Cliente.	TP_Clientes
CLI_NOM	Denominación del Cliente.	TP_Clientes
CLI_Prv	Provincia del Cliente.	TP_Clientes
Precio	Precio sugerido del Producto.	TP_Precio_Sugerido
CantEnvase	Medida de Cantidad.	TP_Productos
Cat_ID	ID de la Caegoría a la que pertenece el Producto.	TP_Productos
DescAdic	Descripción Adicional del Producto.	TP_Productos
DescGen	Descripción Genérica del Producto.	TP_Productos
Marca	Marca del Producto.	TP_Productos
Prod_ID	ID del Producto.	TP_Productos
Proveedor	Proveedor.	TP_Productos
SubCat_ID	ID de la Sub Categoría a la que pertenece el Producto.	TP_Productos
SubCat_Desc	Descripción de la Sub Categoría a la que pertenece el Producto.	TP_Sub_Categoría
CLI_ID	ID del Ciente.	TP_Ventas
SIT_IVA_ID	Situación ante el IVA del Cliente.	TP_Ventas
Venta_Fecha	Fecha de la Venta.	TP_Ventas
Cantidad_UM1	Medida de Cantidad.	TP_Ventas_Prod
Cantidad_UM2	Medida de Cantidad.	TP_Ventas_Prod
Fecha	Fecha de la Venta.	TP_Ventas_Prod
Renglon	Número del renglón de la Venta.	TP_Ventas_Prod
Venta_ID	ID de la Venta.	TP_Ventas_Prod

2. Realización de un script para separar la columna DescAdic, este script contiene la separación de la subcategoría primeramente para después poder separar los atributos que están identificados con la notación:

   Nombre del Atributo: Valor del Atributo

3. Realización de un script (ver Anexo) para generar un .csv con los nuevos registros para la tabla TP_MarcaAtributos en R.
4. Creación de la tabla en Access con las siguientes columnas:
   - marca_ID
   - marca_Nombre
   - marca_AtributoNombre
   - marca_AtributoValor
   - SubCat_ID

Determinamos que casi toda la información contenida en este campo se encuentra codificada dentro de Prod_ID, por lo que decidimos no trabajar a la hora de generar reglas con DescAdic.

- Como medida de volumen, decidimos trabajar con el campo Cantidad UM1
- Clasificamos los clientes siguiendo tres ejes: la cantidad de compras realizadas, el gasto incurrido en las mismas, y la categoría de los productos que compran. Para las dos primeras clasificaciones, utilizamos tres categorías para cada una, a saber:
  - Cantidad de transacciones. Muy Frecuente: $\geq$ 14 transacciones (una compra por mes en promedio o más); Frecuente entre 7 y 13 (una compra cada dos meses o más); Poco Frecuente $<$ 7 (menos de una compra cada dos meses)
  - Gasto. Gasto Alto $\geq$ \$ 500.000. Gasto Medio entre \$ 100.000 y \$ 499.999. Gasto Bajo $<$ \$ 100.000.
  - Adicionalmente, determinamos su rubro (o rubros), a partir de los productos que compran, agrupados por categoría de acuerdo al siguiente esquema:
    * Los clientes que compraron un 50% o más de productos (COUNT DISTINCT Prod_ID & Venta_ID) de una misma categoría, se clasifican con esa categoría
    * A los clientes que no entran en la clasificación anterior, si entre las dos categorías mayoritarias suman más del 80%, se utilizan ambas para su clasificación. Si no, se los clasifica como POLIRUBRO

2.2 Software utilizado

En la elaboración del presente trabajo, se usaron gran variedad de herramientas de análisis de datos. El preprocesamiento de los datos se hizo conjuntamente con las herramientas Microsoft Excel, Microsoft Access y QlikView.

Para la generación y evaluación de las reglas presentadas se exploraron las herramientas *Weka* [Witten and Frank, 2005] y *R* [R Core Team, 2013]. Dada la flexibilidad de programación del software estadístico *R*, se optó por este último.

