Facultad de Ingeniería
Ingeniería Civil Informática

Informe Proyecto Ciencia de Datos
Segunda Entrega

Retail Online

Autores:
Marcelo Guzmán Escobar
Sebastián Herrera Liberona
Pedro Hurtado Lagos
Cristian Meza Cesped
Diego Muñoz Cabrera

Profesor:
Billy Peralta Márquez.

05 de Junio del 2020

June 5, 2020

Contents

1 Introducción

Definición de proyecto

Recordando el hito uno de este proyecto, el objetivo principal de este es encontrar los patrones de compra dentro de un conjunto de datos de ventas online de retail y las distintas relaciones entre los datos de esta. En donde se tenía la siguiente distibución dentro del dataset como se puede apreciar en la siguiente tabla:

invoiceNo	Número de transacción. Si comienza con una C, significa que es una transacción cancelada.
StockCode	Código de identificación para cada producto.
Description	Nombre del producto.
Quantity	Cantidad del producto por transacción.
InvoiceDate	Día y hora en la que la transacción fue efectuada.
UnitPrice	Precio por unidad del producto.
CustomerID	Código de identificación del cliente.
Country	País donde reside el cliente.

Table 1: Distribución dataset

El primer paso para poder realizar los experimentos siguientes es pasar por diferentes etapas, procesos de analisis y preprocesamientos entre los cuales se incluyen: Limpieza de datos/Reducción de datos y eliminación de columnas irrelevantes. Luego se hicieron análisis 1D y 2D utilizando las herramientas como los mapas de calor y gráficos para la distribución de datos.

cabe recordar que los experimentos planteados en la primera fase de este proyecto fueron:

- Encontrar las fechas en que se obtuvo mayor ganancia de dinero. Relacionar Quantity, UnitPrice e InvoiceDate
- Recomendaciones de productos
- Horarios de mayor fluctuación de datos (mayor cantidad de compras)
- Relaciones entre los productos con el uso de algoritmo a-priori (cálculo de confianza, soporte y lift)

Además, dentro de este hito se espera realizar una revisión de trabajos ya existentes los cuales servirán como guía para este proyecto y sus futuros experimentos. A la vez se realizará una descripción sobre los algoritmos a trabajar junto con sus respectivas implementaciones y resultados obtenidos con estos.

2 Revisión de Trabajos

En la lectura de diferentes documentos relacionados con la minería de datos para obtener un mejor planteamiento de los experimentos, algoritmos a utilizar y su implementación respectiva, se logró encontrar documentos clave para la utlización de este segundo hito.

Entre los textos encontrados sobre el tema se logran identificar los siguientes: "Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC" [3] , en el cual logramos extraer algunas técnicas de mineria de datos y con ellas barajar diferentes opciones para avanzar con el proyecto, ver como funcionaba y utilizaban k-means en otros experimentos y ver de que forma logran implementar el Algoritmo Apriori en base al trabajo expuesto en ese documento. En este logramos encontrar anexos sumamente interesantes para el avance del proyecto y mejora del mismo. Dentro de los anexos se logra identificar partes del codigo utilizados para el funcionamiento del programa, como formó los grupos de clientes que posteriormente serían utilizados en k-means y la utilización del algortimo A-priori para un experimento sobre las transacciones de cada grupo de lealtad de clientes. En este estudio se aplicó la metodología CRISP-DM para el preprocesamiento. Luego, para el análisis, se baso en el modelo RFM que según explican es Recencia, Frecuencia, Valor Monetario y bajo este se utilizó el algoritmo de agrupamiento k-means. Finalmente, como ya se mencionó, se utilizó el algoritmo apriori para encontrar asociaciones entre los diferentes productos para cada grupo de clientes. Para todo este trabajo la herramienta que utilizaron fue RStudio.

Luego encontramos una tesis de grado del tema "Obtención del perfil de un cliente fiel en una tienda departamental mediante el diseño de un data warehouse y arboles de decisión" [1] en esta logramos ver otro tipo de alternativas para lograr una solución al problema que tenemos planteado sobre el retail online, con este vemos las ventajas y desventajas de los arboles de decisiones para tomar una mejor decisión con respecto a que algoritmo se utilizará más adelante y cuales son los resultados con este, fue de mucha ayuda ya que nos dio una mejor visión sobre en que algoritmo enfocarse segun nuestro dataset. Este trabajo utiliza la mineria de datos de un data warehouse con esto establencen un árbol de decisión como el método central para lograr determinar los perfiles de un cliente fiel en una tienda departamental. En este se encuentra un capitulo en el qué se habla profundamente sobre la metodología árboles de decisión, dejando en claro sus ventajas y desventajas a la hora de ser utilizado. Nos dejan claro como comprender este diseño, la arquitectura que tiene que tener el árbol, el modelamiento planteado y análisis para tener una mejor toma de decisiones y así cumplir el objetivo principal que es determinar el perfil del cliente fiel.

