UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE RECOMENDACIÓN DE PRODUCTOS APLICADO A UNA EMPRESA DE CUPONES ONLINE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DIEGO ILAN FLORÁS MOSES

PROFESOR GUÍA:
LUIS ABURTO LAFOURCADE

MIEMBROS DE LA COMISIÓN:
TODD PEZZUTI LLOYD
ANDRES MUSALEM SAID

SANTIAGO DE CHILE

2016

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: DIEGO ILAN FLORÁS MOSES
FECHA: 04/03/2016
PROFESOR GUÍA: LUIS ABURTO LAFOURCADE

MODELO DE RECOMENDACIÓN DE PRODUCTOS APLICADO A UNA EMPRESA DE CUPONES ONLINE

En la actualidad, los *e-commerce* han ido evolucionando su forma de atraer a nuevos clientes, y por sobre todo, su forma de fidelizar los que ya han comprado alguna vez. Sin embargo, esta última tarea no ha sido fácil, en especial para sitios con una alta rotación de productos, como los son los sitios de cupones *online*.

Esta memoria propone y genera un prototipo de sistema de recomendación, que permite a los clientes conocer, mediante correos electrónicos, las ofertas que más se ajustan a sus necesidades, y lograr así mejorar los ingresos de la empresa.

Se generaron dos modelos principales para la recomendación, uno basado en filtros colaborativos sobre la base de las preferencias de los usuarios y otro basado en reglas de asociación. El primero consiste en encontrar usuarios con similar comportamiento histórico a cada uno de acuerdo con la información transaccional, para luego recomendarles productos adquiridos por ellos; y el segundo compara las canastas de compras de todos los clientes para ver qué productos o categorías deberían comprarse al haber realizado otra compra.

Los resultados de la experimentación arrojaron que la Tasa de Clics (porcentaje de clics de la cantidad de correos abiertos) y la Tasa de Conversión (porcentaje de compras del total de clics) aumentaría utilizando un modelo de recomendación con un 95% de confianza. El modelo que mejores resultados obtuvo, es el de Reglas de Asociación con Productos Frecuentes, donde aplicando sus cifras a toda la base de datos de la empresa, el correo tendría un tráfico diario adicional de aproximadamente 4.500 clientes, 1.000 adicionales en su sitio *web* y su Tasa de Conversión aumentaría en un 40% en promedio. Además, con la data descriptiva obtenida en los resultados, con este modelo se incrementarían las ventas en casi un 70%.

Dado lo anterior, es recomendable que la empresa implemente un modelo de recomendación en su Newsletter basado en Reglas de Asociación con Productos Frecuentes, pues así, es posible que mejore sus rendimientos económicos como se ha probado en este proceso de experimentación.

Finalmente, a futuro se puede utilizar este mismo modelo para ordenar las ofertas del sitio web de la empresa de acuerdo a las recomendaciones generadas para cada uno de los clientes, y así mismo, se pueden generar recomendaciones asociadas a cada uno de los productos ofertados cuando éstos son visitados.

Para mi familia por siempre creer en mi,

"No creo que las cosas cambien por sí solas,
las tienes que hacer cambiar y
yo voy a hacer lo posible por cambiar"

Rafael Nadal

AGRADECIMIENTOS

Agradezco a mi padre Víctor Hugo Florás y a mi madre Claudia Moses, que a pesar de estar lejos físicamente en la actualidad, han estado conmigo más cerca que nunca. Me han brindado el más grande de los apoyos en todos los momentos, buenos o difíciles. Gracias por sus consejos y sus increíbles charlas motivadoras que a ratos tanto necesité. Gracias por creer en mí en todo momento.

También agradezco a mis hermanas Tamara Florás y Natalia Florás por ser una compañía inquebrantable en todos estos años. Gracias por todas las largas conversaciones que hemos tenido y por todas sus locuras y rarezas que tanto me hacen reír.

Gracias a todos mis amigos por mantenerme firme y por motivarme a no rendirme nunca. A pesar de que algunos intentan llevarme al "lado oscuro", siempre están preocupados por mí, y no me arrepiento de tenerlos como amigos.

No puedo dejar de lado a mi Profesor Guía de este proyecto, Luis Aburto, el cual ha tenido una paciencia insuperable al aceptarme casi todas las semanas con todas mis dudas e inquietudes. También quiero agradecer a mi Profesor Co-guía Todd Pezzuti, que también tuvo una disposición absoluta en recibirme y en contestar cada una de mis dudas en forma rápida y completa.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

1.1. Antecedentes Generales

En el último tiempo, los *e-commerce* han ido evolucionando en su forma de atraer a los clientes. Sin embargo, no siempre las formas adoptadas por estas empresas son las correctas para hacer que éstos sean inducidos a hacer una nueva compra.

En general, un cliente en particular hace una compra cuando encuentra un producto o un servicio que le genera valor, por lo que la forma correcta de hacer que un cliente compre o vuelva a comprar en un *e-commerce*, es ofreciéndole productos o servicios de acuerdo con sus gustos y requerimientos, más aún cuando el comportamiento de compra de los clientes es de forma impulsiva (compra de cupones).

Muchas empresas *online*, han adoptado distintos métodos de personalización para ofrecer sus productos y servicios. Uno de los principales ejemplo existentes es *Amazon*, el cual envía recomendaciones de sus productos a cada cliente de acuerdo con las necesidades de cada uno, utilizando un sistema de recomendación con filtros colaborativos basados en los productos (*ítem-based* en inglés) [1], donde mediante un algoritmo se determinan cuáles productos son similares entre sí, tomando en cuenta las compras y calificaciones de otros clientes. Más adelante en el informe, se detallarán los principales métodos existentes, haciendo una comparación de cuándo y cómo se debe usar cada uno de ellos.

En este caso, se desea hacer un prototipo de sistema de recomendación para una empresa de cupones *online*, como lo es *Cuponatic*. Un sitio como éste, difiere radicalmente de un modelo de negocios como *Amazon* o *Netflix*, ya que no posee la misma cantidad de productos de una misma categoría, ni tampoco posee toda la información que se requiere para poder basarse en las características de los productos para hacer una recomendación, por lo que no se pueden utilizar exactamente los mismos métodos para hacer recomendaciones personalizadas que en los sitios mencionados.

1.2. Descripción y Justificación del Proyecto

En la actualidad, *Cuponatic* envía dos correos electrónicos diarios con descuentos a sus clientes, donde el contenido de éstos son las ofertas más destacadas del día, en el primer correo, y el evento especial del día, en el segundo (por ejemplo, artículos de bebé). El primero se envía a las 08:00 hrs. y el segundo a las 12:00 hrs. aproximadamente.

Al hacer esto, no se toman en cuenta los distintos perfiles de cada cliente, ya que los correos que se envían son los mismos para todos, y sólo varían según su género (masculino, femenino y *default*). Por ejemplo, si a una persona sólo le agrada la música, pero en los productos destacados no hay alguno relacionado con esto, no habrá una respuesta positiva por parte del cliente. Sin embargo, si dentro de ese correo existieran

ofertas de productos y servicios que sean del interés de éste (música en este caso), sí se lo desearía y, por lo tanto, accedería al sitio *web*.

La distribución actual de los correos se pueden ver en la Tabla 1.1. y Tabla 1.2., donde queda en evidencia que existe un formato predeterminado para los correos, donde no se toma en cuenta qué ofertas se debieran ofrecer más que otras.

Categoría	Porcentaje del correo
Destacados	30%
Productos	25%
Health&Beauty	18%
Viajes	18%
Otros	9%
Entretención	0%
Gastronomía	0%

Tabla 1.1: Distribución de *Newsletter* masculino

Categoría	Porcentaje del correo
Destacados	30%
Health&Beauty	25%
Productos	18%
Viajes	18%
Otros	9%
Entretención	0%
Gastronomía	0%

Tabla 1.2: Distribución de *Newsletter* femenino y *default*

Por otro lado, existen clientes que son susceptibles a comprar todo tipo de categorías, como hay algunos que simplemente compran sólo una, por lo que tener una distribución predeterminada para el correo, imposibilita a muchos clientes ver todos los descuentos de su interés.

En la Tabla 1.1 y Tabla 1.2, la distribución del *Newsletter* está predeterminada y se quiere determinar si esta está relacionada con las cantidad de ventas por categorías. Para esto se muestran las ilustraciones 1.1-3, donde se exhiben las categorías que engloban el

78% de las ventas en promedio de los clientes suscritos a las listas *Women* y *Default*, y el 71% de las ventas en promedio de los clientes suscritos a la lista *Men*.

Analizando la Ilustración 1.1 que corresponde a la lista *Women*, se puede observar que la categoría *Health & Beauty* engloba en promedio un 25% de las ventas, lo que coincide totalmente con la distribución predeterminada expuesta en la Tabla 1.2. No obstante, la desviación estándar es de un 17,4%, siendo esto una gran diferencia. Por otro lado, la categoría *Otros* posee ventas en promedio de un 14% del total, acaparando un 5% más que los valores predeterminados, que es un 9%. Su desviación estándar es de 19,6%, lo que se puede ver fácilmente en el gráfico, variando sus ventas desde un 1% hasta un 53%. Finalmente, la categoría de *Productos* tiene un promedio de ventas de un 39%, más del doble del valor predeterminado en el *Newsletter*, que es un 18%. Su desviación estándar es de 16,7%, lo que al igual que las categorías anteriores varían demasiado mes a mes. Para ver el análisis de las listas *Men* y *Default*, ver Anexo 1.

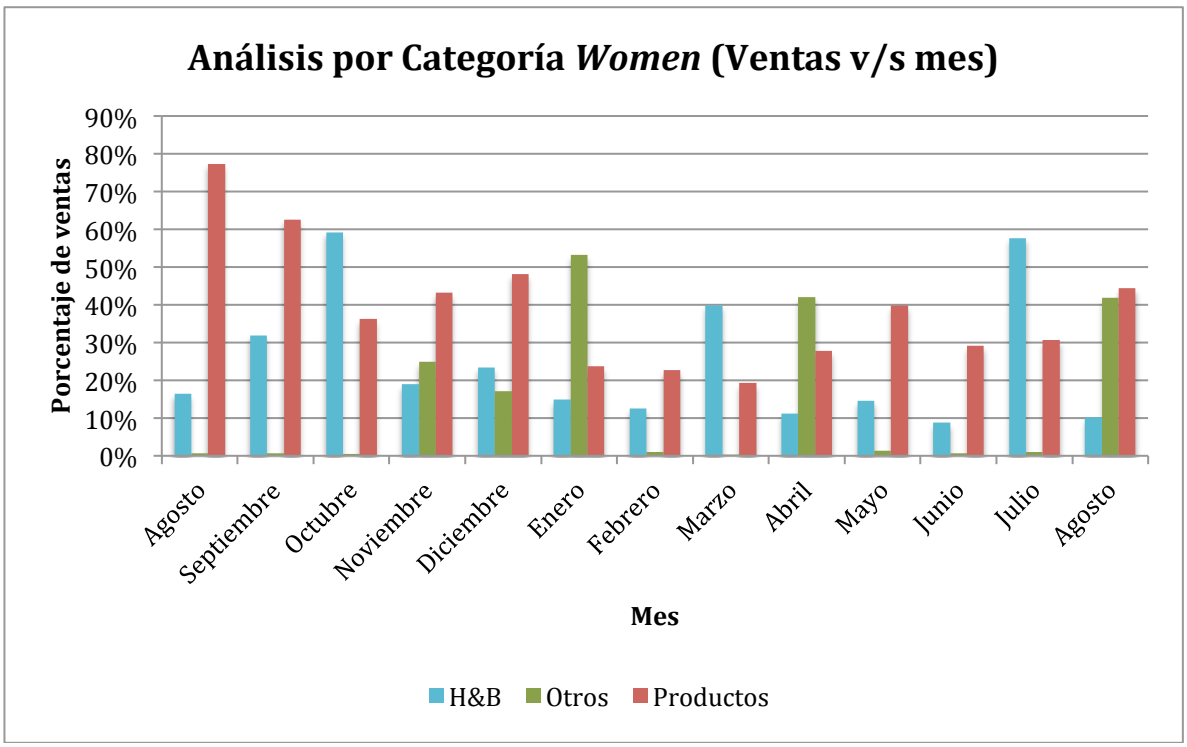


Ilustración 1.1: Análisis por categoría de *Women* (Ventas vs Mes)

Es por esto, que se plantea la realización de un modelo de recomendación que pueda sugerir productos y servicios a cada cliente según su comportamiento histórico, proporcionando así una solución para este problema.

Gran parte de los *retailers* del mundo que poseen una plataforma *web*, tienen un modelo de recomendación desarrollado que ayuda a sus clientes a encontrar lo que necesitan, y que, además, ayuda a la empresa a obtener rendimientos superiores. Sin embargo, Cuponatic no funciona de la misma manera, ya que al ser un sitio web de descuentos diarios (o de cupones de descuento), posee una alta rotación de productos y servicios, además de una amplia variedad de categorías, lo que hace que los modelos existentes no puedan crear una personalización ajustada.

Ahora bien, ¿Cómo hacer esto para más de cuatrocientos mil clientes? Mediante un modelo de recomendación, el cual de forma automática detecte las necesidades de cada cliente y envíe ofertas personalizadas a sus correos.

Actualmente, Cuponatic posee una baja Tasa de Clics y una baja Tasa de Apertura, siendo 2,05% y 12,24% respectivamente. La primera está por sobre el promedio de los sitios de ofertas diarias, correspondiente a un 1,89%, pero está bajo el promedio de los *e-commerce* en el mundo, que es de un 2,66%, cifra que la empresa pretende lograr en el corto plazo. La segunda está por debajo de la tasa de apertura de las empresas de cupones *online* del mundo, la cual es de un 13,69%, y por debajo de la tasa de apertura de los *e-commerce*, que es de un 16,87%. Estas cifras fueron recopiladas por el popular *email sender,* Mailchimp (*www.mailchimp.com*).

Actualmente, se le envía el correo a aproximadamente 400.000 clientes y las visitas al sitio *web* que se inician a través del correo corresponden al 51% del total de ellas, por lo que si se aumenta la tasa de clics en sólo 0,1% y al mismo tiempo se aumenta la tasa de apertura en un 1%, se podría aumentar el tráfico en el sitio web en al menos un 13,4%, teniendo un 13,4% más de clientes dispuestos a comprar.

1.3. Objetivos del Proyecto

1.3.1. Objetivo General

Diseñar un modelo de recomendación de productos y servicios de acuerdo con la información transaccional de cada cliente, aplicado a una empresa de cupones *online*.

1.3.2. Objetivos Específicos

1. Proponer y generar dos prototipos de modelos de recomendación.
2. Extraer recomendaciones de productos y servicios para clientes
3. Diseñar experimentos para testear los dos modelos de recomendación
4. Experimentar y evaluar la efectividad de los prototipos mediante los indicadores *Tasa de Clics* (*Click through rate o CTR* en inglés), *Tasa de Apertura* (*Open Rate* en inglés), *Tasa de Conversión*, y las *Ventas*.
5. Comparar ambos modelos para ver cuál es más efectivo para la empresa

1.4. Hipótesis a Testear

Las hipótesis principales del proyecto son las siguientes:

1. Un modelo de recomendación aumentaría la tasa de clics (*CTR*), ya que los clientes verían productos y servicios que sean acordes a sus gustos y necesidades, y no simplemente los más populares.

2. Un modelo de recomendación aumentaría la *Tasa de Apertura* (*Open Rate*), ya que los clientes al tener un trato personalizado, se sienten más cercanos a la empresa.

3. Un modelo de recomendación aumentaría la *Tasa de Conversión* de los clientes que ingresan a la página a través del correo, ya que los clientes verían productos y servicios que les interesan, lo que haría que ingresen al sitio *web* por un producto que les interesa en realidad.

4. Un modelo de recomendación aumentaría la cantidad de productos y servicios vendidos a través del correo, ya que al ver productos de su interés, los clientes van a ser más propensos a comprar productos o servicios de sus gustos o necesidades.

1.5. Resultados Esperados

Los resultados esperados del proyecto son los siguientes:

1. Creación de modelos de recomendación que determine qué se le debe sugerir a cada cliente.

2. Realización de experimentos que determinen la variación de la *Tasa de Clics* del *newsletter*, la *Tasa de Apertura*, *Tasa de Conversión*, y las Ventas comparando grupos de tratamiento con el grupo de control.

3. Determinación de modelo de recomendación más efectivo para la empresa.

1.6. Alcances del Proyecto

El alcance de esta memoria es el desarrollo de un prototipo de modelo de recomendación. Este prototipo, se realizará tomando en cuenta principalmente el género de los clientes y la información transaccional de cada uno. No se utilizarán las calificaciones que los clientes le asignan a los productos y servicios, ya que los datos de la encuesta, al ser esta enviada dos semanas después de la compra, no estarán disponibles en ese entonces. Además, el total de las calificaciones cubren tan solo el 11% de las ventas, por lo que no son suficientes para generar las recomendaciones y establecer las comparaciones pertinentes.

Debido a que sólo aproximadamente el 40% de los clientes tiene una o más compras, se puede observar que la información transaccional puede no ser determinante para hacer recomendaciones para todos. Es por esto que el grupo de control y el grupo experimental fueron definidos sobre la base de clientes que hayan realizado al menos tres compras históricas.

La evaluación de la efectividad del sistema será observada principalmente en la Tasa de Clics, en la Tasa de Apertura, en la Tasa de Conversión y en la cantidad de productos y servicios vendidos, además de los indicadores estadísticos de *Precision* y *Recall*.

2. MARCO CONCEPTUAL

2.1. Sistemas de Recomendación

Los sistemas de recomendación, son utilizados por empresas para sugerir productos o servicios a los clientes de forma personalizada. Muchas de estas, han demostrado que poseer un sistema de recomendación es una forma efectiva para hacer que los clientes sean satisfechos de acuerdo con sus necesidades, y más importante aún, hacer incrementar las ventas para la misma empresa [2].

Algunas empresas conocidas que poseen sistemas de recomendación, son *Amazon*, *Netflix*, y *Facebook*. El primero, utiliza un sistema de recomendación para sugerir productos de acuerdo con lo que han comprado los usuarios anteriormente, asociándolo y comparándolo con otros clientes. El segundo, utiliza las calificaciones de las películas para sugerir otras similares, de acuerdo al contenido de estas. Y, por último, Facebook utiliza a los "amigos" de cada usuario para poder sugerir otros "amigos" a este.

Para comenzar a hacer el prototipo de sistema de recomendación, se deben definir algunos conceptos para determinar qué tipo de métodos se utilizan en la actualidad. A lo largo de este capítulo, se dará a conocer el estado del arte de los sistemas de recomendación, donde se explicarán todos los algoritmos existentes y cuándo deben ser utilizados.

2.2. Tipos de Sistemas de Recomendación

Existen distintos tipos de sistemas de recomendación, que se utilizan de acuerdo con la información que posee cada empresa. Esta información puede ser:

i. Historial de compras de cada usuario
ii. Historial de navegación de cada usuario
iii. Calificación de productos o servicios
iv. Información (atributos) de productos o servicios

Los distintos tipos de algoritmos existentes en la actualidad para un sistema de recomendación son los siguientes:

1. *Algoritmos de Filtros Colaborativos*: se basan en el historial de actividades asociadas a la empresa (compras, navegación, calificaciones, entre otras).
2. *Algoritmos en base al Contenido*: se basan en la información que se posee de los productos (categorías, subcategorías, características del producto, entre otras).
3. *Algoritmos Híbridos*: se basan en una combinación de ambos algoritmos mencionados anteriormente.

En esta sección, se analizarán estos distintos algoritmos en detalle, donde se explicará cómo funcionan y cuándo es conveniente utilizarlos. Se pondrá más énfasis en los Filtros Colaborativos, debido a que es el método que se utilizará para hacer el prototipo.

2.2.1. Sistemas de Recomendación en base a Filtros Colaborativos

Los filtros colaborativos, se utilizan para asociar clientes o productos con otros similares [3][4][5].Para esto, se pueden utilizar todas las actividades que realiza cada cliente con el sitio *web* de la empresa, tales como el historial de compras, el historial de navegación, entre otros. Típicamente, para una mejor sugerencia personalizada, se utilizan las calificaciones que cada cliente le asigna a cada producto o servicio.

Los dos distintos tipos de filtros colaborativos que existen en la actualidad, son los siguientes:

i. *Filtros Colaborativos basados en la memoria*: Se utiliza todo el historial de actividades de los clientes con el sitio *web* de la empresa.
ii. *Filtros Colaborativos basados en un modelo*: Se utiliza un modelo, el cual va aprendiendo de la información de los clientes y no necesita utilizar siempre toda su información.

En esta sección, se explican detalladamente cada uno de estos.

2.2.1.1. Filtros Colaborativos basados en la memoria

Los algoritmos de filtros colaborativos basados en la memoria, utilizan toda la base de datos que posea información relevante para poder encontrar sugerencias de productos o servicios para cada cliente [6]. En general, son utilizados para encontrar usuarios similares para cada usuario, y productos (servicios) similares para cada producto (o servicio). El primero pertenece a los algoritmos basados en los usuarios, y el segundo a los algoritmos basados en los productos.