En dicho software, se implementaron los códigos necesarios para la generación de los conjuntos de datos en el formato requerido por el algoritmo, el cálculo de las medidas de interés adicionales y la poda de las reglas.

2.3 Justificación de la elección de los parámetros del algoritmo

La determinación del minsup se realizó a partir del análisis exploratorio de los datos. La confianza, a partir de iteraciones sucesivas del algoritmo y observar cuántas reglas interesantes generaba. Trabajamos con valores de 0.005 para el min support, y 0.6 para la confianza.

Adicionalmente, al no contar con conocimiento experto sobre la base de datos, se tomo la decisión de emplear otras medidas, las cuales se presentan en el apartado 2.4.1

2.4 Criterio para la selección de los resultados no técnicos

Para el análisis de las reglas interesantes, se decidió utilizar medidas adicionales a las medidas clásicas encontradas en el libro de [Tan et al., 2005].

2.4.1 Medidas adicionales

Las medidas que fueron elegidas tienen la propiedad de ser *null-invariantes*, es decir, no se ven afectadas por el efecto de la falta de la categoría en el conjunto de datos, adicionalmente, son más robustas que otras medidas con la misma propiedad como por ejemplo la confianza máxima o la confianza total. Las medidas utilizadas principalmente son la medida coseno y la medida de Kulczynsky, las cuales miden la correlación entre el antecedente y el consecuente de la regla, siendo 0 correlación negativa y 1 correlación positiva. Cuando el desbalanceo es muy alto entre los datos, la medida de Kulczynsky tiene valores cercanos a 0.5, mientras que la medida de coseno pierde robustez, en cuyo caso, se usa como soporte la medida de Razón de desbalanceo, la cual permite identificar las reglas interesantes como aquellas que tienen este índice cercano a 1 [Hall et al., 2009]. Estas medidas se presentan a continuación

Medida coseno:

$$cosine(A, B) = \sqrt{P(A|B) \times P(B|A)}$$

Medida de Kulczynsky:

$$Kulc(A, B) = \frac{1}{2} \left(P(A|B) + P(B|A)\right)$$

Razón de desbalanceo:

$$IR = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

3 Resultados no técnicos esperados

Para la selección de las reglas en el presente capítulo, se tuvieron en cuenta las medidas descritas en la sección 2.4.1. Se tomaron las reglas que se consideraron interesantes con respecto a dichas medidas.

3.1 Características más habituales de las ventas de la empresa

La mayoría de las reglas interesantes que se presentan en este apartado son de nivel Categoría o Subcategoría, porque entendemos que resumen mejor la actividad de la empresa.

En este sentido observamos que, consistentemente con lo detectado en el análisis exploratorio de los datos, todas las reglas involucran artículos de camping o pesca (al hablar de pesca nos referimos tanto a PESCA, como a PESCA REELS y PESCA CAÑAS), que son los dos rubros más representativos a nivel ventas, tanto si se lo mide a partir de la cantidad de transacciones como del monto.

Como diferencia metodológica con respecto al resto de los puntos de este apartado, se decidió, para algunos casos, prescindir de la medida de kulczinsky, de manera tal que sobrevivan al filtrado algunas reglas que consideramos aportan información relevante. Estas reglas se presentan en la tabla 2.

3.2 Reglas generadas a partir de las variables demográficas

En general las relaciones existentes entre las variables demográficas y los productos, descripción general de los productos, subcategorías y categorías es muy poca. A nivel de producto, las reglas encontradas suelen no ser muy interesantes pues éstas solamente relacionan botellas con botellas o termos con termos, sin embargo, éstas tienen dos particularidades: todas tienen una correlación positiva entre el antecedente y el consecuente y todas están relacionadas con la localidad y la provincia de Buenos Aires.

En cuanto a los demás grupos (descripción general, subcategoría y categoría), las correlaciones encontradas son negativas mostrando una "repelencia" entre los ítems y las ciudades en las cuales se encuentran. Las reglas nombradas, son presentadas en la tabla 3.