También se analizó "Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales" [4] , si bien en este estudio se utiliza visual studio, se utilizó más que

nada para ver los experimentos que implementaban y resultados que conseguian junto a su modelo. Este proyecto busca encontrar reglas que logren determinar un patrón de consumo de clientes para una distribuidora de suplementos nutricionales, en donde utilizaron tecnicas de mineria de datos bajo el software ya nombrado Visual Studio 2015 y con una base de datos en SQL server 2014. En este las tablas mostradas en resultados fueron utiles para lograr generar una idea con respecto a que reglas de asociación lograr utilizar (obviamente estas estan enfocadas en el rumbo que este experimento tomo).

Finalmente se leyó "Un sistema de recomendación basado en perfiles generados por agrupamiento y asociaciones" [2] con este logramos llegar más a fondo con el algoritmo apriori y reglas de asosiación pertinentes para estos casos, también logramos ver experimentos relacionados con el algoritmo apriori, que será él que ocuparemos para poder seguir avanzando con este trabajo. En el nos habla de todo lo que tenemos que tener en cuenta para poder llegar a unos resultados esperados, desde que tecnicas de mineria se van a implementar que en el caso del trabajo leído fue con el algoritmo k-means, luego las reglas de asociación pertimentes que se llevan a cabo con el algoritmo apriori, por ultimo vemos los experimentos hecho en este trabajo el cual fue primeramente omitir una elección conocida (en este trabajo utilizaron dos dataset, uno llamado LastFm el cual contiene datos sobre preferencia de artistas entre diferentes usuarios y MovieLens la cual contiene clasificaciones y etiquetas a diferentes peliculas por distintos usuarios) y como segundo experimento ven el recomendar un nuevo ítem.

Gracias a la lectura de estos documentos se logra crear una idea más clara con la cual poder trabajar con un algoritmo específico, viendo todas las ventajas y desventajas que estos nos presentarán durante el trabajo.

3 Descripción de algoritmos

Bajo la lectura y analisis ya hecho previamente, se cree pertinente que el algoritmo óptimo a utilizar en este caso es el algoritmo apriori. Este algoritmo trabaja en generar primeramente conjuntos de items frecuentes y, luego de esto, son ordenados para generar las reglas que se utilizarán. Dentro de este una de los principios fundamentales es que cada subconjunto de un conjunto frencuente que tenga el dataset tambien será frecuente.
Este algoritmo necesita pasar muchas veces por el dataset, ya que busca todos los conjuntos frecuentes, contando sus ocurrencias. En términos de coste computacional es costoso cuando se procesa una gran cantidad de datos, debido a que el número de subconjunto frecuente en cada candidato es cada vez mayor, así irá incrementando los niveles del candidato. El tiempo de ejecución de este tambien es alto debido a lo que se mencionó anteriormente en el coste computacional. Finalmente en rendimiento es donde este algoritmo llega a mejorar notoriamente, ya que reduce el número de items que contienen subconjuntos infrecuentes.

Este algoritmo, basado en el mínimo soporte y confianza definidos por nosotros en la búsqueda de los ítems frecuentes, empieza considerando todo el dataset como candidatos al conjunto de elementos frecuentes, y entonces enumera la ocurrencia de cada elemento, y agrupa los que superan el mínimo soporte y confianza definidos. El paso siguiente se basa en el conjunto formado en el paso anterior y lo adopta como conjunto de candidatos que combina entre sí buscando conjuntos frecuentes de 2 elementos que superan los umbrales mínimos. De forma iterativa, esto se repite hasta lograr obtener un solo grupo.

4 Implementación de Algoritmos

A continuación se mostrará como se implementó el algoritmo elegido (Apriori) para este dataset. Para su correcto funcionamiento se debe instalar la librería "apyori" previamente.