2.2.1.1.1. Filtros Colaborativos basados en los Usuarios

Los filtros colaborativos basados en el usuario (*user-based collaborative filtering* en inglés), determinan los clientes que son más parecidos a cada uno, para luego recomendar productos que estos otros han comprado, y que él no. Para esto, se utiliza el historial de compras de cada cliente, las calificaciones que cada cliente le asigna a los productos, el historial de navegación, o cualquier tipo de interacción que se tenga con los productos.

Para llevar a cabo este algoritmo, se deben definir dos elementos esenciales:

i. *Similitud de Usuarios*: Existen distintos tipos de métricas que determinan si un usuario es similar a otro que dependen de distintos elementos. Cada uno de estos son explicados más adelante en este capítulo.

ii. *Cantidad de Usuarios Similares*: Se debe definir cuántos usuarios son considerados como similares para cada uno. Para determinar esta cantidad, típicamente se utilizan los vecinos más cercanos a cada usuario. El cómo decidir cuántos considerar y cómo hacerlo, es explicado más adelante en este capítulo.

En general, un sistema de filtros colaborativos basado en el usuario, utiliza un algoritmo

que se puede explicar simplificándolo de la siguiente manera:

Algoritmo 1: Filtros Colaborativos basados en el usuario [7].

Para cada otro usuario v
calcular una similitud s entre u y v
retener los usuarios más similares, ordenados por s como la vecindad n
Para cada producto i en que algún usuario en n tenga una preferencia,
pero que u todavía no tenga una preferencia
Para cada otro usuario v en n que tenga una preferencia por i
calcular una similitud s entre u y v
incorporar las preferencias de v para i, ponderado por s, en una media móvil

Para explicar esto de una forma más gráfica, se considera el siguiente ejemplo:

	Producto 1	Producto 2	Producto 3	Producto 4	**Similitud con Usuario 1**
Usuario 1	X	-	X	-	**1**
Usuario 2	-	X	-	-	**0**
Usuario 3	X	-	X	X	**0,67**
Usuario 4	-	-	X	-	**0,33**

Tabla 2.1: Filtros Colaborativos basados en el usuario

Las cruces que aparecen en la Tabla 2.1, corresponden a las interacciones de cada usuario con cada producto, ya sea una compra o una visita. La similitud con el usuario 1, está dado por una métrica llamada *Distancia Jaccard*. En el caso que se decidiera encontrar sólo un usuario más cercano al Usuario 1, sería el Usuario 3, por lo que el Producto 4 sería el producto más recomendable para el Usuario 1.

Este tipo de algoritmo, es recomendable utilizarlo cuando la cantidad de usuarios es menor a la cantidad de productos en la base de datos, ya que el cálculo de la similitud entre ellos demora menos que el cálculo entre productos.

2.2.1.1.2. Filtros Colaborativos basados en los Productos

A diferencia de los filtros colaborativos basados en los usuarios, este algoritmo se basa en encontrar productos similares de acuerdo con las interacciones de cada usuario con un sitio *web*. Estas interacciones, pueden ser las mismas utilizadas en el algoritmo anterior, ya sea el historial de compras, el historial de navegación, las calificaciones que usuarios le asignan a producto, o cualquier dato relevante que pueda ser útil a la hora de comparar los usuarios.

Las métricas de similitud de usuario y cantidad de usuarios similares son las mismas utilizadas en ambos algoritmos, por lo que el algoritmo general de los filtros colaborativos basados en los productos se puede caracterizar de la siguiente forma:

Algoritmo 2: Filtros Colaborativos basados en los productos [7]

Para cada producto i que u todavía no tenga una preferencia
    Para cada producto j que u tenga una preferencia
        calcular la similitud s entre i y j
        agregar la preferencia de u por j, ponderado por s, a una media móvil
retornar los items más similares, ordenados por la ponderación de la media móvil

Para explicar este algoritmo de la misma manera que el anterior, se muestra el siguiente ejemplo:

	Producto 1	Producto 2	Producto 3	Producto 4	Producto 5
Usuario 1	5	3	2	3	1
Usuario 2	4	2	1	2	2
Usuario 3	3	3	2	1	4
Similitud con Producto 1	**1**	**0**	**0**	**0,38**	**-0,37**

Tabla 2.2: Filtros Colaborativos basados en los productos

En este caso, los números que aparecen en la Tabla 2.2 corresponden a las calificaciones asignadas por cada usuario a cada producto que previamente este consumió. La similitud con el Producto 1, está dada por una métrica llamada *Coeficiente de Correlación de Pearson*. El producto más similar al Producto 1 es el Producto 4, por lo que si un usuario compra el Producto 1, se le debe recomendar el Producto 4.

Este tipo de algoritmo se debe utilizar cuando la cantidad de productos es menor a la cantidad de usuarios, ya que el tiempo que demora en calcular las similitudes es menor que en el caso contrario.

2.2.1.1.3. Similitud entre Usuarios

Como se ha mencionado anteriormente, existen distintas métricas para determinar si un usuario es similar a otro, o bien, si un producto es similar a otro. En esta sección, se explican los principales métodos existentes y cuándo deben ser utilizados.

i. *Distancia Euclidiana*:

La *Distancia Euclidiana* es la métrica más simple que existe para medir similitudes. Esta es conocida por medir la distancia entre dos puntos en un plano mediante la siguiente fórmula:

$$d(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Para utilizar esta expresión, es necesaria la información de las calificaciones que cada usuario le asigna a los productos, por lo que en caso de no tener esta información, no se puede utilizar. Por otro lado, esta métrica se debe utilizar cuando se desea una medición exacta, ya que no considera una normalización de las calificaciones. Para ver ejemplos de usos de esta métrica, ver Anexo 2.

ii. *Distancia Euclidiana Cuadrática*:

La *Distancia Euclidiana Cuadrática* corresponde, como lo dice su nombre, al cuadrado del valor de la Distancia Euclidiana. Su valor es calculado mediante la siguiente fórmula:

$$d(X,Y) = \sum_{i=1}^{N} (x_i - y_i)^2$$

Esta métrica es utilizada cuando son demasiados valores los que se deben calcular, por lo que la *Distancia Euclidiana* puede tardar más al calcular la raíz cuadrada. En general, para los algoritmos de filtros colaborativos, el resultado de la *Distancia Euclidiana Cuadrática* es el mismo que el de la *Distancia Euclidiana*, ya que a pesar de que los valores van a ser mayores, siempre se mantendrá el orden de estos. Para ver ejemplos de usos de esta métrica, ver Anexo 2.

iii. *Distancia Manhattan*:

La *Distancia Manhattan* también es conocida por medir dos puntos en un plano, pero no calculando la distancia diagonal entre estos (*Distancia Euclidiana*), sino que en distancias horizontales o verticales. Esta distancia es calculada mediante la siguiente fórmula:

$$d(X,Y) = \sum_{i=1}^{N} |x_i - y_i|$$

Es recomendable utilizar la *Distancia Manhattan* bajo las mismas condiciones que la *Distancia Euclidiana*, ya que al igual que esta, es necesaria la información de calificaciones que los usuarios le asignan a los productos, y la distancia puede ser calculada de forma similar. El resultado de la Distancia Manhattan es el mismo que los dos anteriores, a pesar de que los valores encontrados son distintos. Para una completa explicación de usos de esta métrica, ver Anexo 2.

iv. *Distancia Coseno*:

Para utilizar la *Distancia Coseno*, es necesario pensar en cada calificación como un punto en un plano, y trazar un vector desde el origen hasta dicho punto, con el fin de formar un ángulo $\boldsymbol{\theta}$ entre ambos vectores. Para calcular esta distancia se utiliza la siguiente fórmula:

$$d(X,Y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}} = \frac{X \times Y}{||X|| \cdot ||Y||}$$

Esta medida de similitud es de mucha utilidad cuando se desea encontrar patrones en las calificaciones que los usuarios le asignan a los productos. Sin embargo, no considera la distancia entre estas, es decir, pueden seguir un mismo patrón; pero puede que no sean tan parecidos en los productos comprados. Para una mejor explicación de usos de esta métrica, ver Anexo 2.

v. *Distancia de Jaccard*:

A diferencia de las distancias mencionadas anteriormente, que capturan la distancia o el ángulo entre los distintos usuarios, el *Coeficiente de Tanimoto*, o también conocida como la *Distancia Jaccard*, calcula la similitud entre dos usuarios de acuerdo a qué tan probable es que tengan los mismos productos en su canasta de compra (Musalem & Bosch, 2001).

Se calcula mediante la siguiente expresión:

$$T = \frac{\sum_{i=1}^{N}(x_i \wedge y_i)}{\sum_{i=1}^{N}(x_i \vee y_i)}$$

La limitación de esta medida de similitud, es que solo es aplicable para interacciones del usuario con el sitio web, ya sean compras o visitas, y no las calificaciones que cada usuario le asigna a cada producto. Por lo tanto, los resultados difieren de las anteriores al no incluir esta información que en muchos casos es muy valiosa. Sin embargo, para sitios webs de *e-commerce* que no tienen información sobre calificaciones, es una muy buena opción, puesto que esta métrica ignora los productos que no han sido comprados por ninguno de los usuarios, por lo que el resultado es muy ajustado. Para una mejor explicación de usos de esta métrica, ver Anexo 2.

vi. *Coeficiente de Correlación de Pearson*:

El *Coeficiente de Correlación de Pearson* mide la tendencia de dos series de números que se mueven en la misma dirección. En otras palabra, si dos usuarios tienen calificaciones similares o simplemente similares proporcionalmente (al igual que la *Distancia Coseno*), serán similares según esta métrica.

Para calcular este coeficiente, se utiliza la siguiente expresión:

$$d(X,Y) = \frac{N\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)\,(\sum_{i=1}^{N} y_i)}{\sqrt{(N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2)(N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2)}}$$

Esta métrica es una buena medida cuando los usuarios tienen una gran cantidad de productos calificados, ya que no considera los productos que no tienen calificación, y si estos son muchos, el coeficiente podría ser calculado erróneamente. Por otro lado, es recomendable usar esta medida sólo para sitios que posean información de calificaciones, de otra forma el coeficiente no será calculado correctamente.

Finalmente, es una buena medida para regular o escalar las calificaciones, al igual como lo hace la *Distancia Coseno*, ya que el coeficiente toma en cuenta si las calificaciones por ambos son altos o bajos en comparación con sus otras calificaciones. Para una mejor

explicación de usos de esta métrica, ver Anexo 2.

vii. *Correlación de Spearman*:

Esta medida de similitud es una variante del Coeficiente de Correlación de Pearson, ya que se utiliza la misma expresión, pero cambiando las calificaciones de los usuarios. En primer lugar, se reemplaza por un 1 la calificación del último producto comprado, por un 2 el penúltimo producto comprado y así sucesivamente.

Al utilizar esta métrica, se tienen las mismas ventajas y desventajas que al ocupar la métrica anterior, ya que la fórmula es la misma. Sin embargo, el resultado de ambos es distinto, pues el Coeficiente de Spearman calcula la tendencia del orden de las compras, en lugar de obtener similitudes en las calificaciones. Para una mejor explicación de los usos de esta métrica, ver Anexo 2.

viii. *Similitud Log-verosimilitud*:

Esta métrica de similitud funciona de forma similar al *Coeficiente de Tanimoto*, y al igual que esta, no considera los productos que no son comprados por ninguno de los usuarios para calcular la similitud y solo considera las compras (no las calificaciones). Sin embargo, mientras *Tanimoto* calcula la similitud utilizando las compras anteriores de cada uno, la *Similitud Log-verosimilitud* calcula qué tan improbable es que la superposición de productos comprados por ambos usuarios sea debido a la suerte. Por lo tanto, mientras más improbable es, más similares son los usuarios [8].

Por ejemplo, si dos usuarios comparten 3 productos en común, de un total de 5 cada uno, van a ser mucho más similares que si cada uno hubiese comprado 30, donde es probable que esos productos en común se hayan comprado en común debido a la suerte.

La tabla 2.3 muestra los datos que poseen dos eventos, A y B.

	Evento A	Todo menos A
Evento B	A y B juntos (k_{11})	B, sin A (k_{12})
Todo menos B	A, sin B (k_{21})	Ni A ni B (k_{22})

Tabla 2.3: Similitud Log-Verosimilitud

Asumiendo que el evento A corresponde a las compras realizadas por el cliente A y que el evento B corresponde a las compras realizadas por el cliente B, la tabla quedaría formada por:

$$k_{11} = ||A \cap B||$$

$$k_{12} = ||B \backslash A||$$

$$k_{21} = ||A \backslash B||$$

$$k_{21} = \emptyset$$

Para calcular el *ratio Log-verosimilitud*, o la similitud entre ambos usuarios, se utiliza la siguiente fórmula:

$$LLR = 2 * \left(H\left(\sum_{matriz} k \right) - H\left(\sum_{fila} k \right) - H\left(\sum_{col} k \right) \right)$$

Donde $H(\cdot)$ corresponde a la *entropía de Shannon*, que está dado por la siguiente fórmula:

$$H = \sum_{k=i,j} k \left(log\left(\sum_{k=i,j} k \right) \right) - \; k_{ij} \log\left(k_{ij}\right)$$

Al igual que la *Distancia Jaccard*, es recomendable utilizar esta métrica cuando se posee solo la información transaccional o de navegación de parte de los usuarios, ya que estas son las medidas que solo utilizan este tipo de información. Es una buena opción a la hora de encontrar una similitud, ya que omite los productos que no son comprados por ambos usuarios comparados, y además, calcula observando las demás compras si coinciden en estas debido a la suerte o porque son similares.

2.2.1.1.4. Cantidad de Usuarios Similares

Para saber cuántos usuarios son considerados similares a otro, existen dos tipos de medidas, las cuales son descritas en esta sección.

i. *K-Vecinos más cercanos*:

Se utilizan los *K-Vecinos más cercanos* (*K-nn: K-Nearest Neighborhood* en inglés), cuando se desea obtener un número fijo K de usuarios más similares a otro. Por ejemplo, si se desean 10-Vecinos más cercanos, el algoritmo calculará quienes son los 10 usuarios más parecidos a cada uno sin considerar qué tan similares son.

Esta cantidad es utilizada cuando se desea obtener la mayor cantidad de usuarios similares posibles. Sin embargo, cuando es requerido que los usuarios sean ciertamente similares o más, esta no es una buena determinación, ya que es posible que el 10º usuario más cercano sea muy distinto.

ii. *Vecinos cercanos dependientes de la métrica de similitud*:

Esta cantidad de usuarios es utilizada cuando se desea obtener usuarios similares dependiendo de la similitud que se emplee. Por ejemplo, si uno utiliza la *Distancia Coseno* para calcular la similitud entre usuarios, uno puede determinar todos los usuarios que tengan una distancia menor a 0,2, donde dependiendo de los usuarios que existan en la base de datos, pueden ser una gran cantidad o una baja cantidad. Es por esto que, dependiendo de la métrica de similitud, se debe elegir cómo determinar la cantidad de usuarios que se considerarán como similares.

Por lo tanto, es conveniente utilizar esta forma de determinar la cantidad de usuarios similares cuando se sabe que existen muchos usuarios similares entre sí, ya que en caso contrario, puede que para muchos de estos no encuentre ningún vecino más cercano.

Todas las métricas mencionadas anteriormente, fueron explicadas para los algoritmos basados en los usuarios, sin embargo, se pueden aplicar análogamente para los algoritmos basados en los productos.

2.2.1.1.5. Ventajas y Desventajas de utilizar Filtros Colaborativos basados en la memoria

Las principales ventajas de utilizar estos algoritmos, radican en que los resultados son mucho más ajustados a la realidad que los basados en un modelo, debido a que utiliza toda la base de datos y toda la información para generar recomendaciones.

Por otro lado, se necesita un algoritmo relativamente simple para calcular las recomendaciones, por lo que se puede actualizar la base de datos diariamente de forma muy simple.

A pesar de que este método genera buenos resultados al momento de hacer recomendaciones, al utilizar toda la base de datos puede hacer que el cálculo de éstas sea muy lento, por lo que no siempre es conveniente utilizarlo. En otras palabras, si la cantidad de usuarios y de productos es muy grande, el algoritmo puede demorar horas en tener resultados, lo que puede hacer que no funcione adecuadamente para todas las funciones que requiere cada empresa.

Además, puede que este sistema no siempre genere recomendaciones para todos los usuarios, ya que existe la posibilidad de que un usuario no tenga productos comprados o visitados en común con otro.

2.2.1.2. Filtros Colaborativos en base a un modelo

Los Filtros Colaborativos basados en un modelo se crean a partir de una base de datos que contiene calificaciones sobre productos [6][9]. A diferencia de los Filtros Colaborativos basados en la memoria, no utilizan esta información todas las veces que se desean encontrar recomendaciones, sino que el sistema va aprendiendo de los datos que se le entregan, y va formando un modelo que va definiendo qué se le debe recomendar a cada uno sin utilizar toda la información histórica.

En la actualidad, existen tres principales métodos que son utilizados por este tipo de filtros colaborativos:

i. *Modelos Bayesianos*: Es un modelo basado en el Teorema de Bayes, que es utilizado para pronosticar calificaciones de productos no consumidos por un usuario.
ii. *Reglas de Asociación*: Permiten asociar canastas de compras de los clientes para identificar patrones y generar recomendaciones a partir de éstos.
iii. *Segmentación*: Se basa en clasificar y agrupar probabilísticamente a los clientes según su similitud, para luego determinar qué se le debe recomendar a los miembros de cada grupo.

2.2.1.2.1. Redes y Modelos Bayesianos

Los *Modelos Bayesianos* [10][11][12] son una forma de predecir o pronosticar valores en circunstancias de incertidumbre. Algunas veces, puede estar representado por redes o grafos probabilísticos, donde los nodos representan variables, que pueden ser representados mediante usuarios o productos, y donde los arcos representan la relación entre ellas.

En el caso de los Filtros Colaborativos, se utilizan los *modelos bayesianos* para estimar las calificaciones de productos que un usuario le asignaría previamente a su compra, con el fin de recomendar los productos con una mayor calificación.

Este modelo funciona con la idea de agrupar todos los usuarios que posean juicios de productos similares como si tuvieran una distribución de probabilidad de calificaciones idéntica. Para su cálculo, como lo dice su nombre, se utiliza el Teorema de Bayes.

2.2.1.2.2. Reglas de Asociación

Otro modelo muy conocido en las empresas de *retail*, son las *Reglas de Asociación*. Estas consisten en identificar patrones en las canastas de compras, como por ejemplo, si muchos clientes que compran pan y mantequilla, a los clientes que compren pan y no mantequilla, se les recomendaría la mantequilla.

Existen dos principales tipos de identificar estas reglas, las cuales se explican a continuación.

i. *Apriori*:

El algoritmo *Apriori* [13] utiliza la información de las canastas de compras o las compras históricas de cada cliente para luego generar reglas de asociación. Se considera la Tabla 2.4 como un ejemplo de estas compras.

Usuario	Compras
Usuario 1	P1, P2, P3, P4
Usuario 2	P2, P4
Usuario 3	P1, P3, P4
Usuario 4	P2, P3, P4

Tabla 2.4: Ejemplo de canastas de compras de clientes

Una vez teniendo estas canastas de compras de cada cliente, se debe definir cuántas compras de productos son consideradas como relevantes. Por ejemplo, el Producto 4 (P4), fue comprado 4 veces (por Usuario 1, 2, 3 y 4). Se considerará relevante una cantidad de 3 compras, número al que se le llama *Soporte*, explicado en detalle más adelante junto con otras definiciones importantes.

Utilizando el algoritmo *Apriori*, las reglas se definirían de la siguiente manera:

Producto	Cantidad
P1	2
P2	3
P3	3
P4	4

Luego, se realizan todas las combinaciones de dos productos que tengan una cantidad de compras mayor a 3, y se comparan con las canastas de compras mencionadas anteriormente.

Producto	Cantidad
P2, P3	1
P2, P4	3
P3, P4	3

Se hace lo mismo que en la etapa anterior, pero uniendo tres productos.

Producto	Cantidad
P2, P3, P4	2

Como la cantidad 2 es menor a 3, se termina el algoritmo, por lo que una regla de asociación indica que los usuarios que compren uno de esos productos, también deberían comprar los otros dos, por lo que se convertirían en una recomendación.

ii. *FP-Growth*:

Otro conocido algoritmo de Reglas de Asociación, es el *FP-Growth* [14]. Este es un método eficiente que observa patrones frecuentes (*FP*) de canastas de compras sin generar candidatos (como el algoritmo *Apriori*), y lo hace a través de un árbol de patrones frecuentes (*FP-Tree*).

Tomando en cuenta la Tabla 2.4 utilizada como ejemplo en el algoritmo anterior, se debe contar la cantidad de compras totales de cada producto y, además, ordenarlos por prioridad.

Producto	Cantidad	Prioridad
P1	2	4
P2	3	3
P3	3	2
P4	4	1

Para comprender mejor el algoritmo, se ordenan las canastas de compras de acuerdo a la prioridad de los productos.