3.3 Reglas a nivel de monto y cantidad de ventas de la empresa

En este apartado, se buscó seleccionar reglas interesantes desde un punto de vista económico, ponderando las mismas por la cantidad de ventas que involucran y su monto.

El primer factor se encuentra resumido en el soporte, mientras que para el segundo se calculó, para cada regla, el ingreso que le reportó al negocio en el período analizado la venta de los ítems que la componen.

En cuanto a la selección de las reglas, se le dio prioridad, más allá del criterio explicitado en el párrafo anterior, a aquellas generadas a nivel producto, entendiendo que, sin conocimiento de la empresa, son las que pueden resultar menos obvias o triviales.

Por el mismo motivo (desconocimiento del negocio) presentamos un par de reglas que pueden parecer redundantes, por ejemplo: BASTÓN TREKKING y PEDERNAL PARA ENCENDER EL FUEGO a nivel producto (ERNE =>FSTONE01) y luego una regla que involucra estos productos a nivel de Descripción General. Las reglas son presentadas en la tabla 4.

3.4 Reglas generadas de un año a otro

Las reglas presentadas a continuación fueron tomadas con las siguientes condiciones:

- El periodo de comparación del 2014 elegido (de acuerdo a la información disponible para el 2015) va desde el mes de Marzo hasta Mayo.
- Las reglas generadas e identificadas como interesantes reflejan un periodo de comparación similar entre un año y otro, las del 2014 elegidas luego fueron calculadas a mano en el 2015 para ver el impacto de la misma en ese año.

Como se mencionó en los apartados anteriores para la selección de las reglas se uso la medida Kulczynsky, la cual en ambas tablas tiene valores altos y en el caso de algunas reglas con valor de 0.50 se usó la razón de desbalanceo (IR) para ayudar en la decisión.

En el resultado encontrado de las reglas ***por ID de producto***, "Copo Modelo: LN129B Tipo: EXTENSIBLE" resultó en una compra muy frecuente junto al "Copo Modelo: LN117B Tipo: ALUM.ANODIZADO" durante el 2014 sin embargo en el 2015 para el mismo periodo no tuvo un impacto significativo es decir no fue una compra frecuente.

Por otro lado la compra de Binoculares Teens en 2 colores diferentes (Rojo y Negro) son una compra poco frecuente por las casas de óptica, esta regla en el 2015 tampoco tuvo impacto significativo al ver que el porcentaje de transacciones en el que se encuentra es practicamente cero.

Con respecto a la ***descripción del producto***, se puede apreciar una regla bastante interesante que es la compra de 2 tipos de pala: plegable y pala pico, esta como una compra muy frecuente en el 2014 y que también tiene impacto similar en el 2015 para el mismo periodo.

De la misma manera, las líneas monofilamento en cojunto con los reels tipo charger son poco frecuentes con el reel Bellus, esto en el 2015 se ve que es una regla interesante con una asociación negativa, es decir que un cliente que compre un reel tipo charger con una linea monofilamento es poco probable que compre un reel Bellus.

Para las ***subcategorias*** se destacan un par de reglas interesantes referentes a las cañas de tipo Pejerrey y sus correspondiente accesorio del mismo tipo (reel) junto con anzuelos representan clientes con un gasto medio.

Las bolsas de dormir rectangulares con las de tipo sarcofago se asocian con la compra de linternas por casas de pesca, estas reglas en el 2014 tienen una asociación positiva mientras que en el 2015 ocurre un efecto contrario.

Por último la regla de compra de telescopios reflectores por casas de opticas es muy poco frecuente guarda una asociación negativa en el 2014 mientras que en el 2015 su efecto es de asociación positiva.

En las reglas obtenidas por ***categoria***, vemos que destacan las casas que compran muy frecuente, son las casas de pesca y con gastos altos para ambos años. De la misma manera artículos de camping, pesca, cañas de pescar compradas por las casas de pesca implican compra de reels, así como también artículos de camping, indumentaria, cañas de pescar y pesca ocurren frecuentemente para casas de pesca.