El fin de utilizar el algoritmo Apriori, es poder obtener las distintas reglas de asociasión posibles al ir modificiando los parametros de soporte y confianza, luego, con estas reglas se sugiere al usuario una oferta diaria dentro del top 20 de los itemsets más frecuentes.

```
In [8]: a_priori(df_bd_preprocesada,0.01,0.2)

        Soporte:   0.01

        Confianza:   0.2


        Aplicando algoritmo A-priori...


        ---------------top 20 reglas  -----------------
        frozenset({'FANCY FONT BIRTHDAY CARD', ' '})
        frozenset({'KEY FOB ', ' BACK DOOR '})
        frozenset({'KEY FOB ', ' SHED'})
        frozenset({'60 CAKE CASES DOLLY GIRL DESIGN', 'PACK OF 72 RETROSPOT CAKE CASES'})
        frozenset({'60 CAKE CASES VINTAGE CHRISTMAS', 'SET OF 20 VINTAGE CHRISTMAS NAPKINS'})
        frozenset({'72 SWEETHEART FAIRY CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
        frozenset({'PACK OF 60 DINOSAUR CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
        frozenset({'PACK OF 60 PINK PAISLEY CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
        frozenset({'PACK OF 72 RETROSPOT CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
        frozenset({'PACK OF 60 DINOSAUR CAKE CASES', '72 SWEETHEART FAIRY CAKE CASES'})
        frozenset({'PACK OF 72 RETROSPOT CAKE CASES', '72 SWEETHEART FAIRY CAKE CASES'})
        frozenset({'ALARM CLOCK BAKELIKE CHOCOLATE', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'ALARM CLOCK BAKELIKE CHOCOLATE', 'ALARM CLOCK BAKELIKE RED '})
        frozenset({'ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE IVORY'})
        frozenset({'ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE ORANGE'})
        frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE IVORY'})
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE IVORY'})
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE ORANGE'})
        frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED '})

        ---------------Oferta Diaria ------------------------

         la oferta de hoy es  ['60 CAKE CASES DOLLY GIRL DESIGN', 'PACK OF 72 RETROSPOT CAKE CASES']

        ---------------Oferta Diaria ------------------------

        ////////////////////////////////////////////
        Ejecución del Algoritmo A-priori terminada.
        ////////////////////////////////////////////
```

Figure 1: Apriori, Soporte=0.01 - Confianza=0.2

Luego de partir con un soporte mínimo de 0.01 y una confianza de 0.2, se puede observar que existen distintos itemset y los objetos que más se destacan son "ALARM CLOCK BAKELIKE" en distintas variedades de color. Para este caso, la oferta diaria es "60 CAKE CASES DOLLY GIRL DESIGN" y "PACK OF 22 RETROSPOT CAKE CASES" , el cual se ubicaba en la cuarta posición del top 20.

```
In [11]: a_priori(df_bd_preprocesada,0.01,0.4)

         Soporte:   0.01

         Confianza:   0.4


         Aplicando algoritmo A-priori...


         ---------------top 20 reglas  -----------------
         frozenset({'FANCY FONT BIRTHDAY CARD', ' '})
         frozenset({'KEY FOB ', ' BACK DOOR '})
         frozenset({'KEY FOB ', ' SHED'})
         frozenset({'60 CAKE CASES DOLLY GIRL DESIGN', 'PACK OF 72 RETROSPOT CAKE CASES'})
         frozenset({'60 CAKE CASES VINTAGE CHRISTMAS', 'SET OF 20 VINTAGE CHRISTMAS NAPKINS'})
         frozenset({'72 SWEETHEART FAIRY CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
         frozenset({'PACK OF 60 DINOSAUR CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
         frozenset({'PACK OF 60 PINK PAISLEY CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
         frozenset({'PACK OF 72 RETROSPOT CAKE CASES', '60 TEATIME FAIRY CAKE CASES'})
         frozenset({'PACK OF 72 RETROSPOT CAKE CASES', '72 SWEETHEART FAIRY CAKE CASES'})
         frozenset({'ALARM CLOCK BAKELIKE CHOCOLATE', 'ALARM CLOCK BAKELIKE GREEN'})
         frozenset({'ALARM CLOCK BAKELIKE CHOCOLATE', 'ALARM CLOCK BAKELIKE RED '})
         frozenset({'ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE IVORY'})
         frozenset({'ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE ORANGE'})
         frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE GREEN'})
         frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE GREEN'})
         frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE IVORY'})
         frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE IVORY'})
         frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE ORANGE'})
         frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED '})
         frozenset({'PAINTED METAL PEARS ASSORTED', 'ASSORTED COLOUR BIRD ORNAMENT'})

         ----------------Oferta Diaria -------------------------

          la oferta de hoy es  ['ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE ORANGE']

         ----------------Oferta Diaria -------------------------

         ////////////////////////////////////////////
         Ejecución del Algoritmo A-priori terminada.
         ////////////////////////////////////////////
```