Usuario	Compras
Usuario 1	P4, P3, P2, P1
Usuario 2	P4, P2
Usuario 3	P4, P3, P1
Usuario 4	P4, P3, P2

Luego, se crea el árbol de patrones frecuentes. Este árbol, se crea a partir de las transacciones ordenadas, con lo que se obtiene la figura de la Ilustración 2.1.

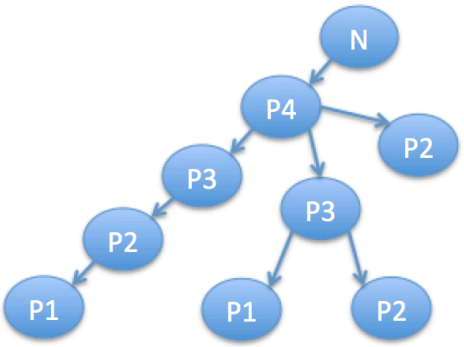


Ilustración 2.1: Ejemplo de árbol de Algoritmo FP-Growth

Aquí se pueden ver los siguientes patrones frecuentes:

P4: {P1, P4 : 2}, {P2, P4 : 2}, {P3, P4 : 2}, {P1, P3, P4 : 2}, {P2, P3, P4 : 2}

P3: {P1, P3 : 2}, {P2, P3 : 2}

P2: ---

P1: ---

Estas son las reglas que se pueden utilizar para recomendar productos si una persona compra uno dentro de estos patrones.

iii. Diferencias entre *Apriori* y *FP-Growth*

Ambos métodos, *Apriori* y *FP-Growth*, dan como resultados reglas muy similares. Sin embargo, el proceso de creación de reglas de cada uno de ellos es distinto. Es por esto, que en la Tabla 2.5 se muestra una comparación en cada uno de los aspectos relevantes a la hora de generar el modelo.

Parámetro	*Apriori*	*FP-Growth*
Memoria utilizada	Debido a que existe una gran cantidad de posibles reglas a generar, se necesita una mucho espacio de memoria.	Debido a la estructura de cómo funciona este algoritmo, y al no generar grandes cantidades de posibles reglas, necesito menos espacio de memoria que el método *Apriori*.
Número de iteraciones	Se necesita hacer muchas iteraciones dependiendo de qué cuántas sean las canastas de compras.	Sólo realiza dos iteraciones de toda la base de datos, ya que sólo se crea el *FP-Tree* y luego se extraen los patrones frecuentes.
Tiempo	Al necesitar una gran cantidad de espacio de memoria y al hacer muchas iteraciones, el tiempo de demora puede ser bastante.	Al contrario del método *Apriori,* se necesita poco espacio de memoria y sólo toma dos iteraciones para generar las reglas, por lo tanto el tiempo de demora es menor.

Tabla 2.5: Comparación de Algoritmos Apriori y FP-Growth

iv. *Soporte*, *Confianza* y *Elevación*:

Para determinar qué cantidad de compras es relevante para crear una regla de asociación, se debe indicar cuánto es el *Soporte (Support)*. Éste viene dado por la cantidad de veces en que se repite una regla dentro del total de canastas de compra. Por ejemplo, considerando las compras de la Tabla 2.8, se tiene que el *soporte* de la regla P2 → P4 es de 0,75, puesto que se han comprado en conjunto en 3 de las 4 compras.

Para determinar qué tan fuerte es una regla, se debe observar la *Confianza (Confidence)*.

Esta determina qué fracción de donde aparece el antecedente, aparece también la consecuencia.

$$confianza(X \rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

Tomando el mismo ejemplo anterior, la regla P2 → P4 tiene una confianza de 1, dado que en todas las transacciones que aparece P2, también aparece P4.

Finalmente, se debe definir el concepto de *Elevación (Lift)*. Este corresponde a qué tanto más probable es que se compre la consecuencia cuando se compró ese antecedente, a que se compre en una transacción normal. Mientras más alto es el valor de la *elevación*, mejor es la regla.

$$elevación(X \rightarrow Y) = \frac{soporte(X \cup Y)}{(soporte(X) * soporte(Y))}$$

En el mismo caso anterior, la regla P2 → P4 tiene una *elevación* de 1.

2.2.1.2.3. Segmentación

Para hacer un modelo que requiera grandes cantidades de elementos, ya sea usuarios o productos, puede ser conveniente segmentarlos. Al hacer esto, se pueden encontrar relaciones entre ellos, pudiendo así por ejemplo, lograr recomendar productos que otro usuario del mismo segmento haya comprado.

Esta técnica puede ser utilizada al mismo tiempo que las descritas anteriormente. Por ejemplo, si se segmenta un grupo al cual le gusta la música, se pueden encontrar reglas de asociación dentro de ese segmento como la siguiente:

$$Guitarra \rightarrow Cuerdas\ de\ guitarra.$$

Existen distintos métodos para lograr una efectiva segmentación, los cuales se explican a continuación.

i. *K-Means*
ii. *X-Means*
iii. *Fuzzy C-Means*
iv. *Expectation-Maximization* (*EM*)

Para ver una explicación de cada uno de estos métodos, véase el Anexo 3.

2.2.1.2.4. Ventajas y Desventajas de los Filtros Colaborativos en base a un modelo

Una de las ventajas más importantes que tienen los sistemas en base a un modelo, recae en la escalabilidad, pues los resultados que se obtienen son mucho más pequeños que la base de datos original. En consecuencia, para hacer recomendaciones a usuarios dentro de bases de datos que son muy grandes, es conveniente utilizar un sistema en base a un modelo.

Por otro lado, un sistema en base a un modelo es mucho más rápido a la hora de hacer recomendaciones que un sistema en base a la memoria, ya que como se dijo anteriormente, no utiliza toda la base de datos cada vez que se utiliza el modelo.

Existen dos principales desventajas en los *sistemas en base a un modelo*. En primer lugar, estos sistemas son muy inflexibles, ya que como estos son utilizados para ahorrar tiempo y para calcular recomendaciones en grandes bases de datos, agregar información a estos es muy difícil.

En segundo lugar, como los *sistemas en base a un modelo* no utilizan toda la información disponible cada vez que calcula las recomendaciones, y se basa sólo en un modelo, la calidad de los pronósticos no es muy buena en comparación con los sistemas en base a la memoria.

2.2.2.Sistemas de Recomendación en base al Contenido

Otros métodos conocidos y estudiados en el Estado del Arte, son los *Sistemas de Recomendación basados en el contenido* [15][16] .En general, este método utiliza toda la información sobre productos previamente comprado por cada usuario, la que puede ser, por ejemplo, descripciones, atributos, propiedades, documentos, precios, categorías, entre otras. Con esto se forma un modelo que determina qué productos, que contienen similar información, son los indicados para recomendar a cada usuario.

Estos sistemas incorporan las siguientes etapas a la hora de filtrar la información de los productos:

i. Análisis de Contenido:

La descripción o la caracterización de los productos está formada por texto, el cual no siempre tiene una estructura determinada. Es por esto que es importante exportar elementos principales de cada descripción para ser utilizado en las siguientes etapas. Así, utilizando distintas técnicas para extraer atributos principales de cada producto, se logra tener un análisis del contenido que sirve para continuar con la creación de un modelo.

ii. Creador de Perfiles Inteligentes:

Una vez teniendo el análisis del contenido de cada producto que cada usuario ya compró, se extrae toda la información relevante de cada interacción. Esta información se ve representada típicamente como calificaciones de productos o si presionó “me gusta”/“no me gusta”. Esto genera un modelo que crea un perfil de cada usuario mediante la información de cada producto que “le gustó”, o bien que calificó de buena manera.

iii. Filtrado de Componentes:

El Filtrado de Componentes utiliza ambas etapas anteriormente explicadas para generar un modelo en el cual se toma la información extraída de los productos y el perfil de “gustos” de cada usuario, y los compara con el perfil de cada uno de los productos disponibles. Es así como se generan recomendaciones personalizadas para cada uno de ellos.

2.2.2.1. Ventajas y Desventajas de los Sistemas de Recomendaciones en base al contenido

La principal ventaja de los Sistemas de Recomendaciones en base al contenido, recaen en que a diferencia de los filtros colaborativos, sí pueden recomendar nuevos productos incorporados recientemente al sitio *web*, ya que este último sólo recomienda productos que ya fueron comprados por otros usuarios.

A pesar de que este sistema es capaz de recomendar nuevos productos que posea el sitio *web*, son muy dependientes de la información para hacer las recomendaciones, por lo que no generaliza patrones de similitud como lo hacen los filtros colaborativos.

Por otro lado, para hacer buenas recomendaciones, se requiere tener toda la información necesaria en cada producto, porque en caso contrario, productos que sean similares y no posean toda la información, no serán recomendados.

Otra desventaja de estos sistemas, es que normalmente son muy exigentes a la hora de comparar productos, por lo que generalmente recomiendan productos que son demasiado similares, lo que no siempre es bueno.

2.2.3.Sistemas de Recomendación Híbridos

Los Sistemas de Recomendación Híbridos [17], son los que combinan de alguna forma los sistemas de filtros colaborativos y los en base al contenido. Las distintas formas de combinarlos son explicadas a lo largo de este capítulo.

2.2.3.1. Sistemas Ponderados

Un sistema se considera ponderado cuando la calificación de un ítem por recomendar, es calculado con una ponderación por cada uno de los sistemas utilizados.

Por ejemplo, uno puede considerar un sistema creado en base a filtros colaborativos que al generar recomendaciones haga un pronóstico de una calificación de un producto, y a la vez, utilizando un sistema en base al contenido haga otro pronóstico. Una ponderación de estas (que pueden ser 0,5 cada uno, es decir, el promedio) da como resultado una nueva calificación, la que se utilizará finalmente para recomendar productos.

2.2.3.2. Sistemas Variables

Un sistema es variable cuando se consideran los sistemas de filtros colaborativos y los en base al contenido según los diferentes resultados.

Por ejemplo, se puede utilizar primero el sistema de filtros colaborativos, pero si este no da los resultados esperados, se puede utilizar el sistema en base al contenido, y viceversa. De esta forma, se pueden obtener más recomendaciones, o bien, recomendaciones con una calificación igual o superior a la realizada con solo un sistema.

2.2.3.3. Sistemas Mezclados

Un sistema de recomendación que utilice sistemas con filtros colaborativos y sistemas en base al contenido al mismo tiempo, se considera un Sistema Mezclado.

Al utilizar ambos sistemas a la vez, se pueden obtener recomendaciones más precisas, ya

que al utilizar filtros colaborativos, se pueden obtener recomendaciones de productos similares o productos comprados por usuarios similares. De igual modo, al usar un sistema en base al contenido, puede tener recomendaciones similares y poder así recomendar nuevos productos también.

2.2.3.4. Sistemas de Atributos Combinados

Este sistema utiliza los resultados de un sistema de filtros colaborativos como un "*input*" para un sistema en base al contenido.

La ventaja de utilizar este método, recae en que se pueden obtener recomendaciones sobre productos similares de acuerdo con la descripción de cada uno principalmente, pero además considera productos o usuarios similares, lo que hace que la recomendación no caiga en la sensibilidad de que muy pocos usuarios hayan comprado o calificado cada producto.

2.2.3.5. Sistemas Cascadas

Los *Sistemas Cascadas*, se utilizan para mejorar la recomendación, mediante el uso de un sistema en primer lugar, ya sea filtros colaborativos o en base al contenido, y después utilizar el otro.

Por ejemplo, si se desea hacer una recomendación en base a filtros colaborativos, los resultados de este se pueden mejorar aún más si estos son un "*input*" para un sistema en base al contenido, donde los resultados pueden ser más precisos.

2.3. Evaluación de Sistemas de Recomendación

En muchas ocasiones, donde se tiene que elegir qué sistema utilizar o qué métrica de similitud se debe emplear, se pueden comparar haciendo evaluaciones de efectividad de estos mismos. En esta sección, se definen y se explican los distintos métodos de evaluación que existen en el Estado del Arte [18]. Dentro de estos se encuentran el *Root Mean Squared Error (RMSE)*, el *Mean Absolut Error (MAE),* la *Precision*, *Recall* y *F-Measure*.

2.3.1. *Root Mean Squared Error* (*RMSE*)

Esta métrica de evaluación es la más popular dentro de los métodos de evaluación de Sistemas de Recomendación. Dado un Sistema que pronostique calificaciones $\hat{\boldsymbol{r}}_{\boldsymbol{ui}}$ en una matriz de $\boldsymbol{N}$ pares de usuarios-productos $(\boldsymbol{u}, \boldsymbol{i})$, donde las verdaderas calificaciones $\boldsymbol{r}_{\boldsymbol{ui}}$ son conocidas, el *RMSE* viene dado por la siguiente fórmula:

$$RMSE = \sqrt{\frac{1}{|N|} \sum_{u,i \in N} (\hat{r}_{ui} - r_{ui})^2}$$

2.3.2.*Mean Absolut Error* (*MAE*)

Otra métrica de evaluación muy popular es el *Mean Absolut Error* (*MAE*). Esta se calcula bajo las mismas condiciones que la anterior, mediante la siguiente fórmula:

$$MAE = \sqrt{\frac{1}{|N|}\sum_{u,i \in N} |\hat{r}_{ui} - r_{ui}|}$$

Como se puede observar en las fórmulas del *RMSE* y del *MAE*, ambas difieren solo en que uno calcula el cuadrado de la diferencia entre las calificaciones pronosticadas y las reales, y que el otro calcula el valor absoluto de ellas. Así, se tiene que el RMSE demuestra más ampliamente los errores. Tomando el mismo ejemplo que en [21], si se tienen dos sistemas con cuatro calificaciones desconocidas, uno con un error de 2 en tres de las calificaciones y de 0 en la cuarta, y otro con un error de 3 en una calificación y 0 en los otros tres, el *RMSE* preferiría el primer sistema, mientras que el *MAE* preferiría el segundo.

2.3.3.Precision

Otras métricas utilizadas frecuentemente al evaluar Sistemas de Recomendación, son *Precision* y *Recall*. Antes de definir que son cada una de estas y cómo funcionan, se debe explicar bajo qué escenarios se utilizan.

Dado un sistema que genera recomendaciones para un usuario ***u***, se pueden tomar parte de cada una de las compras realizadas por este y no incluirlas en el sistema. Así, se puede observar si el sistema recomendaría, según los otros productos comprados, lo que ya alguna vez compró (los no incluidos en el sistema).

Estos resultados, se ingresan en una tabla como la que se muestra en la Tabla 2.6.

	Recomendado	No Recomendado
Usado	Verdadero-Positivo (VP)	Falso-Negativo (FN)
No usado	Falso-Positivo (FP)	Verdadero-Negativo (VN)

Tabla 2.6: Clasificación de posibles resultados de un modelo de recomendación para un usuario

La tabla muestra cuatros distintos posibles resultados: VP, FN, FP y VN. El *Verdadero-Positivo* (VP) refleja la cantidad de productos dejados fuera del sistema y que a la vez fueron recomendados por el mismo, el *Falso-Negativo* (FN) refleja la cantidad de productos dejados fuera del sistema y que no fueron recomendados en él, el *Falso-Positivo* (FP) refleja la cantidad de productos recomendados por el sistema que el usuario nunca había consumido, y el *Verdadero-Negativo* (VN), la cantidad de productos restantes en dicho sistema.

La métrica *Precision* está dada por la siguiente expresión:

$$Precision = \frac{VP}{VP + FP}$$

Esta refleja la tasa de productos que fueron recomendados por el sistema y que a la vez fueron consumidos por el usuario alguna vez, por sobre todos los productos alguna vez consumidos y que se hayan dejado fuera del sistema.

2.3.4.Recall (True Positive Rate)

Bajo las mismas condiciones que la métrica *Precision*, el *Recall* es calculado mediante la siguiente fórmula:

$$Recall = \frac{VP}{VP + FN}$$

Esto refleja la tasa de productos que se dejaron fuera del sistema y que a la vez fueron recomendados por sobre la suma de productos dejados fuera del sistema.

2.3.5.F-Measure

Esta métrica utiliza los métodos explicados anteriormente, *Recall* y *Precision* [19]. Se calcula una media harmónica ponderada utilizando la siguiente expresión:

$$F_1 = 2 * \frac{Recall * Precision}{Recall + Precision}$$

El resultado se manifiesta como un porcentaje al igual que las dos métricas anteriores, y sirve para medir la precisión de un sistema en general. Al utilizar ambas, *Recall* y *Precision*, resulta una buena medida para calcular la eficiencia de este.

3. METODOLOGÍA

La metodología que se utilizará en este proyecto, está basada en el proceso *KDD* (*Knowledge Discovery in Databases* en inglés) [20], el cual se emplea frecuentemente para hacer minería de datos.

Para llevar a cabo la creación del modelo de recomendación, se modificó este proceso para investigar qué es lo que actualmente existe de acuerdo con este tema y se ajustó a los factores relevantes para la empresa.

En esta sección, se explican los pasos en detalle de cómo se realizó el proyecto.

3.1. Investigación del Estado del Arte de Sistemas de Recomendación

En primer lugar, se debe investigar qué es lo que existe en la actualidad en relación con los modelos de recomendación, para así tener una base en la cual trabajar. Es por esto, que se deben estudiar en profundidad los tres métodos principales para hacer recomendaciones, los cuales son:

i. *Algoritmos en base a Filtros Colaborativos*
ii. *Algoritmos en base al Contenido*
iii. *Sistemas Híbridos*

Una vez teniendo en cuenta todas las formas distintas de encontrar recomendaciones para cada usuario, se debe investigar los casos en que se hayan implementado en algún *retailer*, para conocer situaciones de éxito y fracaso, y poder establecer una lista de requerimientos que deben ser tomados en cuenta al momento de utilizar un método de recomendación.

Finalmente, luego de tener conocimiento de cómo funcionan los algoritmos y cómo se han implementado en distintos *e-commerce*, se investigan casos de modelos de recomendación específicamente en empresas de cupones *online*, para comparar los sistemas estudiados con los que se han utilizado en empresas del mismo rubro de *Cuponatic*.

Con esto, se tendrá una perspectiva amplia de cómo se han realizado modelos de recomendación y qué es lo que se puede hacer de una nueva forma.

3.2. Análisis Descriptivo

Antes de analizar qué datos se van a utilizar para el modelo de recomendación, se debe hacer un completo análisis sobre los datos que posee la empresa. Es necesario estudiar qué variables son relevantes y, a la vez, qué información es útil para la realización de un sistema que genere recomendaciones personalizadas.

Es importante destacar, que se debe analizar todo tipo de información indiscriminadamente, ya que de aquí se pueden sacar conclusiones y nuevas ideas para la realización de un novedoso y creativo modelo.

3.3. Procesamiento de Datos

Una vez teniendo los requerimientos relevantes, concluidos de la etapa anterior, se procede a seleccionar los datos que dispone la empresa para utilizarlos en el sistema. Se hace una limpieza de ellos y, posteriormente, si es necesario, una transformación de estos.

3.3.1. Selección de Datos

La selección de los datos, proviene de la determinación de información relevante para utilizar en el modelo. Esto quiere decir que se deben utilizar todos los datos encontrados en la investigación del Estado del Arte que se ajusten al desarrollo de un modelo de recomendación.

Es importante hacer un equilibrio entre los requerimientos de los modelos estudiados con los requerimientos de la empresa, puesto que los datos proporcionados por esta pueden estar de distinta forma en una Base de Datos.

En otras palabras, se deben determinar los datos relevantes para el modelo (qué necesita el modelo, *input*) y qué datos son relevantes para la empresa para considerarlos al momento de hacer recomendaciones. Teniendo esto, se hace la selección y la extracción de los datos para utilizarlos en el modelo.

3.3.2. Limpieza de Datos

La limpieza de los datos se utiliza para que la información que sea utilizada por el modelo no posea errores o datos que no permitan que el modelo funcione de manera correcta.

En primer lugar, se deben eliminar bases de clientes que no contengan algún tipo de información, que por algún motivo no pudo guardarse. Esto ayuda a que el modelo no intente encontrar recomendaciones para clientes que no poseen toda la información necesaria.

Por otro lado, se deben definir criterios que permitan establecer qué información de la seleccionada es la que efectivamente se utilizará. Por ejemplo, para hacer recomendaciones sólo con información transaccional histórica, se requiere que los clientes hayan realizado al menos una compra, pero a la vez sólo una compra no es determinante para encontrar recomendaciones. Es por esto que se deben definir elementos específicos para utilizar en el modelo y quitar de la Base de Datos a los demás.

3.3.3. Transformación de Datos

Como se mencionó anteriormente, existen distintas formas de almacenar la información en una Base de Datos, por lo que en una pueden estar explícitamente los datos y en otra no, pero que sí se pueden identificar.

Por ejemplo, si se desea obtener la edad de los clientes, pero en la Base de Datos no existe una columna con esta información, y sí existe la fecha de nacimiento, es posible determinar la edad de estos restando la fecha actual con su fecha de nacimiento.