La tabla 5 presenta las 10 reglas interesantes encontradas para el 2014, mientras que la tabla 6 contiene el cálculo de las medidas a mano del 2015 para las mismas reglas que resultaron interesantes en el 2014, ambas tablas ordenadas por la medida Kulczynsky.

Table 2: Reglas más habituales

Regla	**Soporte**	**Confianza**	**Lift**	**Coseno**	**Kulczinsky**	**IR**	**Grupo**
CAMPING,INDUMENTARIA,PESCA,PESCA CAÑAS =>PESCA REELS	0.016	0.831	3.768	0.248	0.452	0.897	Categoría
CAMPING,INDUMENTARIA,PESCA REELS,TIRO Y DEFENSA =>PESCA	0.006	0.946	2.862	0.126	0.481	0.981	Categoría
CAMPING,PESCA CAÑAS =>PESCA	0.072	0.722	2.185	0.396	0.470	0.645	Categoría
CAMPING,INDUMENTARIA,PESCA REELS,TIRO Y DEFENSA =>PESCA CAÑAS	0.005	0.919	3.996	0.147	0.471	0.972	Categoría
CAMPING,PESCA CAnAS =>PESCA REELS	0.077	0.773	3.508	0.519	0.561	0.498	Categoría
INDUMENTARIA,INDUMENTARIA PESCA,PESCA,PESCA REELS =>PESCA CAÑAS	0.008	0.909	3.953	0.177	0.472	0.959	Categoría
PESCA,PESCA CAÑAS =>PESCA REELS	0.103	0.778	3.532	0.604	0.623	0.352	Categoría
INDUMENTARIA PESCA,PESCA REELS,TIRO Y DEFENSA =>PESCA	0.007	0.933	2.824	0.137	0.477	0.977	Categoría
INDUMENTARIA PESCA,PESCA REELS =>PESCA CAÑAS	0.030	0.802	3.487	0.321	0.465	0.814	Categoría
PESCA REELS =>PESCA	0.135	0.613	1.855	0.501	0.511	0.265	Categoría
PESCA REELS =>PESCA CAÑAS	0.168	0.764	3.321	0.748	0.748	0.034	Categoría
CAÑAS VARIADA,CARPAS =>REELS VARIADA	0.007	0.738	5.214	0.193	0.394	0.915	SubCategoría
CAÑAS VARIADA,REELS PEJERREY =>REELS VARIADA	0.045	0.764	5.397	0.491	0.539	0.535	SubCategoría
COCINA OUTDOOR =>JARROS, BOTELLAS Y TERMOS	0.014	0.634	3.913	0.236	0.361	0.819	SubCategoría
ACCESORIOS PESCA,CAnAS PEJERREY,REELS VARIADA =>CAÑAS VARIADA	0.005	0.914	7.400	0.194	0.478	0.951	SubCategoría
CAÑAS VARIADA =>REELS VARIADA	0.077	0.625	4.415	0.584	0.585	0.096	SubCategoría
REELS PEJERREY =>REELS VARIADA	0.077	0.601	4.246	0.571	0.571	0.072	SubCategoría
LiNEA MONOFILAMENTO,SOMBRERO DE ALA =>CAP CON VISERA	0.006	0.851	14.804	0.307	0.481	0.854	Desc. General
ACCESORIOS FLY,CAnAS FLY,LINEAS FLY =>REELS FLY	0.005	0.917	30.221	0.398	0.545	0.799	SubCategoría
CAÑAS FLY,LINEAS FLY =>REELS FLY	0.009	0.763	25.160	0.481	0.533	0.550	SubCategoría
REELS FLY =>CAÑAS FLY	0.019	0.634	18.051	0.589	0.591	0.103	SubCategoría
PASAHILOS =>PUNTERA	0.015	0.612	18.886	0.528	0.534	0.198	Desc. General