Figure 2: Apriori, Soporte=0.01 - Confianza=0.4

Luego, se decidió aumentar la confianza mínima para la ejecución del algoritmo, en este caso, la confianza fue de 0.4 y se mantuvo el mismo soporte de 0.01. No hubo un mayor cambio dentro del top 20, salvo en la última posición, donde se cambió [”ALARM CLOCK BAKELIKE PINK”,”ALARM CLOCK BAKELIKE RED”] por [”PAINTED METAL PEARS ASSORTED” , ”ASSORTED COLOUR BIRD ORNAMENT”], deduciendo así que el primero no cumplía con la confianza mínima solicitada. En este caso la oferta diaria fue ”ALARM CLOCK BAKELIKE GREEN” y ”ALARM CLOCK BAKELIKE ORANGE”

```
In [2]: a_priori(df_bd_preprocesada,0.01,0.6)

        Soporte:   0.01

        Confianza:   0.6


        Aplicando algoritmo A-priori...


        ---------------top 20 reglas  -----------------
        frozenset({'FANCY FONT BIRTHDAY CARD', ' '})
        frozenset({'KEY FOB ', ' BACK DOOR '})
        frozenset({'KEY FOB ', ' SHED'})
        frozenset({'ALARM CLOCK BAKELIKE CHOCOLATE', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'ALARM CLOCK BAKELIKE CHOCOLATE', 'ALARM CLOCK BAKELIKE RED '})
        frozenset({'ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE ORANGE'})
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE IVORY'})
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE ORANGE'})
        frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED '})
        frozenset({'PAINTED METAL PEARS ASSORTED', 'ASSORTED COLOUR BIRD ORNAMENT'})
        frozenset({'BAKING SET 9 PIECE RETROSPOT ', 'BAKING SET SPACEBOY DESIGN'})
        frozenset({'TOILET METAL SIGN', 'BATHROOM METAL SIGN'})
        frozenset({'BLUE HAPPY BIRTHDAY BUNTING', 'PINK HAPPY BIRTHDAY BUNTING'})
        frozenset({'RED STRIPE CERAMIC DRAWER KNOB', 'BLUE STRIPE CERAMIC DRAWER KNOB'})
        frozenset({'WHITE HANGING HEART T-LIGHT HOLDER', 'CANDLEHOLDER PINK HANGING HEART'})
        frozenset({'CHARLOTTE BAG PINK POLKADOT', 'RED RETROSPOT CHARLOTTE BAG'})
        frozenset({'CHILDRENS CUTLERY DOLLY GIRL ', 'CHILDRENS CUTLERY SPACEBOY '})
        frozenset({'CHRISTMAS CRAFT WHITE FAIRY ', 'CHRISTMAS CRAFT LITTLE FRIENDS'})
        frozenset({'SET 3 RETROSPOT TEA', 'COFFEE'})
        frozenset({'SUGAR', 'COFFEE'})

        ----------------Oferta Diaria ------------------------

         la oferta de hoy es  ['CHILDRENS CUTLERY DOLLY GIRL ', 'CHILDRENS CUTLERY SPACEBOY ']

        ----------------Oferta Diaria ------------------------

        ////////////////////////////////////////
        Ejecución del Algoritmo A-priori terminada.
        ////////////////////////////////////////
```