Por otro lado, es probable que para hacer recomendaciones, se necesiten varios datos de un producto o servicio para incorporar en el modelo, pero que deben agregarse en forma conjunta. Es por esto que se pueden crear nuevos *ID's* de ellos. Pongamos por caso que se desea agregar la categoría '10' a un producto '20', para tal efecto, se puede crear un ID '20-10', el cual será utilizado por el modelo como un *SKU* normal.

3.4. Generación de Modelo de Recomendación

Una vez completados los pasos anteriores, se conocerán todos los modelos existentes y se tendrá toda la información relevante para hacer un modelo. Por lo tanto, se debe elegir el o los modelos estudiados que más se ajusten a la empresa.

Al definir un modelo, se procederá a generar este mismo, programando en caso de ser necesario. Existen distintas implementaciones o herramientas que ya están disponibles para su uso, por lo que sólo bastaría adaptar los métodos a la empresa de cupones *online*.

Una vez definido el modelo, se deben preparar los *item sets* o, en otras palabras, lo que se le proporcionará al modelo para que haga recomendaciones (*input*), utilizando los datos transformados en la etapa anterior.

3.5. Diseño Experimental

Cuando el modelo de recomendación esté listo para ser empleado, se seleccionarán clientes para utilizarlos en los experimentos, dividiéndolos en grupos de control y grupos de tratamiento. Un grupo de control, es un segmento de clientes similares al resto de clientes, con el fin de observar el comportamiento de ellos en condiciones normales. Por otro lado, un grupo de tratamiento es una muestra similar al resto de los clientes, los que se utilizan para testear un cambio en las condiciones.

La distribución de ellos en cada grupo debe ser equivalente en sus datos. Por ejemplo, las cantidades de compra deben ser similares en ambos grupos, ya que si uno de ellos posee más compras que el otro, tendría mejores clientes, por lo que la experimentación se vería perjudicada. Por otro lado, es importante tomar en cuenta la cantidad de hombres y mujeres en estos grupos, ya que como estas últimas son más propensas a hacer una compra[1], una desproporción en los grupos podría afectar también la experimentación.

El modelo generará un *output*, que determinará qué productos y servicios se le deben recomendar a cada cliente, o bien, qué productos y servicios se deben recomendar luego

[1] Fuente: Base de Datos de la empresa.

de haber realizado otra compra. Es por esto, que se debe tomar esta información (*output*) y se debe utilizar para poder generar las recomendaciones personalizadas.

Luego, se debe diseñar cómo se hará el experimento. En este caso, al ser una modificación al *Newsletter* de de la empresa, se generará una plantilla (*template*) de este para ser utilizada para enviar recomendaciones personalizadas.

Así, se le enviará el *Newsletter* original al grupo de control y el Newsletter con recomendaciones al (los) grupos(s) de experimentación, para así contrastar los resultados en la etapa siguiente.

3.6. Evaluación de Resultados

En primer lugar, se esperan desarrollar correctamente dos modelos de recomendación en base a dos algoritmos distintos, utilizando información transaccional histórica de cada uno de los clientes.

Por otro lado, se espera diseñar una experimentación para poder probar cada uno de estos modelos. Para esto se deberán desarrollar distintos grupos de tratamiento y de control.

Una vez testeado el modelo de recomendación mediante el experimento, se deberán obtener los siguientes resultados:

1. Variación de la Tasa de Clics del *newsletter*, comparando el grupo de tratamiento con el grupo de control.
2. Variación de la Tasa de Apertura del *newsletter*, comparando el grupo de tratamiento con el grupo de control.
3. Variación de la Tasa de Conversión de clientes provenientes del *newsletter*, comparando el grupo de tratamiento con el grupo de control.
4. Variación de la cantidad de ventas provenientes del *newsletter*, comparando el grupo de tratamiento con el grupo de control.

Con estos resultados, se puede medir la efectividad del modelo, concluyendo qué indicadores son los que más se ven beneficiados con este.

Ilustración 3.1: Metodología del Proyecto

4. DESARROLLO METODOLÓGICO

4.1. ANÁLISIS DESCRIPTIVO

En esta sección, se describen los datos extraídos de la empresa con sus principales conclusiones y justificaciones de por qué se debe hacer un modelo de recomendación. En primer lugar se hará una descripción de cómo están compuestos los clientes que reciben estos correos por género. Posteriormente, se hará una comparación entre las compras por categorías realizadas por cada género, comparándolo con el formato predeterminado que posee el actual *newsletter*.

4.1.1. Comportamiento de Clientes por *Newsletter*

Actualmente, existen 341.594 clientes activos, los cuales reciben el *Newsletter* diario. Estos se dividen en tres listas: *Men*, *Women* y *Default*. La primera corresponde a la lista de hombres, la segunda corresponde a mujeres y la última, a suscritos que no tienen género declarado.

La lista *Men*, posee 110.737 suscritos, los cuales tienen un promedio de 2,06 compras cada uno. En la Ilustración 4.1, se puede observar que la mayoría de los clientes tiene cero compras, y la cantidad de clientes baja muy bruscamente cuando se aumenta una compra.

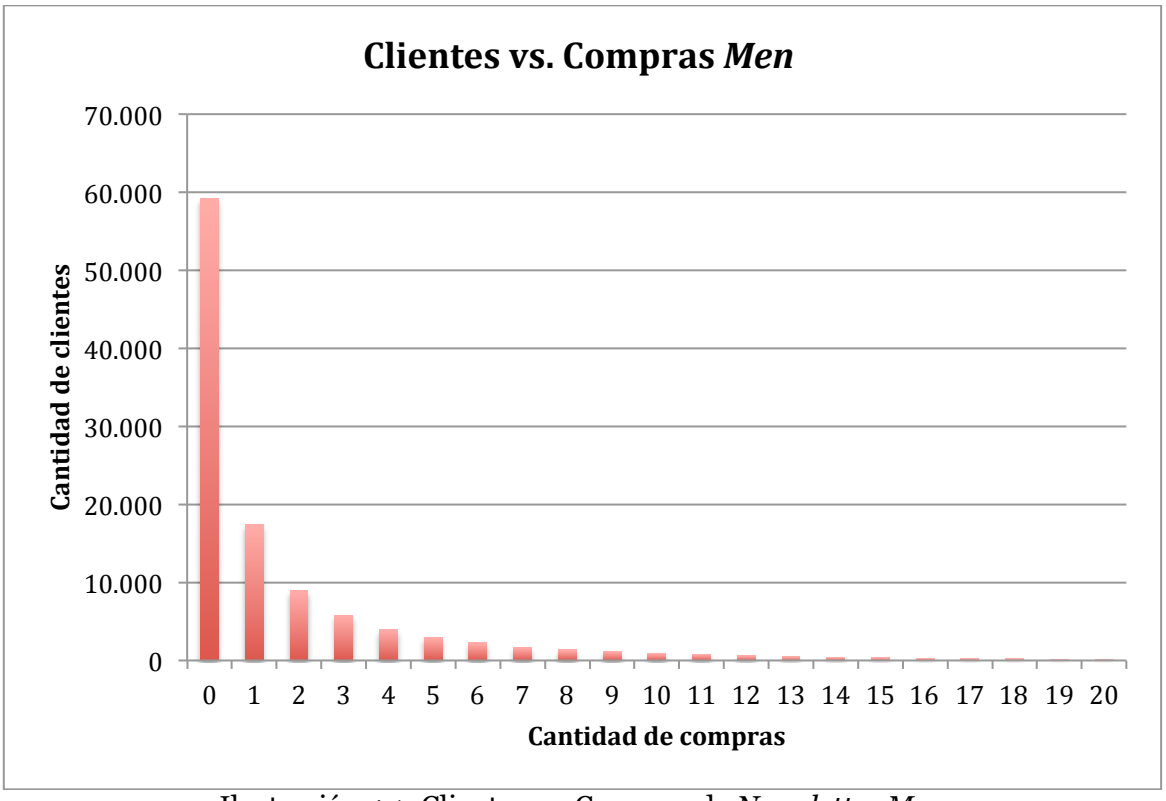


Ilustración 4.1: Clientes vs. Compras de *Newsletter Men*

La Tabla 4.1 muestra los indicadores actuales que posee actualmente la lista *Men*, donde, como se explicó en al Capítulo 1.2, hay un amplio margen en el cual se puede mejorar.

Indicador	Porcentaje
Open Rate	17,7%
Unique Open Rate	13,7%
CTR (enviados)	1,97%
CTR (abiertos)	14,3%
TCTR	14,4%
Tasa de Conversión	3,1%

Tabla 4.1: Cifras actuales de *Newsletter Men*

Las listas *Women* y *Default*, tienen un comportamiento similar, y estas pueden verse en el Anexo 4.

4.1.2. *Ticket* Promedio por Tipo de *Newsletter*

Para analizar el comportamiento histórico de los clientes que están registrados o suscritos en las distintas listas, es necesario saber cuál es el *ticket* promedio de inversión en productos o servicios.

De los 110.737 clientes pertenecientes a la lista *Men*, un 25,6% ha realizado alguna vez una compra. Observando la Ilustración 4.2, el 72,09% de ellos están dispuestos a comprar un cupón entre $0 y $10.000. Por otro lado, las ofertas mayores a $30.000 no representan a un segmento de clientes importante en comparación con el resto. Es por esto, que se debe analizar profundamente si es conveniente que se le ofrezcan cupones personalizados a este grupo que está dispuesto a comprar productos de mayor valor.

Las listas *Women* y *Default*, tienen un comportamiento similar, y estas pueden verse en Anexo 5.

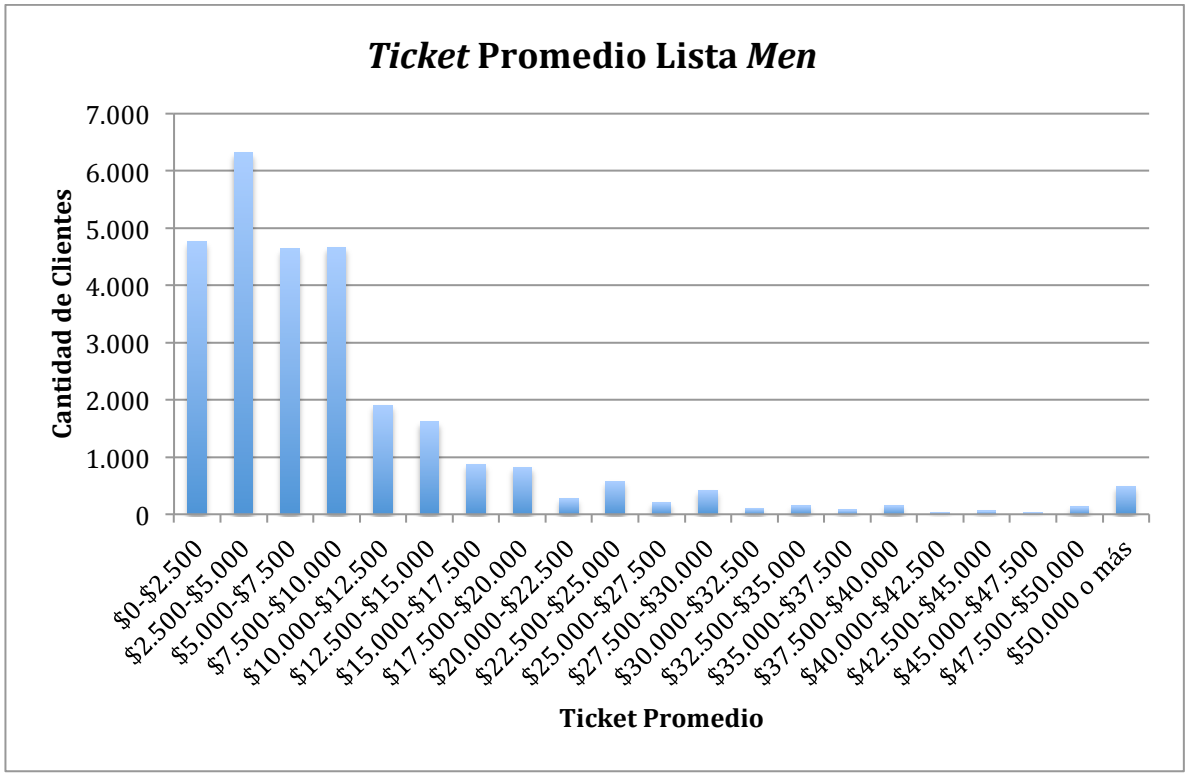


Ilustración 4.2: *Ticket* Promedio de lista *Men*

4.2. PROCESAMIENTO DE DATOS

En este capítulo, se muestran qué datos son los que pueden ser utilizados por el modelo de recomendación, explicando todo el proceso de selección, limpieza y su posterior transformación para ser utilizado por el sistema.

4.2.1. Selección de Datos

La selección de los datos se basó en encontrar la mayor cantidad de información que puede ser utilizada para generar recomendaciones. En este caso, se posee información de distintos ámbitos, de los cuales se extrajo lo más relevante para la generación de un modelo, y estos se explican a continuación.

i. Tabla de Clientes

ID Cliente	ID al cual se le asocia a un cliente
Nombre	Nombre del cliente
Correo Electrónico	Correo electrónico al cual recibe el *Newsletter* el cliente
Género	Género del cliente (Masculino o Femenino)
Fecha de Nacimiento	Fecha de nacimiento del cliente
Ciudad	Ciudad en la que reside el cliente

Tabla 4.2: Información por Cliente

ii. Tabla de Compras

ID Compra	ID asociado a la compra
ID Cliente	ID asociado a un cliente
ID Descuento	ID asociado al descuento comprado
Fecha Compra	Fecha en la que el cliente realizó la compra
Monto del ticket	Cuanto pagó el cliente al hacer la compra
Fuente	Cómo ingresó al sitio (directo, *google*, *newsletter*)

Tabla 4.3: Información por Compra

iii. Tabla de Encuestas

ID Encuesta	ID asociado a la encuesta
ID Compra	ID asociado a la compra
Calificación Compra	Calificación o *Rating* del descuento comprado
Comentarios Compra	Comentarios del descuento comprado

Tabla 4.4: Información por Encuesta

iv. Tabla de Descuentos

ID Descuento	ID asociado al descuento
ID Producto	ID asociado a un producto (nulo en caso de servicio)
ID Categoría	ID de la categoría asociada al descuento
ID Subcategoría	ID de la subcategoría asociada al descuento
Fecha Inicio Descuento	Fecha en la cual comienza el descuento
Fecha Término Descuento	Fecha en la cual termina el descuento

Tabla 4.5: Información por Descuento

v. Tabla de Categorías

ID Categoría	ID asociado a la categoría
Nombre Categoría	Nombre de la categoría
ID Subcategoría	ID asociado a la subcategoría
Nombre Subcategoría	Nombre de la subcategoría

Tabla 4.6: Información por Categoría

vi. Tabla de Productos

ID Producto	ID asociado al producto
Nombre Producto	Nombre del producto
ID Subcategoría 1	ID de la subcategoría de primer nivel asociada al producto
Nombre Subcategoría 1	Nombre de la subcategoría de primer nivel asociada al producto
ID Subcategoría 2	ID de la subcategoría de segundo nivel asociada al producto
Nombre Subcategoría 2	Nombre de la subcategoría de segundo nivel asociada al producto
ID Subcategoría 3	ID de la subcategoría de tercer nivel asociada al producto
Nombre Subcategoría 3	Nombre de la subcategoría de tercer nivel asociada al producto

Tabla 4.7: Información por Producto

vii. Tabla de Empresas

ID Empresa	ID vinculado a la empresa
Nombre Empresa	Nombre de la empresa

Tabla 4.8: Información por Empresa

4.2.2. Limpieza de Datos

Una vez que se hayan seleccionado los datos, reunidos los correos de cada cliente, las compras, las fechas de compra, las calificaciones, el género, la edad, entre otros, se debe hacer una limpieza de los datos de la siguiente forma:

i. Eliminar de la base de datos a los empleados de la empresa.

Se realizó esta limpieza, ya que se observó que personas que trabajan en la empresa, utilizan esta plataforma para hacer la mayoría de las compras, por lo que utilizar a estos mismos clientes para hacer recomendaciones, pueden confundir al modelo.

ii. Eliminar de la base de datos a todos los clientes con 2 o menos compras o intentos de compra.

Para poder hacer recomendaciones en base a las transacciones históricas de los clientes, se necesitan usuarios que tengan un mínimo de transacciones, o de lo contrario, no se podrían generar. Para esto, se definió que se iba a enfocar en clientes con un mínimo de tres compras históricas, con lo que se obtiene una base de datos de más de 95.000 clientes.

iii. Eliminar todas las filas de datos que contengan celdas vacías.

Una vez que se han obtenido cerca de 95.000 filas de información, se debe verificar qué datos no se encuentran disponibles o qué datos se encuentran en blanco. Todas estas filas se deben eliminar para que el modelo pueda utilizar todos los datos disponibles. Así, los datos finales muestran 90.222 filas, las que corresponden a esa cantidad de clientes con tres o más compras o intentos de compras.

4.2.3. Transformación de los Datos

Para utilizar estos datos como *input* para el modelo que se realizará, se deben crear criterios para tener información equivalente en el caso de productos o servicios, y en el caso de categorías. Es importante señalar que los datos explicados a continuación son falsos y solo contribuyen a explicar el proceso al que se sometieron.

En primer lugar, existe un *ID* asociado a los productos que se venden en un cupón de descuento, el cual se puede vender en más de un cupón; pero no existe un *ID* asociado a un servicio. Es por esto que se procedió a generar *ID's* para los servicios publicados en los cupones. Éstos, fueron creados concatenando los ID*'s* de la empresa y de la subcategoría que está asociada al cupón. Por ejemplo, "Belleza Spa", tiene un *id_empresa* = 8234, y ofrece un cupón correspondiente a la subcategoría "Tratamientos reductivos" de *id_subcategoría* = 2324, por lo tanto, el *id_producto* sería 8234-2324, logrando así identificar futuros cupones que se publiquen utilizando esta empresa y la misma subcategoría.

Por otro lado, para los productos existe un nivel de subcategoría de tres estratos, pero para servicios sólo existe un nivel, es por esto que para lograr identificar las categorías como si fuesen iguales se determinaron de la siguiente manera:

i. Categorías de Productos

A los productos se les asignó en primer lugar un 1 al principio, que corresponde a la categoría de "Productos", luego se le concatena el *ID* de la subcategoría de tercer nivel y luego el *ID* de la subcategoría de primer nivel. Por ejemplo, para un producto (*id* = 1), correspondiente la subcategoría de tercer nivel "*iPod* o *MP3*" (*id_subcategoría3* = 543) y correspondiente a una subcategoría de primer nivel "Música (*id_subcategoría1* = 10), quedaría con un *id_categoría* "1-543-10".

ii. Categorías de Servicios

Para los servicios, el método fue similar, ya que a todos los servicios se le asignó el número 2, luego el *ID* de la categoría y luego el *ID* de la subcategoría. Por ejemplo, para un servicio (*id* = 2), correspondiente a "Entretención" (*id_categoría* = 3) y correspondiente a "Conciertos" (*id_subcategoría* = 23), quedó con un *id_categoría* = "2-3-23"

4.3. GENERACIÓN DEL MODELO

En el Capítulo 2, se mencionaron todos los algoritmos y sistemas existentes en la actualidad, y que son parte del Estado del Arte. Sin embargo, se debe seleccionar qué tipos de algoritmos o modelos se deben utilizar para una empresa con una amplia variedad de categorías y con una alta rotación de ofertas por cada una de ellas.

Como se mencionó en ese capítulo, existen tres principales sistemas de recomendación: los *Filtros Colaborativos*, los *Sistemas en base al Contenido* y los *Sistemas Híbridos*. Los *Filtros Colaborativos* utilizan la información transaccional histórica, las calificaciones de ellas, las ofertas vistas, entre otros aspectos, y la empresa posee alguna de esta información, como el historial de compras y algunas calificaciones, pudiendo así realizar un modelo basado en estos datos. Por otro lado, los *Sistemas en base al Contenido* no pueden utilizarse de manera correcta, ya que las características que posee cada oferta no son suficientes para encontrar similitudes entre ellas y, además, las descripciones que se publican en cada oferta están limitadas a cada producto y no siempre existen similitudes basadas en ellas. Finalmente, los *Sistemas Híbridos* tampoco pueden ser utilizados, ya que por definición corresponden a una combinación entre los *Filtros Colaborativos* y los *Sistemas en base al Contenido*, donde este último no se puede llevar a cabo correctamente.

Como fue mencionado anteriormente, sólo se hará un modelo basado en *Filtros Colaborativos*, pero recordando en qué consisten estos métodos, se pueden dividir en dos tipos distintos: los basados en la memoria y los basados en un modelo. Por lo anterior, se llevará a cabo un modelo de cada uno de estos, para posteriormente comparar resultados con el fin de establecer cuál es más efectivos de los dos.

En esta sección, se explicará cómo se llevaron a cabo los modelos, mencionando las razones de por qué se utilizaron ciertos algoritmos y por qué no se utilizaron otros.