Table 3: Reglas con información demográfica

Reglas	**Soporte**	**Confianza**	**Lift**	**Coseno**	**Kulczinsky**	**IR**	**Grupo**
BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,Capital Federal-Prov =>BTP4S79-5SC	0,006	1,000	49,024	0,525	0,638	0,724	Producto
LCM1406NB,Buenos Aires-Prov =>LCM1406NR	0,005	0,944	49,001	0,517	0,614	0,689	Producto
BTP4S79-75BLC,Capital Federal-Prov =>BTP4S79-75RC	0,010	0,878	29,089	0,551	0,612	0,579	Producto
LINEAS FLY,Capital Federal-Prov =>CAnAS FLY	0,006	0,609	17,324	0,329	0,394	0,635	Subcategoría
GENERAL ROCA-Loc =>PESCA	0,008	0,627	1,892	0,126	0,326	0,946	Categoría
Tucuman-Prov =>PESCA	0,009	0,621	1,875	0,133	0,325	0,938	Categoría
RESISTENCIA-Loc =>CAMPING	0,005	0,630	1,135	0,079	0,320	0,979	Categoría
EL PALOMAR-Loc =>PESCA	0,008	0,613	1,849	0,121	0,318	0,947	Categoría
TORTUGUITAS-Loc =>CAMPING	0,005	0,618	1,114	0,078	0,314	0,978	Categoría
MONTE GRANDE-Loc =>CAMPING	0,005	0,607	1,094	0,077	0,308	0,978	Categoría

Table 4: Reglas a nivel monto

Regla	**Soporte**	**Conf.**	**lift**	**Coseno**	**Kulcz.**	**IR**	**Monto**	**Grupo**
CAÑAS VARIADA,NYLON =>REELS VARIADA	0.028	0.726	5.132	0.378	0.461	0.679	$ 19,485,499	SubCategoría
MOCHILAS DSICOVERY,MOCHILAS URBANAS =>MOCHILAS SUPER MOUNTAIN	0.005	0.842	17.915	0.302	0.475	0.854	$ 6,453,944	Desc. General
BINOCULAR ORBITAL,LUPA PROFESIONALES =>LUPA DE MANO	0.007	0.854	23.591	0.392	0.517	0.766	$ 2,693,876	Desc. General
LCM1406NR,TA1001A =>LCM1406NB	0.005	0.943	47.497	0.499	0.603	0.709	$ 1,987,561	Producto
SOMBRERO DE ALA =>CAP CON VISERA	0.026	0.763	13.273	0.583	0.604	0.367	$ 1,895,555	Desc. General
BINOCULAR TRAVEL II,LUPA DE MANO =>LUPA PROFESIONALES	0.007	0.788	20.601	0.366	0.479	0.750	$ 1,210,547	Desc. General
LiNEA FLY SINKING BLACK =>LiNEA FLY FLOATING ORANGE	0.010	0.753	36.200	0.592	0.609	0.331	$ 961,779	Desc. General
BASToN TREKKING,PEDERNAL PARA ENCENDER FUEGO =>MANTA DE EMERGENCIA	0.006	0.809	66.989	0.636	0.654	0.341	$ 852,082	Desc. General
BALLESTA =>ARCO	0.006	0.673	44.127	0.509	0.529	0.360	$ 788,392	Desc. General
LUPA PROFESIONALES,TERMoMETRO =>LUPA DE MANO	0.006	0.830	22.917	0.377	0.500	0.767	$ 596,152	Desc. General
LUPA DE CAMPO,LUPA PROFESIONALES =>LUPA DE MANO	0.007	0.837	23.109	0.388	0.508	0.758	$ 482,369	Desc. General
JACO10 =>NOMA10	0.006	0.632	59.359	0.583	0.584	0.114	$ 307,650	Producto
ERNE =>FSTONE01	0.006	0.648	48.016	0.517	0.530	0.298	$ 285,939	Producto
WASABI10 =>MIRAZUR	0.007	0.776	81.427	0.763	0.763	0.027	$ 278,881	Producto
MIRAZUR =>NOMA10	0.006	0.633	59.524	0.599	0.600	0.079	$ 268,426	Producto
MG1968,TN4X30-1 =>TN4X30-3	0.007	0.811	33.391	0.478	0.546	0.613	$ 261,638	Producto

