Trabajo Práctico 1

Nicolás Muschitiello
Roberto Vélez
Jairo Jiménez

June 27, 2015

1 Introducción

2 Resultados técnicos esperados

2.1 Preprocesamiento

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

- Comenzamos aplanando la base, usando MS Access
- Removimos de la base aplanada todos aquellos renglones que no tenían transacciones asociadas (entre ellos, 1 Categoría y 16 SubCategorías)
- Eliminamos las transacciones con cantidades negativas, ya que representan devoluciones que totalizan 363
- 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

```
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;
```

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 Anexos) 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 la generación y evaluación de las reglas presentadas en este trabajo se exporaron 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.

Adicionalmente, al no contar con conocimiento experto sobre la base con la que trabajamos, decidimos 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 [Tan et al., 2005]. Con la ayuda de éstas, se determinaron las reglas más interesantes para las categorías, subcategorías, etc.

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. Las medidas utilizadas son la medida coseno, la medida de Kulczynsky y la razón de desbalanceo [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

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 del negocio.

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 sección tercera, 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 1.

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, como por ejemplo el caso del jarro térmico en la provincia de Buenos Aires y el termo. Las reglas nombradas, son presentadas en la tabla 2.

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 las 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 3.

3.4 Reglas generadas de un año a otro

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

- El periodo de comparación fue todo el 2014 con respecto a los meses de marzo a Mayo del 2015
- Las reglas generadas y catalogadas como interesantes reflejan un periodo de comparación bastante diferente entre un año y otro por lo que muchas encontradas en el año 2015 no se cumplían en 2014 y viceversa.

Es muy probable que tomando el mismo periodo de meses del 2015 para el 2014 resulten en mayor cantidad de reglas similares en caso de existir comportamientos de venta en común.

En el resultado encontrado de las reglas generadas en ambos años, por ID de producto las lupas con diferentes aumentos comprados por casas de optica es una caracteristica que sobresale en ambos años, por otro lado la compra de botellas de color rojo y azul de 500cc y 750cc con botellas color plateado de 500cc, esto también se refleja como transacciones muy frecuentes, y finalmente otra regla bastante interesante es la compra de termos de tipo rojo y azul en medidas de 400cc por las casas de pesca.

Con respecto a la descripción del producto, se puede apreciar que la reglas interesantes comprenden compra de binoculares del tipo travel II, lupas de mano y lupas profesionales por casas de optica; de la misma manera sombreros y capas por las casas de pesca. Una regla bastante interesante es la de la compra de bolsas para dormir tipo Gravity que generan un gasto medio y son compradas por casas de camping que siguen un patron de compra bastante frecuente en ambos años.

Por el lado de las subcategorias se destacan un par de reglas interesantes referentes a las cañas de tipo fly y sus correspondientes accesorios del mismo tipo (Líneas y reels) comprados por casas de pesca, mientras que las bolsas de dormir rectangulares siempre tienen un gasto alto y son compras muy frecuentes. Similarmente la regla de compra de binoculares, linternas por casas de opticas indica la compra de lupas en general.

En las reglas obtenidas por categoria, vemos que destacan las compras muy frecuentes (muchas transacciones) x las casas de pesca y con gastos altos para ambos años. De la misma manera articulos de camping, pesca, cañas de pescar compradas por las casas de pesca implican compra de reels y de la misma manera articulos de camping, indumentaria, cañas de pescar y pesca ocurren frecuentemente para casas de pesca.

La tabla 4 refleja lo explicado con respecto a las medidas que se utilizaron para el cálculo de reglas interesantes.

Table 1: 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 2: 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	48,438	0,526	0,638	0,723	Producto
BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,C.A.B.A.-Loc =>BTP4S79-5SC	0,005	1,000	48,438	0,504	0,627	0,746	Producto
BTP4S79-5BLC,BTP4S79-75RC,BTP4S79-75SC,Capital Federal-Prov =>BTP4S79-5SC	0,006	0,972	47,093	0,512	0,621	0,718	Producto
LCM1406NB,Buenos Aires-Prov =>LCM1406NR	0,006	0,947	48,501	0,527	0,620	0,680	Producto
REELS PEJERREY,REELS VARIADA,C.A.B.A.-Loc =>CAÑAS TELESC. PEJERREY	0,010	0,693	6,684	0,254	0,393	0,831	Subcategoría
LINEAS FLY,Capital Federal-Prov =>CAÑAS FLY	0,006	0,609	17,363	0,328	0,393	0,638	Subcategoría
LINEAS FLY,C.A.B.A.-Loc =>CAÑAS FLY	0,006	0,625	17,808	0,315	0,392	0,682	Subcategoría
REEL CHARGER,Buenos Aires-Prov =>REEL BELLUS	0,007	0,652	11,313	0,284	0,388	0,760	Desc. General
BINOCULAR TRAVEL II,C.A.B.A.-Loc =>BINOCULAR ORBITAL	0,008	0,615	9,271	0,266	0,365	0,759	Desc. General
JARRO TÉRMICO,Buenos Aires-Prov =>TERMO	0,009	0,628	6,668	0,239	0,359	0,811	Desc. General
EL PALOMAR-Loc =>PESCA	0,008	0,613	1,853	0,120	0,318	0,947	Categoría
TORTUGUITAS-Loc =>CAMPING	0,005	0,618	1,113	0,078	0,314	0,978	Categoría
MONTE GRANDE-Loc =>CAMPING	0,005	0,607	1,093	0,077	0,308	0,978	Categoría