4.3.1. Selección de Algoritmos de Filtros Colaborativos

Existen diversos tipos de Filtros Colaborativos en base a la Memoria y en base a un Modelo. Los principales métodos explicados en el Capítulo 2, correspondientes a algoritmos basados en la memoria, son los basados en los productos y los basados en los usuarios, y los principales métodos basados en un modelo, son los modelos bayesianos, las segmentaciones y las reglas de asociación. Todas ellas fueron explicadas anteriormente.

Antes de decidir qué tipos de algoritmos utilizar, es importante mencionar la información en que se basarán los modelos para hacer recomendaciones. Como se explicó anteriormente, existen datos de información transaccional histórica y calificaciones de las compras realizadas. Sin embargo, las calificaciones poseen un bajo porcentaje del total, por lo que solo el 11% de las compras realizadas por todos los clientes poseen una calificación. Es por esto que se decidió utilizar solo la información transaccional, ya que con tan solo una pequeña cantidad de calificaciones, las

recomendaciones podrían estar cargadas a ciertos productos que no tienen tanta relación como otros.

En primer lugar, se debe definir qué tipo de algoritmo de filtros colaborativos basados en la memoria conviene utilizar: uno que identifique similitudes entre productos, o bien, similitudes entre usuarios. Anteriormente se señaló que se deben utilizar algoritmos para encontrar similitudes entre productos cuando la cantidad de productos es menor a la de los usuarios, y similitudes entre usuarios cuando estos son menores que la cantidad de productos ofertados. Sin embargo, esta empresa de cupones *online*, no funciona de la misma manera que el común de los *e-commerce*, ya que las ofertas tienen un tiempo limitado (promedio de 5 días) y se tienen más de 900 ofertas diarias. Por esta razón, al considerar la cantidad de usuarios activos que reciben diariamente el correo con el total de ofertas que históricamente se han publicado, no existe una clara diferencia de qué algoritmo utilizar, por lo que se decidió utilizar un algoritmo que detecte similitudes entre usuarios, determinando así clientes con gustos similares para recomendarles lo que ellos han comprado. Esta decisión se tomó principalmente porque se quiere obtener ofertas lo más personalizadas posibles, ya que detecta un número determinado de ofertas que son mucho más probables que el otro método.

Luego, se debe definir qué tipo de Filtros Colaborativos basados en un modelo se debe utilizar. Como se ha referido anteriormente, solo se ocuparán los datos transaccionales y no las calificaciones, por lo que no se necesita predecir valores en casos de incertidumbre. De este modo, se descartan los modelos y redes bayesianas, y, como la empresa se basa en compras impulsivas, existe mucha variabilidad en lo que se compra, por lo que al hacer segmentos, no se hacen grupos definidos. Por lo anteriormente expuesto, se decidió utilizar Reglas de Asociación, las cuales, sobre la base de las transacciones históricas de los clientes, generan patrones de compra que definen qué se le debería recomendar a cada cliente dada la compra de cierto producto.

A continuación, se describen en detalle cada uno de los modelos creados y cómo se hizo cada uno, tomando en cuenta todas las suposiciones necesarias.

4.3.2. Filtros Colaborativos basados en el Usuario

Una vez que se decidió utilizar los Filtros Colaborativos basados en el Usuario, se tuvo que definir qué tipo de similitud o qué distancia entre usuarios se debe efectuar en consideración de los datos que posee la empresa.

Para recordar las distancias entre usuarios o los indicadores de similitud explicados en el Capítulo 2.2.1, son mencionados a continuación:

i. *Distancia Euclidiana*
ii. *Distancia Euclidiana Cuadrática*
iii. *Distancia Manhattan*
iv. *Distancia Coseno*
v. *Distancia Jaccard*
vi. *Coeficiente de Correlación de Pearson*
vii. *Correlación de Spearman*

viii. *Similitud Log-verosimilitud*

Cuando estos fueron abordados, se describieron los escenarios en que se deberían utilizar de la forma correcta. En particular, debido a que sólo se utilizarán los datos transaccionales históricos para realizar el modelo, se deben tomar en cuenta las distancias o similitudes que mejor se ajusten a estos datos, la Distancia Jaccard y la Similitud Log-verosimilitud, ya que las demás funcionan de mejor manera utilizando calificaciones de las compras históricas.

Ahora bien, conviene definir cuál de estas dos opciones es la más recomendable para utilizarla en un modelo. Para esto, se decidió testear estadísticamente cada una de ellas utilizando algunas de las métricas de evaluación explicadas en el Capítulo 2: *Precision* y *Recall*. Para esto, se consideraron los 90.222 clientes que posee la base de datos con tres o más compras históricas o intentos de compra. Esta muestra se separó aleatoriamente en grupos compuestos de la siguiente forma: 30% de control y 70% de tratamiento. Así, se obtuvieron los siguientes resultados:

i. *Distancia Jaccard*:

$$Precision = 8{,}33\%$$
$$Recall = 8{,}33\%$$

ii. *Similitud Log-verosimilitud*:

$$Precision = 13{,}33\%$$
$$Recall = 12{,}5\%$$

De acuerdo con lo anterior, podemos concluir que la mejor opción entre estas dos métricas de similitud entre usuarios, es la *Similitud Log-verosimilitud*. A pesar de que estos valores son pequeños, una precisión de un 13,33% es buena, ya que esto solo determina cuáles de las ofertas recomendadas por el modelo, el usuario ya ha comprado anteriormente, pero puede ser que éste nunca había visto estas ofertas.

Por otro lado, es importante definir la cantidad de usuarios similares que se tomarán en cuenta para hacer las recomendaciones. Como se explicó anteriormente, existen dos formas de definir esta cantidad: una determinada por una distancia o una similitud específica y otra con un número fijo de usuarios similares independientes de sus distancias. En este caso particular, se decidió utilizar esta segunda opción, ya que permitía ofrecer productos y servicios en una cantidad mucho más similar a todos, contrario a lo que sucedería si fuera definida por un cierto valor de similitud. Por ejemplo, cuando se determina que se deben encontrar todos los usuarios similares que tengan una distancia menor a 0,3, se pueden encontrar cien clientes parecidos, o bien, cero. La cantidad fija de usuarios similares utilizados por el modelo es de 20, ya que fue la cantidad que arrojó la *Precision* y el *Recall*.

Para explicar de una mejor manera cómo funcionan los filtros colaborativos basados en el usuario, se consideran los siguientes clientes con sus respectivas compras, empleando la forma “cliente, compra”:

Juan, Mochila
Juan, Pendrive
Juan, Parlante
Diego, Mochila
Diego, Audífonos
Diego, Pendrive
Diego, Taza
Pablo, Funda Notebook
Pablo, Chocolates
Pablo, Audífonos
Pablo, Bebidas

Para buscar el usuario más similar, se genera la siguiente tabla:

	Mochila	Pendrive	Parlante	Audífonos	Taza	Funda Notebook	Chocolates	Bebidas
Juan	X	X	X	-	-	-	-	-
Diego	X	X	-	X	X	-	-	-
Pablo	-	-	-	X	-	X	X	X

Tabla 4.9: Ejemplo de Similitud Log-Verosimilitud

Luego, se calcula la similitud *Log-verosimilitud* (véase Capítulo 2.2) para cada uno de ellos. De esta forma, se obtiene que el usuario más similar a Juan es Diego, con una similitud de -0,22, frente a -∞, que es la similitud de Juan con Pablo (es importante recordar que mientras más cercano a cero es el valor, más similares son los usuarios). Así, las recomendaciones para Juan, considerando sólo un usuario similar, serían los productos que Diego ha comprado y que Juan aún no, vale decir, los Audífonos y la Taza. Este proceso es análogo para todos los usuarios.

4.3.3. Reglas de Asociación

Para hacer las Reglas de Asociación, es importante definir bajo qué método se harán, *Apriori* o *FP-Growth*. Como se describió en el Capítulo 2, el método *FP-Growth* es mucho más rápido y utiliza menos memoria. Sin embargo, dicho método requeriría hacerlo en *RapidMiner*, mientras que *Apriori* se puede realizar utilizando los paquetes *arules* y *arulesviz* del programa *R*. Estos paquetes, dan como resultado las mismas reglas que el método *FP-Growth*, pero tienen una opción para visualizar las reglas de una mejor forma que en *RapidMiner*. Es por esto, que se decidió utilizar el método *Apriori* en el programa *R*, donde el código utilizado se puede ver en Anexo 6.

Este modelo, se decidió hacer sobre la base de categorías y no de productos y servicios específicos, para así poder enfocar distintas visiones en dos métodos distintos.

Para llevar a cabo estas reglas, es necesario determinar los valores de *Soporte* (*Support*), *Confianza* (*Confidence*) y *Elevación* (*Lift*), que fueron explicados anteriormente. Se quieren generar las suficientes reglas para hacer recomendaciones para 199 categorías,

por lo que se fue probando con distintos números de *confianza*, desde 0,8 hasta 0,2, hasta generar la mayor cantidad de reglas distintas y que a la vez sean coherentes. Así, el *soporte* mínimo para la creación de reglas se fijó en 35 personas y la *confianza*, en un 0,2. Esto quiere decir que la regla debe estar compuesta en su totalidad por al menos 35 clientes distintos y que, a partir de todas las personas que compraron el antecedente de una regla, el 0,2 de ellos (o el 20%) debe haber comprado también la consecuencia. La *elevación* se calcula utilizando los *soportes* de las reglas y los *soportes* de los antecedentes y consecuencias por separado (ver Capítulo 2 para más detalles), y predice qué tanto más probable es que una persona compre una categoría compuesta por una regla específica que en condiciones normales. En consecuencia, las reglas generadas con una mayor *elevación*, son consideradas mejores, y es por esta razón que se ordenan de acuerdo con este indicador. El número total de reglas generadas en la base de datos de la empresa, fue de 740.

En la Ilustración 4.7, se puede apreciar un gráfico que muestra la concentración de reglas respecto de su *soporte*, su *confianza* y su *elevación*. Se observa que, a medida que el *soporte* es menor, la *elevación* aumenta (se torna más oscuro), lo que significa que encuentra pequeñas relaciones de grupos reducidos que son más propensos a comprar ciertas categorías por haber comprado otras. Un ejemplo de estas reglas es la siguiente:

Regla	Soporte	Confianza	Elevación
{Postres - Despensa} → {Abarrotes - Despensa}	0,0006	0,5	15,8042

En este caso, se puede ver que es una regla que está en un el extremo izquierdo inferior del gráfico, con un tono oscuro (por la alta elevación).

Las reglas que están más lejanas a la concentración, tienen un *soporte* y una *confianza* mayor, pero una *elevación* baja, lo que se explica mediante la popularidad de las categorías pertenecientes a este segmento o, en otras palabras, quiere decir que no es mucho más probable que los clientes las compren debido a estas reglas, sino que también lo es que las compren debido a su popularidad. Un ejemplo de éstas se muestra a continuación:

Regla	Soporte	Confianza	Elevación
{Entretención - Paseos /Aventuras/ Tours} → {Entretención - Cine}	0,07244	0,5703	1,596

En este ejemplo, se puede apreciar que el soporte y la elevación son totalmente contrarias al ejemplo anterior, que corresponde al pequeño grupo de reglas de el extremo derecho superior del gráfico.

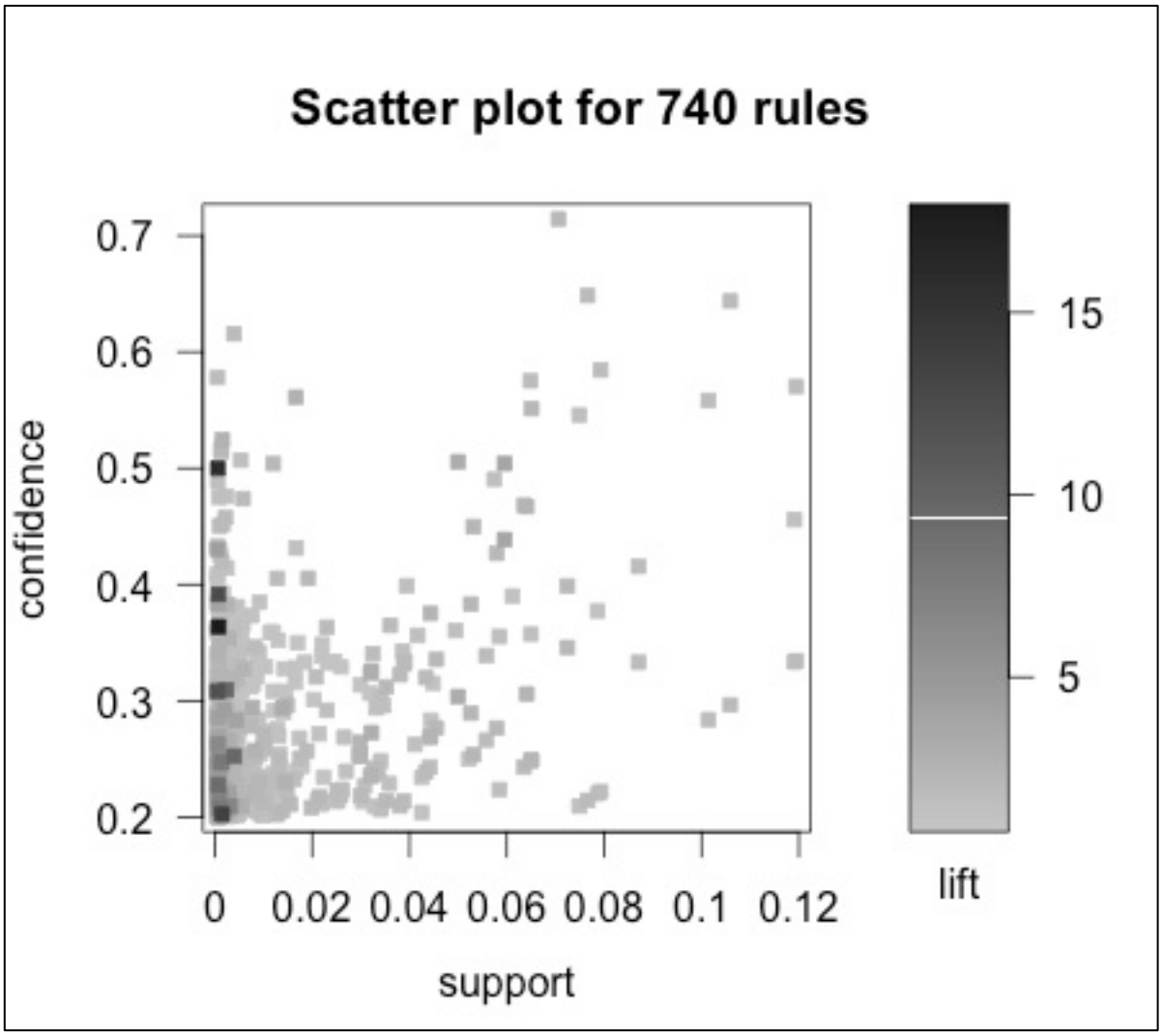


Ilustración 4.3: Gráfico de Dispersión de Reglas de Asociación

Para observar de otra manera algunas de las reglas generadas por el modelo, se muestra la Ilustración 4.8, donde se pueden apreciar las 20 reglas con más *elevación*. En este gráfico agrupado se observa la relación entre el *soporte* y la *elevación*, donde el tamaño del círculo corresponde a la cantidad de *soporte* (mientras más *soporte*, más grande el círculo) y donde la *elevación* está dada por el color (mientras más oscuro, más elevación). Es importante destacar que el eje horizontal (*LHS*, *Left Hand Side*), corresponde al antecedente de la regla; y el eje vertical (*RHS*, *Right Hand Side*), a la consecuencia de la regla. Esto quiere decir que las reglas son de la forma *LHS* → *RHS*.

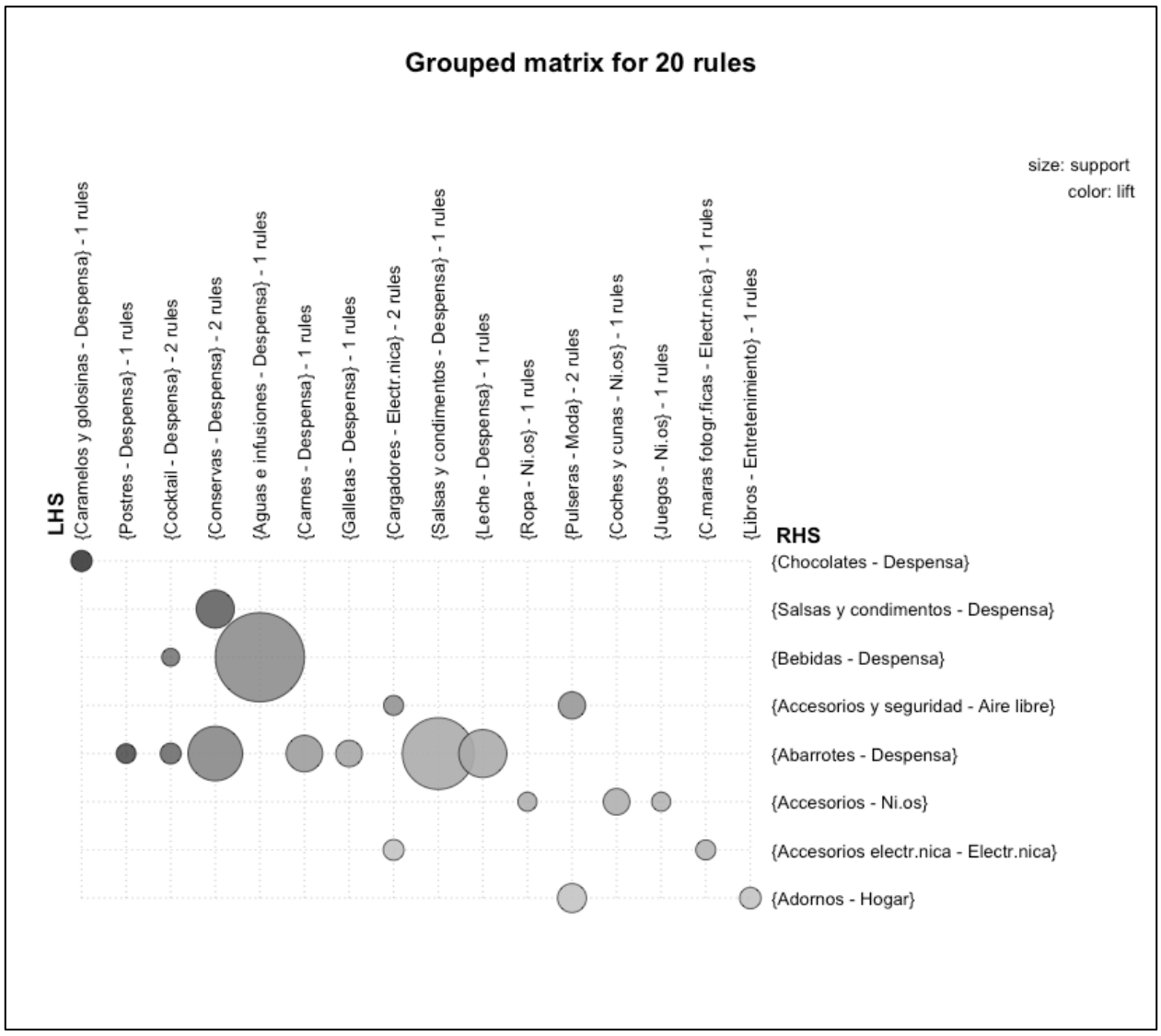


Ilustración 4.4: Gráfico agrupado de Reglas de Asociación

Para apoyar la explicación de cómo funcionan las reglas de asociación aplicadas a una empresa de cupones *online*, podemos apreciar el siguiente ejemplo de canastas de compras (o transacciones de compras históricas) por categorías:

1) Alimentos, Tratamientos Reductivos, Bebidas
2) Tratamientos Reductivos, Tratamientos Tonificantes, Tratamientos Faciales
3) Tratamientos Tonificantes, Tratamientos Reductivos, Tratamientos Faciales
4) Alimentos, Bebidas, Comida Peruana, Comida Rápida
5) Tratamientos Reductivos, Tratamientos Tonificantes, Comida Peruana

Dado este historial, correspondiente a cinco clientes, se pueden generar las siguientes reglas ordenadas por la *elevación*, como se puede observar en la Tabla 4.12.

Reglas	Soporte	Confianza	Elevación
Alimentos → Bebidas	0,4	1	2,5
Bebidas → Alimentos	0,4	1	2,5
Tratamientos Reductivos → Tratamientos Tonificantes	0,6	1	1,6
Tratamientos Reductivos → Tratamientos Faciales	0,4	0,6	1,6

Tabla 4.10: Ejemplos de Reglas de Asociación

Entonces, una persona que compró la categoría "Alimentos" compraría con una probabilidad de 2,5 veces más alta la categoría "Bebidas" que en una persona en condiciones normales, por lo que a todos los que hayan comprado esta categoría se le debería recomendar la segunda.