Table 5: Reglas Interesantes del 2014

Reglas	**S**oporte	**C**onfianza	**L**ift	**C**oseno	**K**ulczinsky	**I**R	**G**rupo
LN129B,Muy Frecuente =>LN117B	0.007	1.0	100.625	0.866	0.875	0.250	Producto
TN4X30-1,Poco Frecuente,CASA DE OPTICA =>TN4X30-4	0.006	1.0	89.444	0.745	0.778	0.444	Producto
PALA PLEGABLE,Muy Frecuente =>PALA PICO	0.010	0.800	32.200	0.566	0.600	0.455	Desc. General
LiNEA MONOFILAMENTO,REEL CHARGER,Poco Frecuente =>REEL BELLUS	0.012	1.000	11.500	0.378	0.571	0.857	Desc. General
ANZUELOS,CAÑAS VARIADA,REELS PEJERREY,Gasto Medio =>CAÑAS PEJERREY	0.006	1.000	21.184	0.363	0.566	0.868	Subcategoria
BOLSAS DORMIR SARCOFAGO,LINTERNAS,CASA DE PESCA =>BOLSAS DORMIR RECTANGULAR	0.007	1.000	9.471	0.266	0.535	0.929	Subcategoria
TELESCOPIOS REFLECTORES,Poco Frecuente =>CASA DE OPTICA	0.010	0.800	5.963	0.243	0.437	0.891	Subcategoria
PESCA REELS,Gasto Alto,CASA DE PESCA =>PESCA CAÑAS	0.088	0.855	2.981	0.513	0.581	0.609	Categoria
PESCA REELS,TIRO Y DEFENSA,CASA DE PESCA =>PESCA	0.021	1.000	2.639	0.236	0.528	0.944	Categoria
CAMPING,INDUMENTARIA,Gasto Medio,CASA DE PESCA =>Frecuente	0.010	1.000	5.476	0.233	0.527	0.946	Categoria

Table 6: Reglas del 2015 con valores calculados a mano en base a las Interesantes del 2014

Reglas	**S**oporte	**C**onfianza	**L**ift	**C**oseno	**K**ulczinsky	**I**R	**G**rupo
LN129B,Muy Frecuente =>LN117B	0.000	N/A	N/A	N/A	N/A	1.000	Producto
TN4X30-1,Poco Frecuente,CASA DE OPTICA =>TN4X30-4	0.000	0.000	N/A	N/A	N/A	1.000	Producto
PALA PLEGABLE,Muy Frecuente =>PALA PICO	0.023	0.676	17.624	0.640	0.641	0.082	Desc. General
LiNEA MONOFILAMENTO,REEL CHARGER,Poco Frecuente =>REEL BELLUS	0.004	0.667	5.893	0.154	0.351	0.930	Desc. General
ANZUELOS,CAÑAS VARIADA,REELS PEJERREY,Gasto Medio =>CAÑAS PEJERREY	0.003	0.333	7.174	0.147	0.199	0.712	Subcategoria
BOLSAS DORMIR SARCOFAGO,LINTERNAS,CASA DE PESCA =>BOLSAS DORMIR RECTANGULAR	0.003	0.429	7.577	0.152	0.241	0.817	Subcategoria
TELESCOPIOS REFLECTORES,Poco Frecuente =>CASA DE OPTICA	0.014	1.000	9.519	0.367	0.567	0.865	Subcategoria
PESCA REELS,Gasto Alto,CASA DE PESCA =>PESCA CAÑAS	0.117	0.859	2.281	0.517	0.585	0.607	Categoria
PESCA REELS,TIRO Y DEFENSA,CASA DE PESCA =>PESCA	0.026	0.722	1.788	0.217	0.394	0.888	Categoria
CAMPING,INDUMENTARIA,Gasto Medio,CASA DE PESCA =>Frecuente	0.006	0.429	2.269	0.117	0.230	0.887	Categoria