Figure 3: Apriori, Soporte=0.1 - Confianza=0.6

En una tercera variación del experimento, con un cambio en la confianza a 0.6 pero con el mismo soporte, se observa una diferencia considerable en el top 20 de regla, manteniendose los puestos más altos pero con una variación en las posiciones inferiores, más especificamente desde el puesto 12 en adelante. Esto indica que gran parte de las reglas no superó una confianza mínima de 0.6 establecida. De acuerdo a lo anterior, para el siguiente ejercicio se decidió variar el soporte en busca de nuevos cambios.

```
In [3]: a_priori(df_bd_preprocesada,0.02,0.6)

        Soporte:   0.02

        Confianza:   0.6


        Aplicando algoritmo A-priori...


        ---------------top 20 reglas  -----------------
        frozenset({'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED '})
        frozenset({'DOLLY GIRL LUNCH BOX', 'SPACEBOY LUNCH BOX '})
        frozenset({'GARDENERS KNEELING PAD KEEP CALM ', 'GARDENERS KNEELING PAD CUP OF TEA '})
        frozenset({'PINK REGENCY TEACUP AND SAUCER', 'GREEN REGENCY TEACUP AND SAUCER'})
        frozenset({'GREEN REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER '})
        frozenset({'JUMBO BAG RED RETROSPOT', 'JUMBO BAG PINK POLKADOT'})
        frozenset({'JUMBO BAG RED RETROSPOT', 'JUMBO BAG STRAWBERRY'})
        frozenset({"PAPER CHAIN KIT 50'S CHRISTMAS ", 'PAPER CHAIN KIT VINTAGE CHRISTMAS'})
        frozenset({'PINK REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER '})
        frozenset({'RED HANGING HEART T-LIGHT HOLDER', 'WHITE HANGING HEART T-LIGHT HOLDER'})
        frozenset({'', 'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE GREEN'})
        frozenset({'', 'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED '})
        frozenset({'', 'DOLLY GIRL LUNCH BOX', 'SPACEBOY LUNCH BOX '})
        frozenset({'', 'GARDENERS KNEELING PAD KEEP CALM ', 'GARDENERS KNEELING PAD CUP OF TEA '})
        frozenset({'', 'PINK REGENCY TEACUP AND SAUCER', 'GREEN REGENCY TEACUP AND SAUCER'})
        frozenset({'', 'GREEN REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER '})
        frozenset({'', 'JUMBO BAG RED RETROSPOT', 'JUMBO BAG PINK POLKADOT'})
        frozenset({'', 'JUMBO BAG RED RETROSPOT', 'JUMBO BAG STRAWBERRY'})
        frozenset({'', "PAPER CHAIN KIT 50'S CHRISTMAS ", 'PAPER CHAIN KIT VINTAGE CHRISTMAS'})
        frozenset({'', 'PINK REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER '})

        ----------------Oferta Diaria ------------------------

         la oferta de hoy es  ['DOLLY GIRL LUNCH BOX', 'SPACEBOY LUNCH BOX ']

        ----------------Oferta Diaria ------------------------

        ////////////////////////////////////////////
        Ejecución del Algoritmo A-priori terminada.
        ////////////////////////////////////////////
```

Figure 4: Apriori, Soporte=0.02 - Confianza=0.6

Finalmente, al aplicar un cambio en el soporte mínimo de 0.01 0.02 se obtiene una variación total en el top 20 de reglas de asociación, manteniendo solo dos de estas. De lo anterior se puede deducir que las antiguas reglas nisiquiera cumplen con el soporte mínimo. Estos conjuntos presentados en este top 20 demuestra que al menos están presente dentro de un 2% de las compras realidas dentro del dataset.

5 Otros experimentos aplicados

Además del algoritmo Apriori utilizado para obtener los dataset más frecuentes se realizaron distintos experimentos. A los planteados en la primera entrega del proyecto se sumaron diferentes ideas de experimentos que se llevaron a cabo, con el fin de poder obtener resultados que aporten al entendimiento del proyecto.

Los nuevos experimentos agregados fueron:

- Saber cuántos artículos en promedio se compra en cada transacción

- Promedio de libras gastadas por compra
- Dividir grupos de clientes según dinero gastado(premium y normal)
- intervalos en hora de menor y mayor número de transacciones

a continuación se mostrarán los resultados de estos experimentos.

5.1 Fechas de mayor ganancia

Este experimento fue planteado en la primera parte del proyecto, y para esto se creó una nueva columna llamada "Monto", la cual nace a raíz de la multiplicación de la columna "Quantity" y "UnitPrice", obteniendo así la cantidad de libras gastadas en cada compra. Luego de esto se compara con las fechas agrupadas originalmente por meses para obtener los meses de mayor ganancia en libras.