Como este modelo genera reglas en forma de categorías, y se busca recomendar productos y servicios específicos, se toman las ofertas más vendidas hasta el momento de cada una de las categorías activas y se recomiendan. Esto se hace sobre la base de cada una de las compras realizadas por cada cliente. Entonces, tomando el mismo ejemplo anterior, una persona que compre un "Alimento" se le recomendará el producto más vendido en la categoría de "Bebidas".

Algunos ejemplos de reglas generadas utilizando la base de datos de la empresa, se pueden observar en la Tabla 4.4.

Rules	*Support*	*Confidence*	*Lift*
{Caramelos y golosinas - Despensa} => {Chocolates - Despensa}	42	0,36	17,89
{Postres - Despensa} => {Abarrotes - Despensa}	36	0,5	15,8
{Conservas - Despensa} => {Salsas y condimentos - Despensa}	90	0,2	13,47
{Cocktail - Despensa} => {Abarrotes - Despensa}	42	0,39	12,38
{Cocktail - Despensa} => {Bebidas - Despensa}	35	0,31	11,61
{Conservas - Despensa} => {Abarrotes - Despensa}	138	0,31	9,78
{Aguas e infusiones - Despensa} => {Bebidas - Despensa}	240	0,25	9,51
{Cargadores - Electrónica} => {Accesorios y seguridad - Aire libre}	35	0,23	9,04
{Pulseras - Moda} => {Accesorios y seguridad - Aire libre}	60	0,21	8,22
{Carnes - Despensa} => {Abarrotes - Despensa}	84	0,25	7,8

Tabla 4.11: Ejemplo de Reglas de Asociación aplicadas a la Base de Datos de la Empresa

4.4. DISEÑO EXPERIMENTAL

En la capítulo anterior, se explicó cómo se llevó a cabo el modelo, la teoría que está detrás de él y cómo se va a utilizar con los datos recopilados. Sin embargo, fuera de lo teórico y de lo que estadísticamente resulta, se debe testear en la realidad. Es por esto que se realizan experimentos para probar si, efectivamente, el modelo generado funciona correctamente y si da resultados positivos con respecto a los indicadores expuestos en el Capítulo 1.3: la *Tasa de Clics*, la *Tasa de Apertura*, la *Tasa de Conversión* y las *Ventas*.

A continuación, se explica qué clientes son los utilizados en la experimentación, qué distintos grupos se crean y cómo se realiza esta.

4.4.1. Clientes para Experimentación

En primer lugar, para realizar experimentos y testear efectivamente si el modelo generado funciona correctamente, es necesario probarlo en la realidad. Es por esto que se decidió hacer experimentos con clientes que tengan suficiente historial de transacciones.

Existen 90.222 clientes que poseen 3 compras o intentos de compras en la empresa, a partir de mayo de 2014. Se decidió utilizar esta fecha, ya que a partir de aquí se comenzó a guardar la información por productos y no solamente por oferta. Esto quiere decir que un producto puede estar en varias ofertas, y una oferta puede ser distinta a otra a pesar de que el producto es el mismo. Esto se debe a que como la empresa vende cupones de descuentos, estos tienen una duración corta, por lo que al publicar de nuevo el cupón, el *ID* de la oferta es distinta.

De esta cantidad de clientes, se hizo una limpieza respecto de quienes están actualmente activos en el *Newsletter* de la empresa, es decir, quienes de ellos actualmente están recibiendo el correo todas las mañanas. Es importante cruzar esta información, ya que si se le envía un correo a clientes que no están activos o que cancelaron su suscripción anteriormente, el correo puede caer en *spam* y afectaría la reputación[2] de la empresa. Así, la cantidad de clientes que se consideró se redujo a aproximadamente 60.000.

Posteriormente, tomando en cuenta que se harían 10 grupos distintos en la experimentación (los cuales serán explicados en el subcapítulo siguiente), se decidió utilizar aproximadamente 2.000 clientes en cada uno. Por lo tanto, la cantidad de clientes final que se seleccionó es de 20.507, conformada por:

- 13.325 clientes de género femenino, correspondiente a un 64,98% del total de ellos.
- 7.182 clientes de género masculino, correspondiente a un 35,02% del total de ellos.
- El promedio de compras o intentos de compras de ellos es de 5,38, donde el promedio de compras de género masculino es de 5,39 y el de género femenino, es de 5,38.

[2] Los principales administradores de correo electrónicos, como *Hotmail*, *Gmail* y *Yahoo*, poseen un estándar de reputación, por lo que si esta es inferior a un 95%, automáticamente trasladan el correo entrante al buzón de *Correo No Deseado*.

4.4.2. Grupos de Control y de Experimentación

Como se mencionó anteriormente, se harán 10 distintos grupos para realizar la experimentación. En esta sección, se justifican las decisiones tomadas respecto de la cantidad de grupos y la definición y conformación de estos.

En primer lugar, es necesario definir qué es lo que se quiere estudiar y cómo se evaluará; en este caso, la *personalización* y la *recomendación*. La primera se estudia a partir de la *Tasa de Apertura* y la segunda, a partir de la *Tasa de Clics*, la *Tasa de Conversión* y las *Ventas*.

Es por esto, que se decidió utilizar un experimento de dos factores. El primer factor corresponde a la personalización, medible en dos niveles, si es un asunto personalizado o no; y el segundo factor es la recomendación, la cual tiene cinco niveles: el correo original y cuatro distintos tipos de recomendación. Para ver claramente cómo se definen estos factores en conjunto con sus niveles, se muestra la Ilustración 4.9.

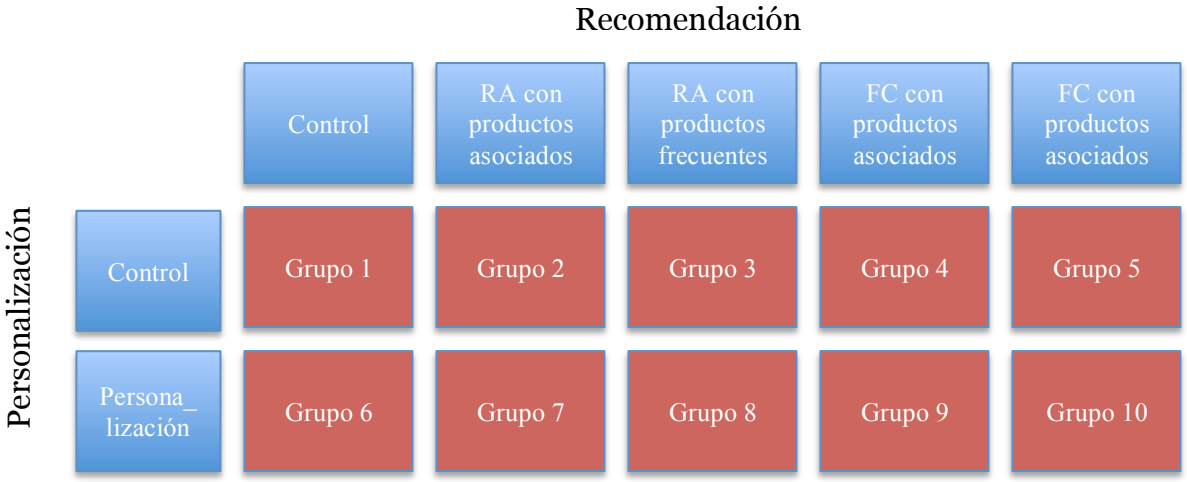


Ilustración 4.5: Diseño Experimental

Para explicar de mejor manera esta tabla, se debe mencionar que el eje vertical corresponde a la personalización del *Asunto (Subject)* del correo y el eje horizontal alude al tipo de recomendación que se emplea. Por ejemplo, el Grupo 8 está conformado por un *Asunto* personalizado y por recomendaciones elaboradas por el modelo de Reglas de Asociación con Productos Frecuentes.

Los distintos grupos están conformados por clientes y cantidades similares. A continuación se muestra cómo están compuestos cada uno de ellos.

i. Cantidad de Clientes:

La cantidad de clientes por cada grupo, está conformada según el siguiente cuadro:

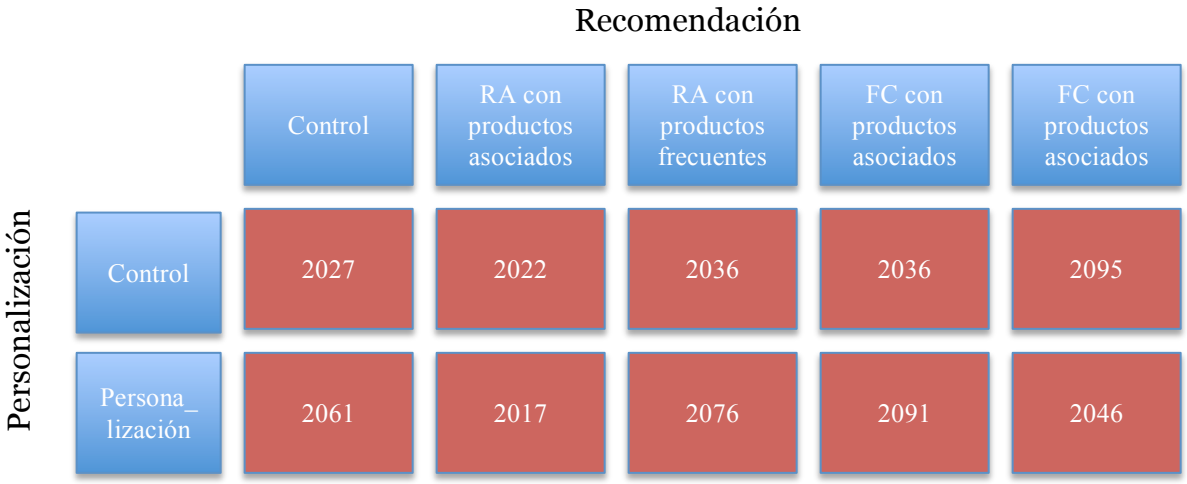


Ilustración 4.6: Cantidad de Clientes por Grupo de Experimentación

ii. Cantidad Promedio de Compras por Cliente:

La cantidad promedio de compras o intentos de compra por cada uno de los clientes asociados a cada grupo, se muestra a continuación:

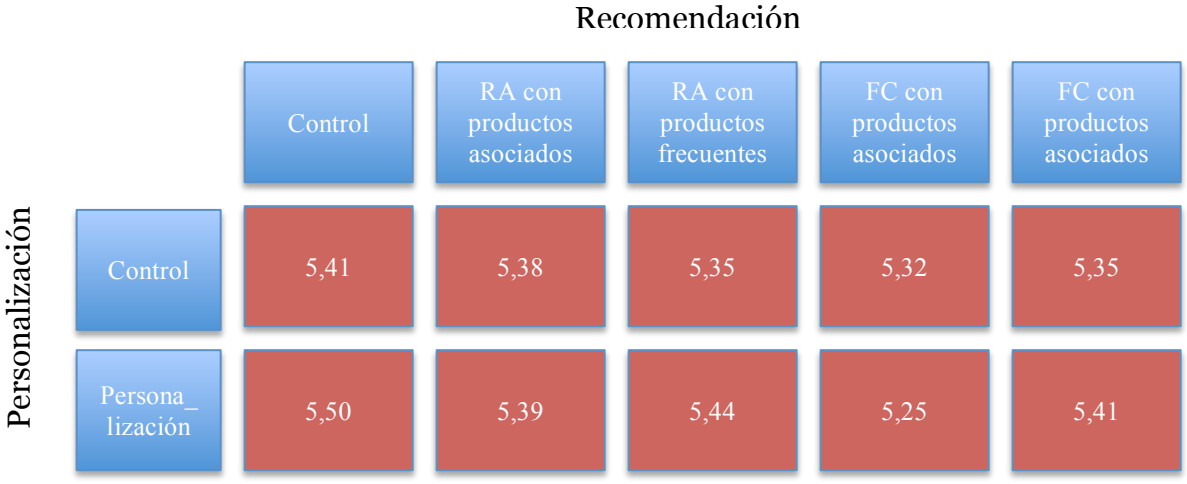


Ilustración 4.7: Promedio de Compras por Grupo de Experimentación

iii. Cantidad de Clientes Masculinos:

La cantidad de clientes de género masculino por cada grupo, junto con su porcentaje del total, se muestra en el siguiente cuadro:

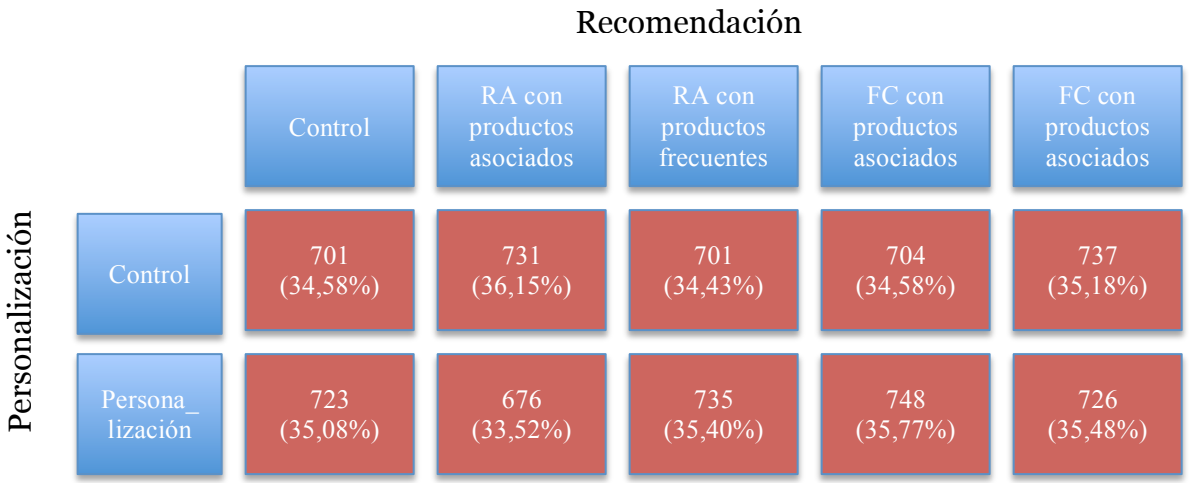


Ilustración 4.8: Cantidad de Clientes Masculinos por Grupo de Experimentación

iv. Cantidad de Clientes Femeninos:

La cantidad de clientes de género femenino por cada grupo, acompañada del porcentaje del total, se muestra en el siguiente cuadro:

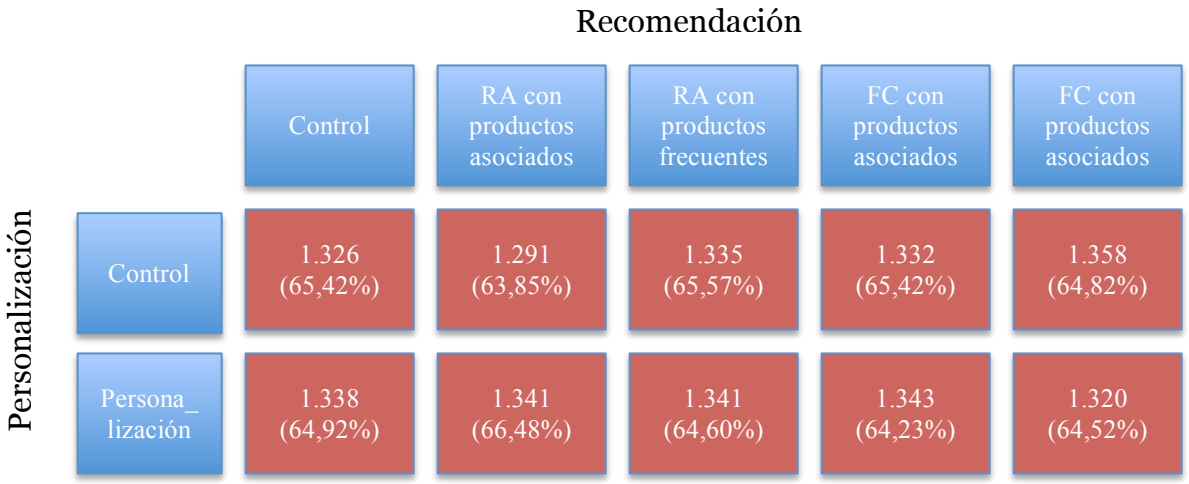


Ilustración 4.9: Cantidad de Clientes Femeninos por Grupo de Experimentación

4.4.3. Diseño de Experimentación

En la sección anterior, se definieron los distintos grupos que se iban a tratar y qué tipo de experimentación se le aplicaría a cada uno. Ahora bien, es necesario definir los parámetros de cómo se realizará el experimento en sí, es decir, cómo será el formato y cómo se aplicarán los cambios al correo para que la experimentación sea testeada correctamente.

El *Newsletter* original que se envía todas las mañanas a los clientes activos, está compuesto por 33 ofertas y el asunto del correo corresponde al título de las tres primeras ofertas del correo. En la Ilustración 4.14, se puede apreciar cómo ve el correo cada cliente.

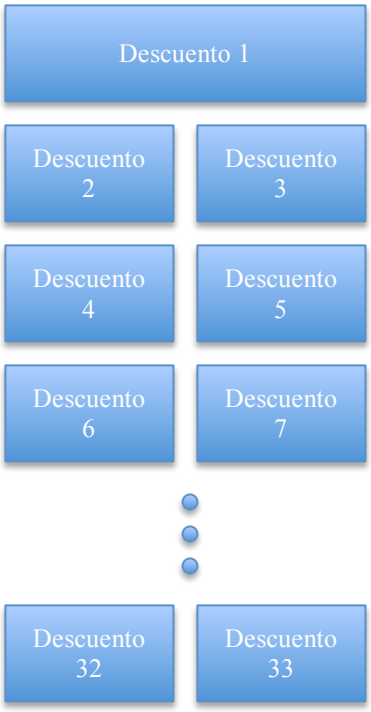


Ilustración 4.10: Estructura Original de Newsletter

Entonces, para testear la personalización del correo y qué efecto tiene en la apertura de este, se define un *asunto control* y un *asunto personalizado*. Por ejemplo:

i. Asunto Control:

"✪ 100 Fotos 13x18 Kodak Express $8.900 ◤◥ Toldo Plegable $34.990 ≃ Ictioterapia + Reiki + Alineación de Chakras $8.990"

ii. Asunto Personalizado:

"[Nombre], ¡Disfruta de los mejores descuentos que tenemos para ti!"

Estos asuntos son los que diferencian las celdas 1-5 (asunto control) de las celdas 6-10 (asunto personalizado), por lo que se testea como un total de ellos. Cabe destacar que los asuntos expuestos anteriormente son solo ejemplos, ya que estos varían día a día.

Por otro lado, para testear las recomendaciones generadas por cada uno de los modelos, es necesario definir la estructura del correo. El *Newsletter* original tiene una estructura de la forma expuesta en la Ilustración 4.15, es decir, tiene un orden específico de acuerdo con las ofertas destacadas. Es por esto, que el grupo de control (celdas 1 y 6) tendrán el mismo orden que el correo original.

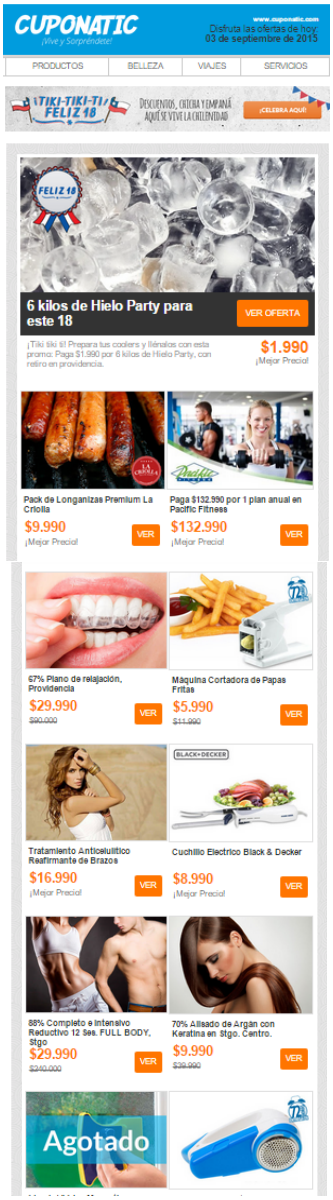


Ilustración 4.11: Vista Original de Newsletter Original

Luego, se tiene que definir cómo se mostrarán las recomendaciones de los siguientes grupos. Estos tendrán la misma estructura que el *Newsletter* original, pero diferenciando el orden en cómo se muestran las ofertas. En la Ilustración 4.16, se muestra un cliente que posee cinco recomendaciones. Como se puede observar, se

muestran en primer lugar las recomendaciones y luego se sigue con el orden original del correo. Esta estructura se utilizará para todos los grupos, exceptuando las celdas 1 y 6.

Ilustración 4.12: Estructura de Newsletter con Recomendaciones

Ahora bien, se debe diferenciar si los clientes hacen clic o compran un producto de los recomendados o de los productos destacados originales. Por esta razón, se ocupará *Google Analytics* para poder identificar si la persona hizo clic en un producto recomendado o no, y se guardará en la base de datos si un producto comprado fue recomendado o no.