Table 3: 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 4: Reglas del 2014 reflejadas en el 2015

Reglas	**S**oporte	**C**onfianza	**L**ift	**C**oseno	**K**ulczinsky	**I**R	**G**rupo
MG1970,CASA DE OPTICA =>MG1968	0.008	1	52.485	0.651	0.712	0.576	Producto
BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,Muy Frecuente =>BTP4S79-5SC	0.011	0.950	30.470	0.578	0.651	0.618	Producto
LCM1406NR,CASA DE PESCA =>LCM1406NB	0.007	0.923	47.023	0.571	0.638	0.600	Producto
BTP4S79-5RC,BTP4S79-75RC,CASA DE CAMPING =>BTP4S79-5BLC	0.009	0.833	32.803	0.533	0.587	0.553	Producto
HPR50600,Gasto Bajo,CASA DE OPTICA =>HPR50360	0.007	0.706	30.565	0.460	0.503	0.511	Producto
BINOCULAR TRAVEL II,LUPA DE MANO,CASA DE OPTICA =>LUPA PROFESIONALES	0.008	1	27.492	0.454	0.603	0.794	Desc. General
PASAHILOS,CASA DE PESCA =>PUNTERA	0.019	0.660	15.877	0.550	0.559	0.247	Desc. General
CAJA DE POLIPROPILENO,CAJA TRANSPARENTE =>CASA DE PESCA	0.010	0.900	1.676	0.132	0.460	0.976	Desc. General
SOMBRERO DE ALA,CASA DE PESCA =>CAP CON VISERA	0.013	0.733	9.408	0.346	0.448	0.734	Desc. General
BOLSA DE DORMIR GRAVITY - 230 x 80 x 55,Gasto Medio =>CASA DE CAMPING	0.008	0.636	3.444	0.167	0.340	0.909	Desc. General
CAÑAS FLY,LINEAS FLY,CASA DE PESCA =>REELS FLY	0.016	0.903	20.858	0.581	0.638	0.564	Subcategoria
Gasto Bajo,CASA DE PESCA =>Poco Frecuente	0.074	0.705	3.666	0.523	0.546	0.388	Subcategoria
TELESCOPIOS REFRACTORES,Gasto Bajo =>CASA DE OPTICA	0.022	0.884	9.165	0.448	0.556	0.721	Subcategoria
BOLSAS DORMIR RECTANGULAR,Gasto Alto =>Muy Frecuente	0.028	1	1.605	0.213	0.523	0.955	Subcategoria
BINOCULARES CLASICOS,LINTERNAS,CASA DE OPTICA =>LUPAS	0.008	0.929	13.865	0.323	0.520	0.872	Subcategoria
CAÑAS VARIADA,REELS PEJERREY,Gasto Medio =>CASA DE PESCA	0.028	0.860	1.601	0.213	0.456	0.931	Subcategoria
Muy Frecuente,CASA DE PESCA =>Gasto Alto	0.246	0.742	1.860	0.676	0.679	0.139	Categoria
PESCA,PESCA CAÑAS,CASA DE PESCA =>PESCA REELS	0.143	0.832	2.643	0.614	0.642	0.417	Categoria
CAMPING,PESCA,PESCA CAÑAS,CASA DE PESCA =>PESCA REELS	0.083	0.856	2.721	0.474	0.559	0.664	Categoria
CAMPING,INDUMENTARIA,PESCA CAÑAS,CASA DE PESCA =>PESCA	0.027	0.979	2.359	0.253	0.522	0.932	Categoria
OPTICA,Gasto Alto,CASA POLIRUBRO =>Muy Frecuente	0.009	1	1.605	0.122	0.507	0.985	Categoria

4 Anexos

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

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

La estructura del repositorio esta formada por cuatro carpetas:

Analisis: 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 multiples consolidados que se usaron para generar las reglas.

Resultados: Aqui se puede encontrar la base final consolidada de la cual se tomó como partida para usarse en el algoritmo que genero las reglas asi como también todos los archivos en formato de excel con las reglas para cada uno de los items del informe.

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

References

[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.