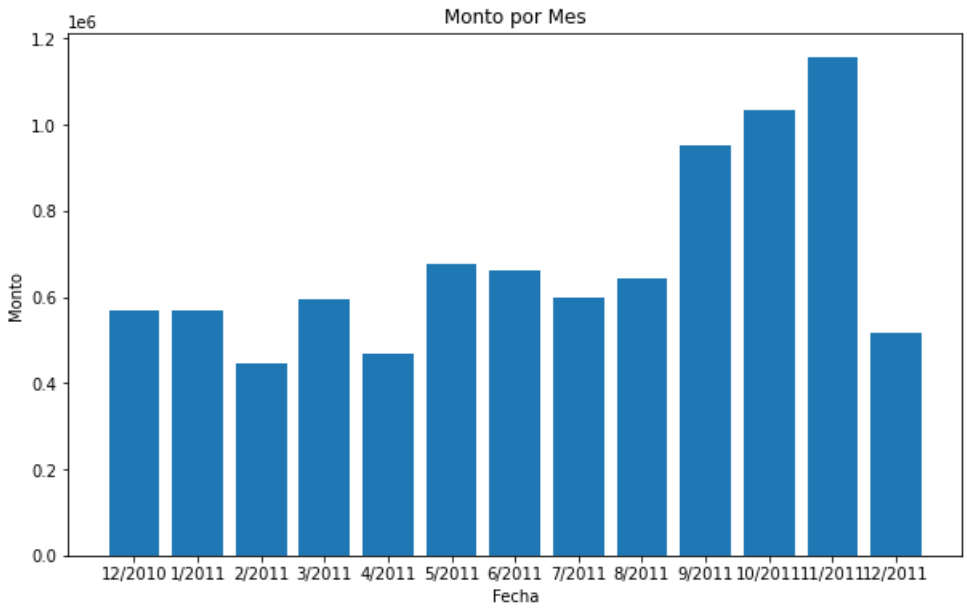

Figure 5: gráfico de compras por mes

5.2 Recomendaciones de productos

Luego de ejecutar el algoritmo Apriori se obtiene los itemset más frecuentes y dentro de un top 20 de distintos itemset frecuentes se entregarán distintas

recomendaciones de productos como la "oferta del día". Este resultado va variando de forma ramdon en cada ejecución, todo esto dentro de los itemset frecuentes. Para ambos ejemplos se utilizó un soporte = 0.01 y una confianza = 0.2

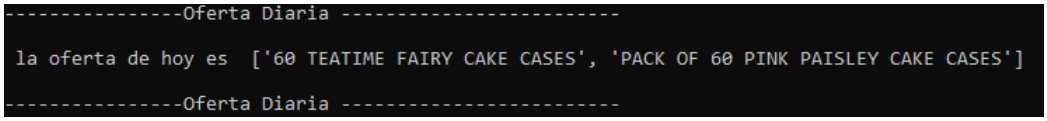

Figure 6: Oferta diaria ejemplo 1

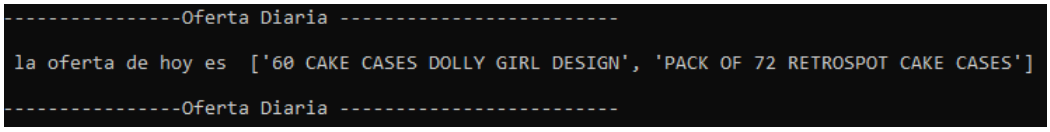

Figure 7: Oferta diaria ejemplo 2

5.3 Horario de mayor fluctuación

La columna "InvoiceDate" se divide en fecha y hora para luego agrupar por hora, buscando así graficar el tramo del día donde hay un mayor flujo de ventas