Esta experimentación se hará tres veces con una frecuencia de una vez por semana. Una vez que se tengan todos los resultados de los indicadores mencionados en el Capítulo 1.3, se analizarán mediante un Test de Proporciones de Poblaciones se estudiará la significancia de la Tasa de Apertura, de la Tasa de Clics y de la Tasa de Conversión y mediante un Test de ANOVA se estudiará la significancia de las Ventas, para después concluir si los modelos producen o no un cambio significante en los clientes. Se harán estos tests distintos para cada una de las métricas ya que sólo se tiene información desagregada para las ventas.

4.5. Evaluación de Resultados

En esta sección, se darán a conocer los resultados de los distintos envíos que se hicieron del experimento, para luego analizar y comparar cada uno de los tratamientos y poder observar y concluir qué método o qué modelo es el más conveniente, según los distintos escenarios, ya sea para aumentar la venta de la empresa, o bien, aumentar el tráfico de su sitio *web*.

4.5.1. Experimentos Realizados

Para analizar los resultados obtenidos de los distintos experimentos realizados, se analizará la cobertura de las recomendaciones sobre el total de las ofertas publicadas, el desempeño global de cada uno de los grupos, acorde a cada métrica descrita en los objetivos, y, por último, la significancia de cada uno de estos resultados.

i. Primer Envío

El primer experimento se envió el jueves 22 de octubre de 2015 a las 08.00 hrs. En tal ocasión, los *Asuntos* originales del correo fueron los siguientes:

- Masculino:

  ♥ Pack Excite Preservativos LifeStyles $7.990 ☐ 6 Meses Criolipólisis $44.990 ☻ Limpieza Dental Full $7.990

- Femenino:

  ☐ Lipocavitación Garantizado Menos Cms $12.990 ☼ Depilación IPL Axila + Bozo $10.990 ☐ 20 Sesiones Tonificante Abdomen + Glúteos $7.990

- *Default*:

  ♥ Pack Excite Preservativos LifeStyles $7.990 ☐ 20 Sesiones Tonificante Abdomen + Glúteos $7.990 ★ Primera Capa Hombre $10.990

El *Asunto* utilizado para medir la personalización de este experimento fue el siguiente:

- Personalización:

  [Nombre], ¡Disfruta de los descuentos que tenemos para ti!

Como se dijo anteriormente, la hipótesis bajo este factor de experimentación es que los clientes se sentirán más atraídos cuando las ofertas son personalizadas a cada una de las personas, por lo que uno se tienta a abrir el correo.

Por otro lado, se tienen cuatro distintos métodos para obtener recomendaciones, los que fueron explicados en el Capítulo 4.4. La cobertura de cada uno de ellos sobre el total de las ofertas del sitio *web* es la siguiente:

Modelo de Recomendación	% del Total
Reglas de Asociación con Productos Asociados	2,01%
Reglas de Asociación con Productos Frecuentes	10,3%
Filtros Colaborativos con Productos Asociados	34,6%
Filtros Colaborativos con Productos Frecuentes	47,2%

Tabla 4.12: Cobertura de Modelos en Primer Envío

Como se puede observar, el modelo que más cobertura otorga es el de Filtros Colaborativos con Productos Frecuentes, recomendando 446 ofertas de los 945 productos disponibles en esta fecha.

El promedio de recomendaciones por personas de los distintos modelos se observa en la siguiente tabla:

Modelo de Recomendación	Promedio de Recomendaciones
Reglas de Asociación con Productos Asociados	3,55
Reglas de Asociación con Productos Frecuentes	7,01
Filtros Colaborativos con Productos Asociados	4,77
Filtros Colaborativos con Productos Frecuentes	3,31

Tabla 4.13: Promedio de Recomendaciones por Cliente en Primer Envío

El modelo que más recomendaciones genera por cada uno de los clientes, es el de Reglas de Asociación con Productos Frecuentes, sobrepasando en una gran cantidad a los otros modelos.

ii. Segundo Envío

El segundo experimento se envió casi una semana después, el día miércoles 28 de octubre de 2015 a las 08.00 hrs. Los *Asuntos* originales del *Newsletter* fueron los siguientes:

- Masculino:

  ☼ Pack 3 Productos Australian Gold $17.990 ▣ Alfombra Shaggy $21.990 ☐ Tonifica Abdomen y Pectorales $10.990

- Femenino:

  ☐ Pack 4 Productos Emuline Facial $13.990 ☼ Pack 3 Productos Australian Gold $17.990 ☐ Levantamiento Glúteos + Anticelulítico $18.990

- *Default*:

  ☐☐ Paquetes de Emuwipes Premium $18.790 ☼ Pack 3 Productos Australian Gold $17.990 ☐ Tratamiento Control de Peso Formoline $24.990

El *Asunto* utilizado para medir la personalización de este segundo experimento fue el siguiente:

- Personalización:

  [Nombre], ¡Mira los descuentos que tenemos para ti!

Al igual que en el envío anterior, también se experimentó con los cuatro distintos modelos de recomendación, donde la cobertura de las ofertas recomendadas se expone en la siguiente tabla:

Modelo de Recomendación	% del Total
Reglas de Asociación con Productos Asociados	2,4%
Reglas de Asociación con Productos Frecuentes	11,02%
Filtros Colaborativos con Productos Asociados	36,2%
Filtros Colaborativos con Productos Frecuentes	48,4%

Tabla 4.14: Cobertura de Modelos en Segundo Envío

Por otro lado, la cantidad promedio de recomendaciones enviadas a cada cliente se puede observar a continuación:

Modelo de Recomendación	Promedio de Recomendaciones
Reglas de Asociación con Productos Asociados	4,98
Reglas de Asociación con Productos Frecuentes	9,72
Filtros Colaborativos con Productos Asociados	4,52
Filtros Colaborativos con Productos Frecuentes	5,92

Tabla 4.15: Promedio de Recomendaciones por Cliente en Segundo Envío

Se obtienen las mismas conclusiones que en el primer envío, donde Filtros Colaborativos con Productos Frecuentes tiene una mayor cobertura de las ofertas, y las Reglas de

Asociación con Productos Frecuentes tiene un mayor promedio de recomendaciones por cada uno de los clientes.

iii. Tercer Envío

El tercer experimento se envió el viernes 13 de noviembre de 2015 a las 08:00 hrs., dos semanas a partir del último. Los *Asuntos* originales del correo diario eran los siguientes:

- Masculino:

  ☐ Potenciador Sexual MaxMan $12.990 ☐ Termoventilador Mural Bano Valory $13.990 ☐ Cuerpo de Gladiador $39.990

- Femenino:

  ☐ Termoventilador Mural Baño Valory $13.990 ☐ 20 Sesiones Tonificante $8.990 ♣ Set de Picnic para 2 $15.990

- *Default*:

  ☐ Termoventilador Mural Baño Valory $13.990 ☐ Potenciador Sexual MaxMan $12.990 ☐ 20 Sesiones Tonificante $8.990

El *Asunto* utilizado para medir la personalización de este segundo experimento fue el siguiente:

- Personalización:

  [Nombre], ¡Descubre los mejores descuentos que tenemos para ti!

Al igual que los envíos anteriores, también se utilizaron los cuatro distintos modelos para hacer las recomendaciones. La cobertura de cada uno de ellos se puede observar en la siguiente tabla:

Modelo de Recomendación	% del Total
Reglas de Asociación con Productos Asociados	2,3%
Reglas de Asociación con Productos Frecuentes	12,47%
Filtros Colaborativos con Productos Asociados	30,6%
Filtros Colaborativos con Productos Frecuentes	41,9%

Tabla 4.16: Cobertura de Modelos en Tercer Envío

Además, la cantidad promedio de recomendaciones asociadas a cada usuario, se puede advertir a continuación:

Modelo de Recomendación	Promedio de Recomendaciones
Reglas de Asociación con Productos Asociados	4,07
Reglas de Asociación con Productos Frecuentes	8,24
Filtros Colaborativos con Productos Asociados	3,29
Filtros Colaborativos con Productos Frecuentes	4,51

Tabla 4.17: Promedio de Recomendaciones por Cliente en Tercer Envío

A pesar de que las cifras varíen entre uno y otro envío, se puede observar que el mismo patrón se cumple: los Filtros Colaborativos con Productos Frecuentes supera a los distintos modelos, y por otro lado, las Reglas de Asociación con Productos Frecuentes superan con creces a la cantidad promedio de recomendaciones.

4.5.2. Resultados de Personalización

En la sección anterior, se obtuvo como resultado que los Filtros Colaborativos con Productos Frecuentes son las que más ofertas recomiendan de las disponibles y, que los Reglas de Asociación con productos Frecuentes es la que más genera recomendaciones en promedio por cada uno de los clientes.

En esta sección, se analizarán los resultados obtenidos a partir del factor de Personalización, donde se estudiará si hay una diferencia en el comportamiento de los clientes cuando el trato es personalizado.

Según la Ilustración 4.9, se tienen 10 grupos distintos para los cuales se les hicieron tres envíos de experimentos. Para medir el factor de personalización, se juntaron las filas de la figura, con el propósito de medir el efecto completo de la personalización del *Asunto* del *Newsletter*. En la Ilustración 4.17, se pueden observar los nuevos grupos.

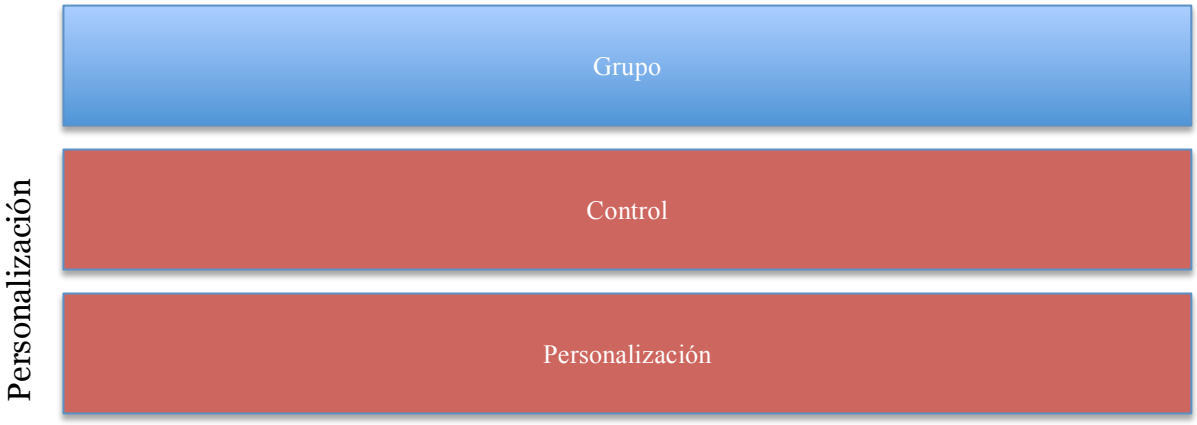


Ilustración 4.13: Grupos por Factor Personalización

En primer lugar, es posible advertir los resultados para la Tasa de Apertura (*Open Rate* en inglés), los cuales tienen un patrón común en todos los envíos, en los que, en un

mayor porcentaje, el correo fue abierto por clientes a quienes se les envió un *Asunto* personalizado con su nombre.

Para denotar si la diferencia de cada uno de los grupos es significativo (Tabla 4.18 a Tabla 4.49), se utilizará la siguiente nomenclatura:

*** Coeficiente significativo al nivel p < 0,01
** Coeficiente significativo al nivel p < 0,05
* Coeficiente significativo al nivel p < 0,1

En caso de que la diferencia no sea significativa, no tendrá un símbolo adherido.

i. Primer Envío

Grupo	Tasa de Apertura (%)
Control	22,19%
Personalización	23,34% (+ 1,15%)

Tabla 4.18: Tasa de Apertura en Primer Envío (Factor Personalización)

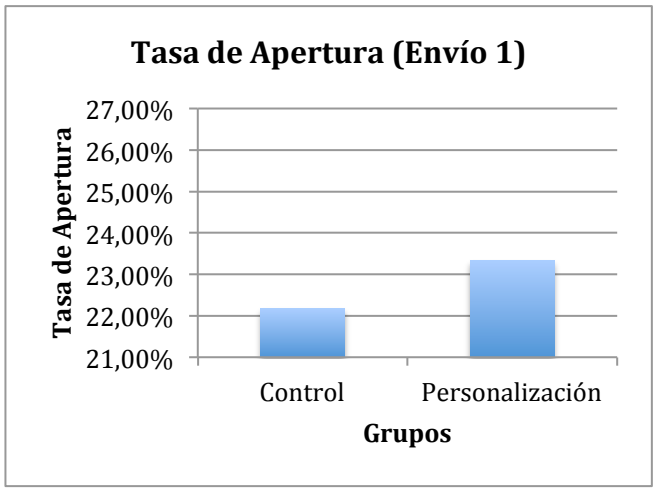


Ilustración 4.14: Tasa de Apertura en Primer Envío (Factor Personalización)

ii. Segundo Envío

Grupo	Tasa de Apertura (%)
Control	25,33%
Personalización	26,79% (+ 1,46%)*

Tabla 4.19: Tasa de Apertura en Segundo Envío (Factor Personalización)

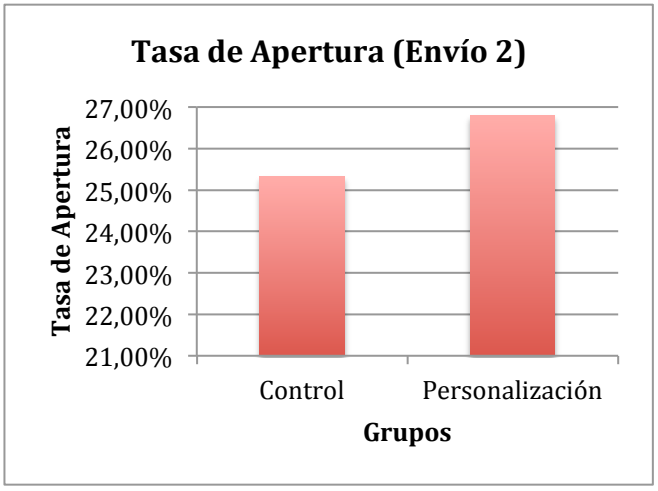


Ilustración 4.15: Tasa de Apertura en Segundo Envío (Factor Personalización)

iii. Tercer Envío

Grupo	Tasa de Apertura (%)
Control	25,08%
Personalización	26,36% (+ 1,28%)*

Tabla 4.20: Tasa de Apertura en Tercer Envío (Factor Personalización)

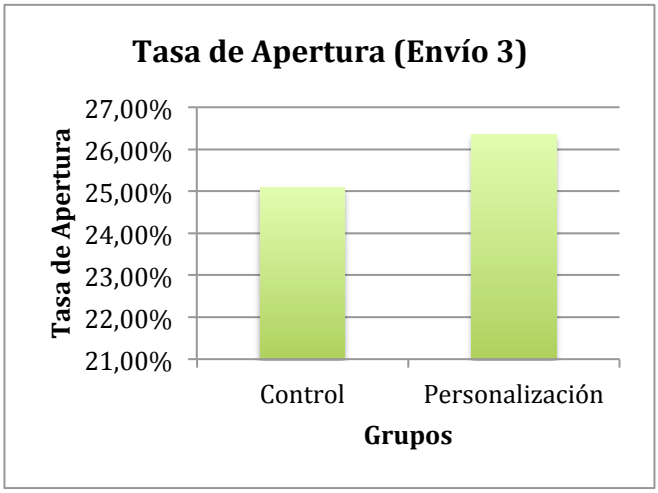


Ilustración 4.16: Tasa de Apertura en Tercer Envío (Factor Personalización)

Teniendo este patrón entre los resultados, se puede decir que, en promedio, la personalización es un factor que puede hacer que los clientes tengan una *Tasa de Apertura* mayor. En la Tabla 4.23 y en la Ilustración 4.21, se puede observar el resultado promedio de esta métrica, donde la diferencia entre los porcentajes, equivale a aproximadamente 4.500 clientes más que abren el correo diariamente.

Grupo	Tasa de Apertura Promedio (%)	Desviación Estándar (%)
Control	24,20%	1,75%
Personalización	25,50% (+ 1,3%)*	1,88%

Tabla 4.21: Tasa de Apertura Promedio (Factor Personalización)

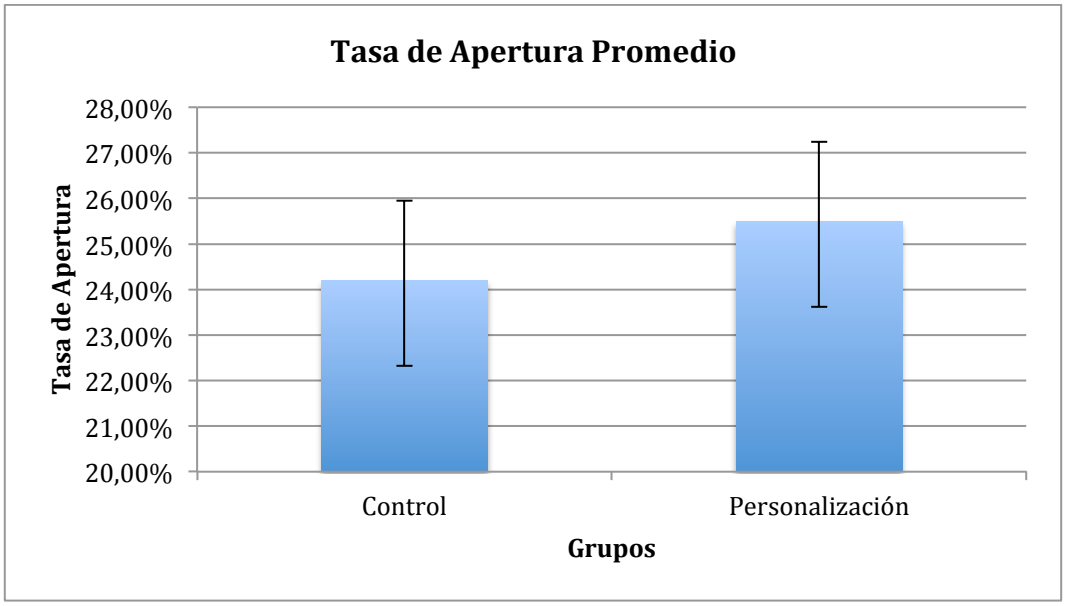


Ilustración 4.17: Gráfico de Tasa de Apertura Promedio (Factor Personalización)

Pues bien, como estos porcentajes tienen una diferencia no muy grande, es necesario poder calcular la significancia de esta. Utilizando el Test de Proporciones de Poblaciones, se puede concluir que esta diferencia no es significativa con un 95% de confianza, pero sí con un 90%, por lo que este resultado sí puede adaptarse para mejorar el tráfico en el *Newsletter* diario.