4 Bonus

Basado en todas las reglas generadas entre la tabla 2, 3, 5 y 6 pero de manera especial en las siguientes:

- PESCA REELS =>PESCA
- BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,Capital Federal-Prov =>BTP4S79-5SC
- CAÑAS VARIADA,NYLON =>REELS VARIADA
- PESCA REELS,Gasto Alto,CASA DE PESCA =>PESCA CAÑAS
- BOLSAS DORMIR SARCOFAGO,LINTERNAS,CASA DE PESCA =>BOLSAS DORMIR RECTANGULAR

Proponemos una promoción que consista en que:

- Las compras realizadas en artículos de pesca de preferencia cañas de pescar de diferente tipo o enfocado a un tipo en particular (Ej. Pejerrey, Dorado, Fly, etc) se les realice un descuento sobre las mismas por la compra de items como botella de 500 o 750cc de cualquier color (rojo, azul o plateado) o items no tan comunes que se compran con la caña de pescar pero que se dan frecuentes con otros items como las linternas y las bolsas de dormir, todo esto para los clientes de Capital Federal.

  Esto tiene una ventaja de que como se conoce las cañas de pescar tiene un precio alto, se incentiva a comprarlas en conjunto con otros articulos frecuentes para que el negocio tenga más salida y venta en esos otros articulos y el cliente a su vez también gana con la compra de los items en conjunto.

5 Conclusiones

- En un primer análisis exploratorio de los datos encontramos cierto ruido (transacciones con cantidades negativos, productos con precio de referencia cero, etc.) que era necesario trabajarlo previo a poder aplicar un algoritmo para encontrar items frecuentes en las transacciones derivadas de las ventas, esto ratifica que es importante conocer los datos porque se corre el riesgo de encontrar reglas que no muestren la verdadera dinámica del negocio.

- A pesar de que no se conoce todo el negocio en su complejidad, creemos que una gran parte de las reglas encontradas son triviales, sin embargo cabe mencionar que las medidas que utilizamos nos permitieron discrimar de mejor manera entre este tipo de reglas.

- Consideramos que por todo el conocimiento generado entre el análisis exploratorio y las reglas obtenidas el perfil de la mayoria de los clientes apunta a distribuidores mayoristas, por lo que pensamos que esta sea la razon por la cual hallamos muchas reglas triviales.

6 Anexos

El presente trabajo fue realizado utilizando una herramienta de desarrollo colaborativo basado en el control de versiones como lo es github.

En el siguiente enlace pueden encontrarse los recursos utilizados para la elaboración del informe.

La estructura del repositorio está formada por cuatro carpetas:

Análisis: Contiene notas hechas por el grupo para la elaboración del informe, primera version de la base de datos consolidada y una base con la separación de la descripción adicional del producto en una nueva tabla como parte del preprocesamiento.

Insumos: Contiene todos los recursos necesarios (archivos de excel de las diferentes tablas) utilizados para formar los múltiples consolidados que se usaron para generar las reglas.

Resultados: Aquí se puede encontrar la base final consolidada la cual se tomó como entrada el algoritmo que generó las reglas así como también todos los archivos en formato de excel con las reglas para cada una de las secciones del informe.

Sintaxis: Contiene los scripts en R utilizados para generar las reglas y para separar el campo descripción adicional, con la generación del .csv para importarse como nueva tabla al modelo original en Access.

References

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Anexos

Anexo 1

```
SELECT TP_Categoria.Cat_ID, TP_Categoria.Cat_Desc,
   TP_Sub_Categoria.SubCat_ID, TP_Sub_Categoria.
   SubCat_Desc, TP_Productos.DescGen, TP_Productos.
   DescAdic
FROM TP_Sub_Categoria INNER JOIN (TP_Categoria INNER
    JOIN TP_Productos ON TP_Categoria.Cat_ID =
   TP_Productos.Cat_ID) ON TP_Sub_Categoria.
   SubCat_ID = TP_Productos.SubCat_ID;
```