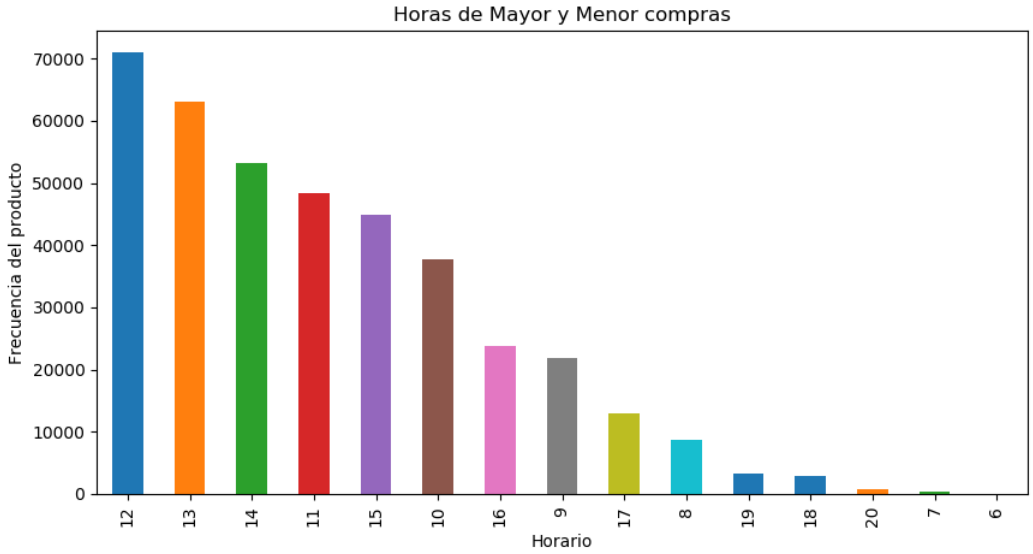

Figure 8: Horario de compras

Como se puede apreciar, la hora donde más compras se realizan es entre las 12 y 13 horas, y el tramo donde hay un menor flujo es entre las 6 y 7 horas

5.4 Artículos y Libras promedio por transacción

Para este experimento se opto por realizar una media recortada, ya que los valores no eran significativos. Como ejemplo, la media en artículos superaba los 300 y en libras gastadas eran más de 4000 en promedio. Estas cifras no resultan representativas del promedio, ya que la media fue recortada a un 20% mediante el uso de la librería stats de scipy. De lo anterior, se puede deducir que ese porcentaje eleva bastante el resultado final y puede llevarnos a interpretaciones erroneas de los resultados, siendo perjudicial para el resultado del experimento

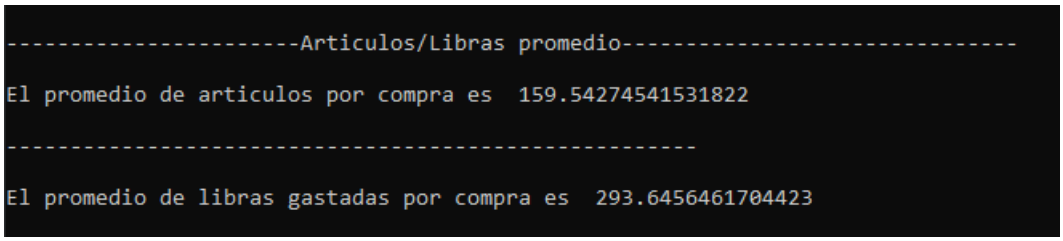

Figure 9: Artículos y Libras promedio

5.5 Grupos de clientes

Se decidió como experimentos agrupar los datos según los id de los clientes dados en la columna "CustomerID" y tomar todas las transacciones de los usuarios en el tramo de fechas compredidos por el dataset para comparar el dinero gastado por cada uno. Luego de esto, quienes superaran la media de gasto se ubicaban como clientes premium y quienes no como clientes normales

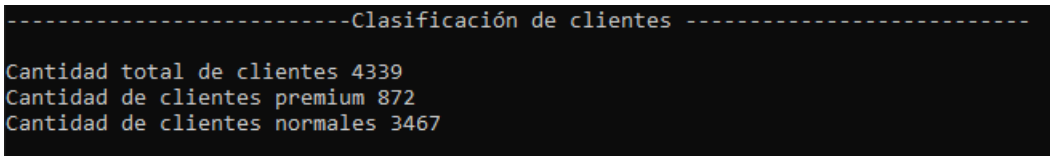

Figure 10: Grupos de clientes

De lo anterior se obtuvo un total de 4339 clientes, los cuales se separaron en 872 Premium y 3467 Normales.

References

[1] Victor Rolando Valencia Valverde Daniel Santiago Aguirre Morocho, Juan Carlos Gallo Ante. *Obtención del perfil de un cliente fiel en una tienda departamental mediante el diseño de un data warehouse y arboles de decisión.* 2010.

[2] Enio Walid Ghobar. *Un sistema de recomendación basado en perfiles generados por agrupamiento y asociaciones.* 2016-2017.

[3] Sairy Fernanda Chamba Jiménez. *Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC.* 2015.

[4] Miguel Angel Grández Márquez. *Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales.* 2017.