Por otro lado, se quiere probar si la personalización del *Asunto* afecta el comportamiento de los clientes al ver el correo independientemente del contenido de este. Así, se pueden observar los siguientes resultados para la Tasa de Clics de cada uno de los envíos.

i. Primer Envío

Grupo	Tasa de Clics (%)
Control	18,08%
Personalización	18,51% (+ 0,43%)

Tabla 4.22: Tasa de Clics en Primer Envío (Factor Personalización)

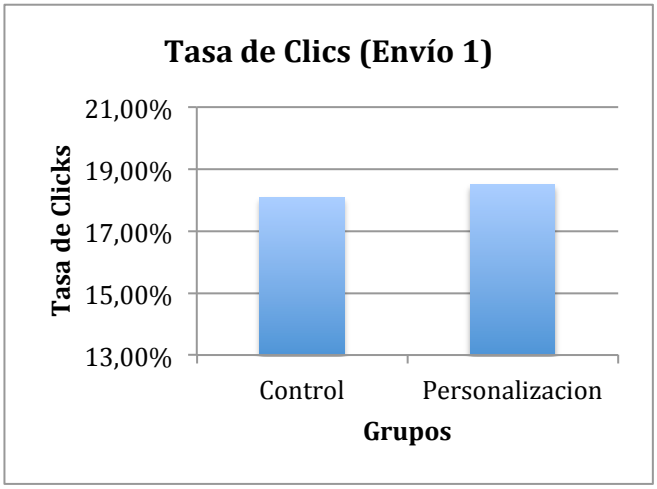


Ilustración 4.18: Tasa de Clics en Primer Envío (Factor Personalización)

ii. Segundo Envío

Grupo	Tasa de Clics (%)
Control	15,04%
Personalización	19,49% (+ 4,45%)***

Tabla 4.23: Tasa de Clics en Segundo Envío (Factor Personalización)

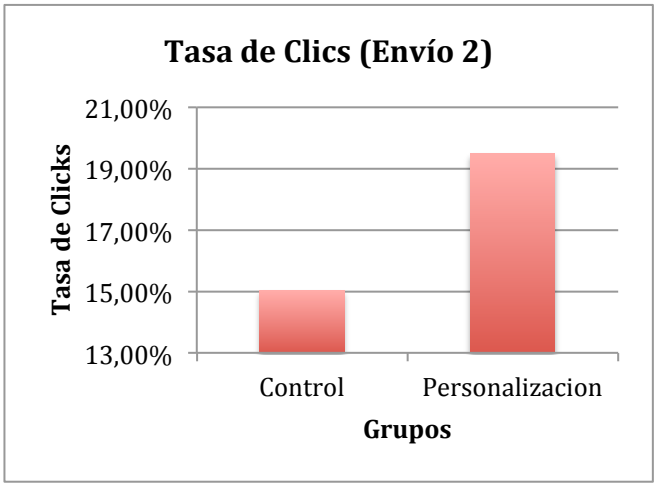


Ilustración 4.19: Tasa de Clics en Segundo Envío (Factor Personalización)

iii. Tercer Envío

Grupo	Tasa de Clics (%)
Control	16,97%
Personalización	17,34% (+ 0,37%)

Tabla 4.24: Tasa de Clics en Tercer Envío (Factor Personalización)

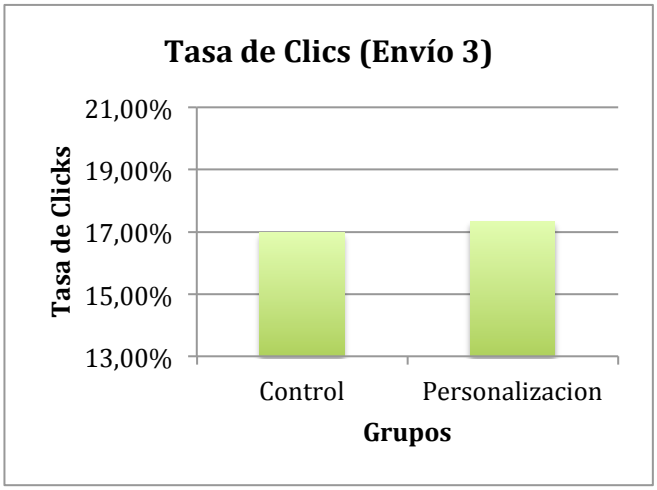


Ilustración 4.20: Tasa de Clics en Tercer Envío (Factor Personalización)

Como se puede apreciar en las tablas y gráficos anteriores, existe un patrón en el comportamiento de los clientes cuando el trato es personalizado. Si bien, solamente el segundo envío es el que tiene una gran diferencia, cerca de 4,5%, los demás envíos también fueron superiores. Para estudiar si este patrón es determinante, se calculará un promedio de ellos para observarlo más claramente.

Grupo	Tasa de Clics Promedio (%)	Desviación Estándar (%)
Control	16,70%	1,54%
Personalización	18,45% (+ 1,75%)**	1,07%

Tabla 4.25: Tasa de Clics Promedio (Factor Personalización)

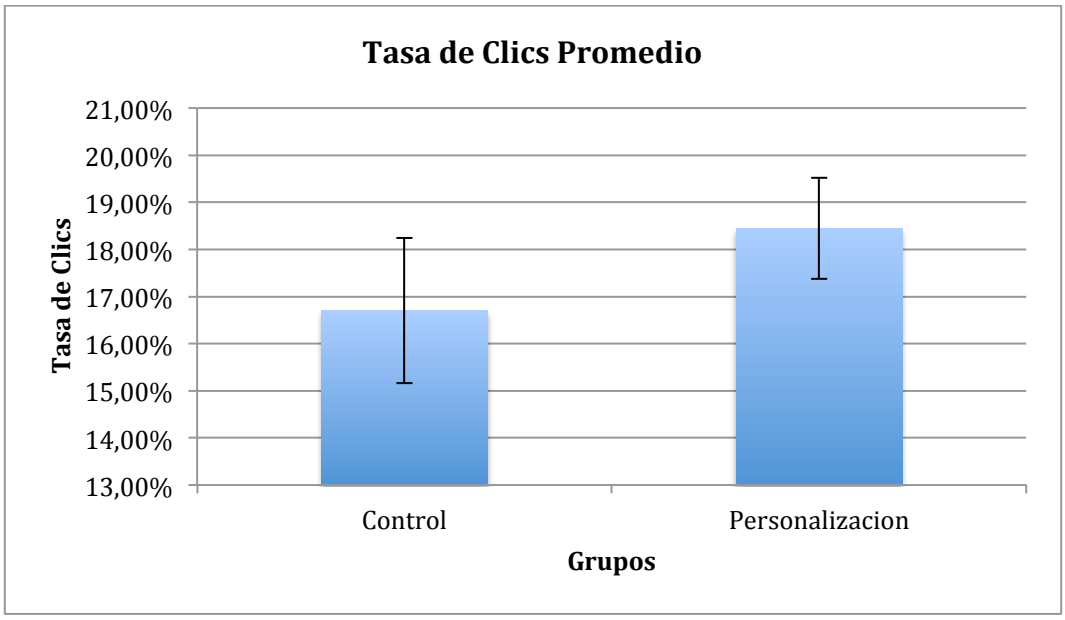


Ilustración 4.21: Gráfico de Tasa de Clics Promedio (Factor Personalización)

En el gráfico anterior, se puede observar que en promedio la Tasa de Clics aumentó en un 1,65%. Esta cifra, llevándola a toda la base de datos, corresponde a cerca de 1.600 clientes adicionales que ingresarían al sitio *web* por día. Ahora bien, es importante calcular la significancia de este resultado, donde también se realizó un Test de Proporciones de Población, en el que, con un 95% de confianza, la personalización del *Asunto* aumenta la Tasa de Clics.

Además, se quiere concluir si la personalización afecta el comportamiento de los clientes una vez que ellos hacen clic o muestran interés en una oferta, mediante la compra de estas. Es por esto que se observó la cantidad de Ventas que se hicieron en cada uno de los envíos en ambos grupos, las cuales se muestran a continuación.

i. Primer Envío

Grupo	Cantidad de Ventas
Control	14
Personalización	11 (- 3)

Tabla 4.26: Cantidad de Ventas en Primer Envío (Factor Personalización)

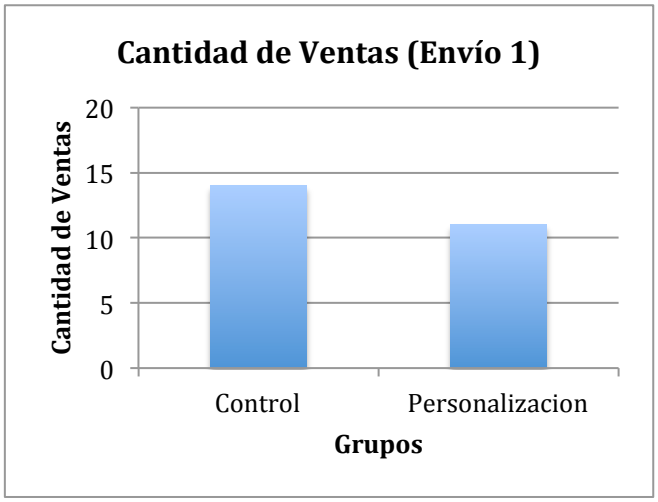


Ilustración 4.22: Cantidad de Ventas en Primer Envío (Factor Personalización)

ii. Segundo Envío

Grupo	Cantidad de Ventas
Control	15
Personalización	16 (+ 1)

Tabla 4.27: Cantidad de Ventas en Segundo Envío (Factor Personalización)

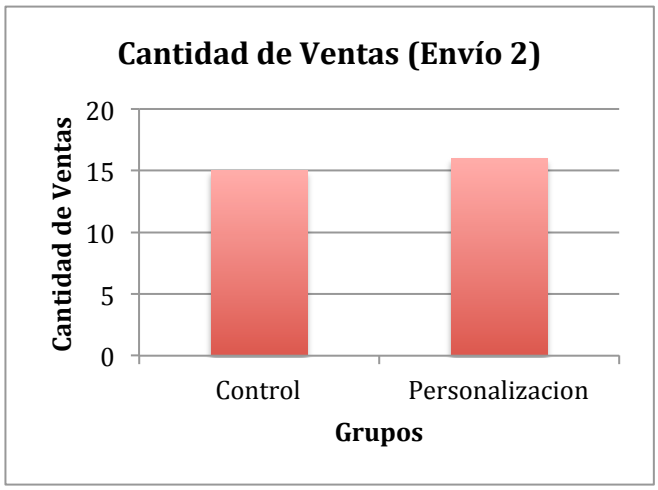


Ilustración 4.23: Cantidad de Ventas en Segundo Envío (Factor Personalización)

iii. Tercer Envío

Grupo	Cantidad de Ventas
Control	18
Personalización	7 (- 11)

Tabla 4.28: Cantidad de Ventas en Tercer Envío (Factor Personalización)

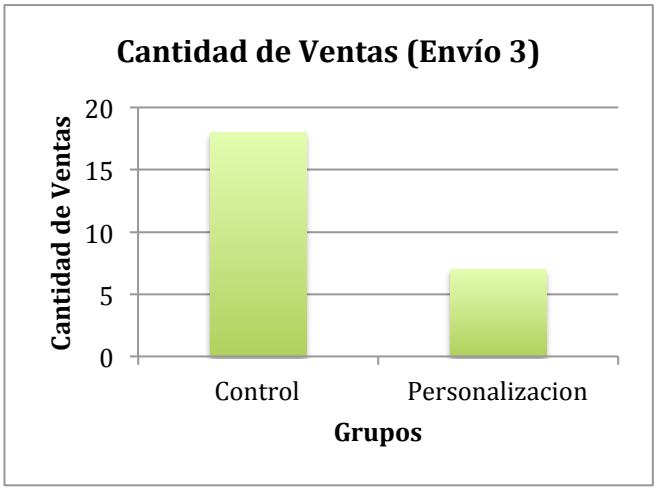


Ilustración 4.24: Cantidad de Ventas en Tercer Envío (Factor Personalización)

En relación con los tres envíos, no se puede determinar un patrón en cantidades de ventas de acuerdo con el trato personalizado. A continuación, se muestran las cantidades de ventas en promedio para poder observar si en realidad existe un cambio en el comportamiento respecto de este factor.

Grupo	Cantidad de Ventas Promedio	Desviación Estándar
Control	15,67	2,08
Personalización	11,33 (- 4,34)	4,51

Tabla 4.29: Cantidad de Ventas Promedio (Factor Personalización)

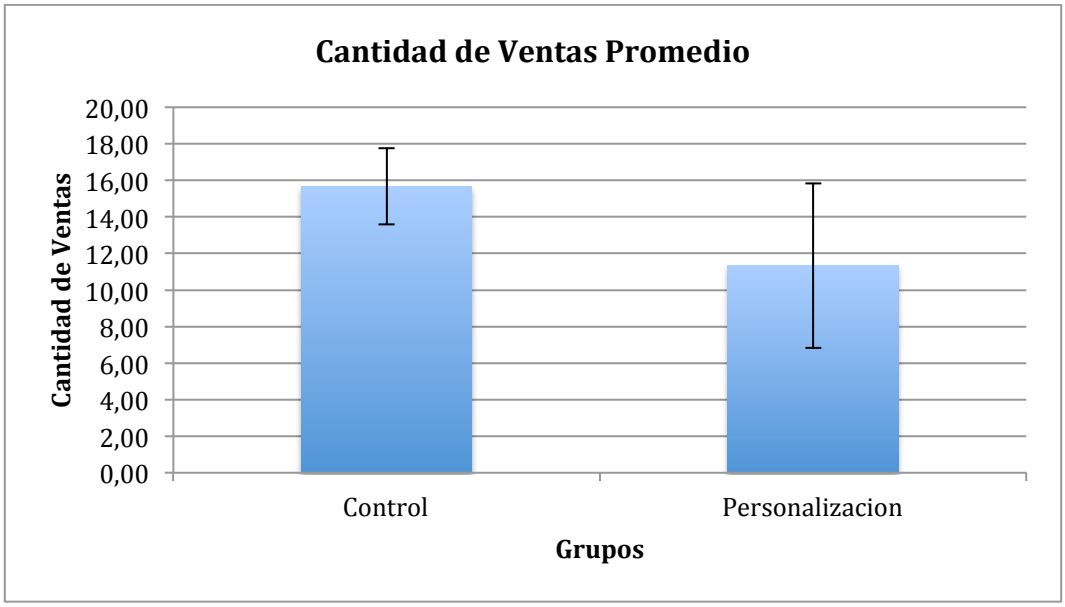


Ilustración 4.25: Gráfico de Cantidad de Ventas Promedio (Factor Personalización)

Como se observa en el gráfico anterior, la personalización no contribuye a mejorar las cantidades de ventas de la empresa, a pesar de que la Tasa de Clics sí aumente. Esto se puede deber al factor "suerte" que existe en los experimentos, ya que las cantidades de ventas son muy pocas respecto del total de clientes.

Finalmente, se desea concluir si la Tasa de Conversión de los clics provenientes del correo aumenta o no al personalizar el *Asunto* de estos. Para tal propósito, se guardó la cantidad de los clics efectuados y la cantidad de las ventas, con lo que se pudo calcular esta tasa, la cual se muestra a continuación.

i. Primer Envío

Grupo	Tasa de Conversión (%)
Control	2,57%
Personalización	1,84% (- 0,73%)

Tabla 4.30: Tasa de Conversión en Primer Envío (Factor Personalización)

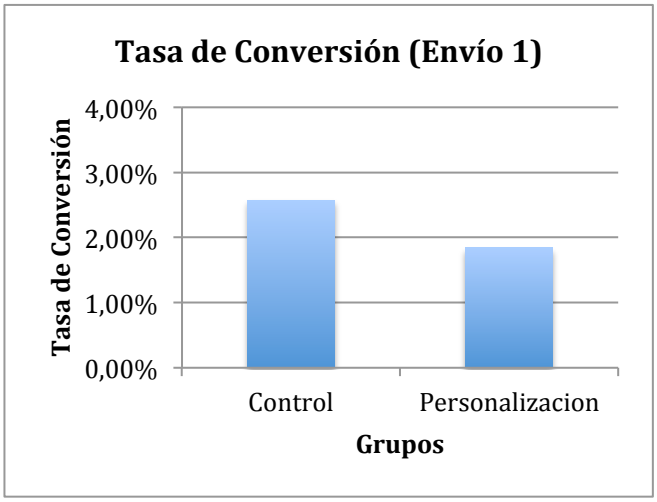


Ilustración 4.26: Tasa de Conversión en Primer Envío (Factor Personalización)

ii. Segundo Envío

Grupo	Tasa de Conversión (%)
Control	2,98%
Personalización	2,33% (- 0,65%)

Tabla 4.31: Tasa de Conversión en Segundo Envío (Factor Personalización)

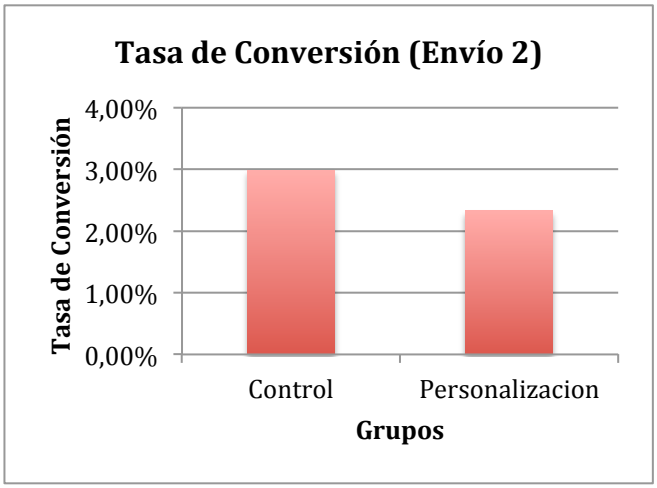


Ilustración 4.27: Tasa de Conversión en Segundo Envío (Factor Personalización)

iii. Tercer Envío

Grupo	Tasa de Conversión (%)
Control	3,18%
Personalización	1,18% (- 2,0%)

Tabla 4.32: Tasa de Conversión en Tercer Envío (Factor Personalización)

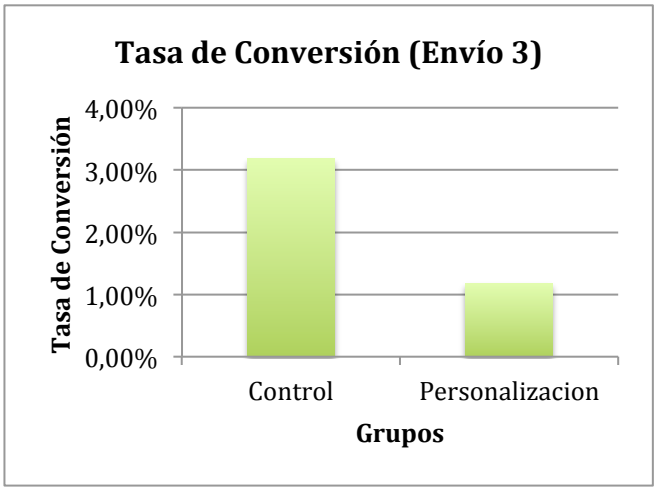


Ilustración 4.28: Tasa de Conversión en Tercer Envío (Factor Personalización)

En los tres envíos, se puede observar que la Tasa de Conversión es mayor en el grupo de control. Esto se debe a que, como se mostró anteriormente, las cantidades de ventas fueron superiores en este último, y la Tasa de Clicks fue mayor en el grupo personalizado, logrando así que la Tasa de Conversión sea menor en todos los envíos en el grupo experimental. A continuación, se muestra el promedio de todos los envíos respecto de esta métrica.

Grupo	Tasa de Conversión Promedio (%)	Desviación Estándar (%)
Control	2,91%	0,31%
Personalización	1,78% (- 1,13%)	0,57%

Tabla 4.33: Tasa de Conversión Promedio (Factor Personalización)

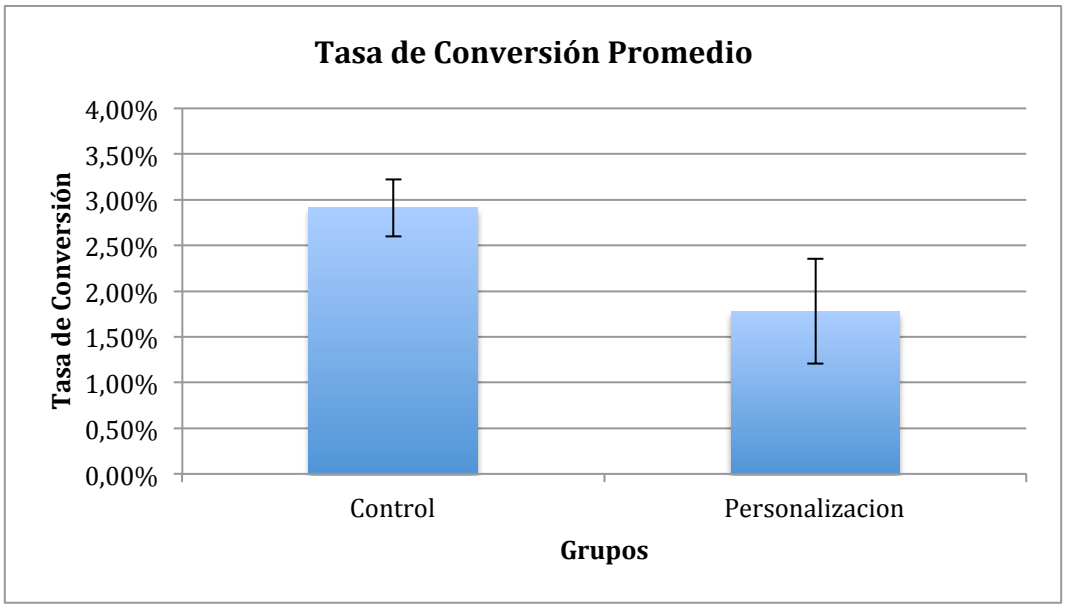


Ilustración 4.29: Gráfico de Tasa de Conversión Promedio (Factor Personalización)

Al igual que en las cantidades de ventas, el grupo de control supera con creces al grupo personalizado. Como se explicó anteriormente, este resultado es esperable, debido a que la Tasa de Clics aumenta y las cantidades de ventas son muy variables y atribuidas a la suerte. Lo anterior provoca que la Tasa de Conversión disminuya, por lo que no se puede concluir que la personalización afecta de manera positiva y significativa para aumentar esta métrica.

Por lo tanto, se puede establecer que el trato personalizado es un aporte significativo para aumentar la Tasa de Apertura y la Tasa de Clics de los clientes que reciben el *Newsletter*. Sin embargo, con respecto a las cantidades de ventas y a la Tasa de Conversión, no se puede concluir que esto ayudaría a incrementarlas.

4.5.3. Resultados de Recomendación

En esta sección, se analizarán los resultados obtenidos a partir del factor de Recomendación, donde se estudiará si hay una diferencia en el comportamiento de los clientes al enviarle ofertas de acuerdo con su historial de transacciones.

Recordando la Ilustración 4.9, se tienen 10 grupos distintos para los cuales se les hicieron tres envíos de experimentos. Para medir el factor de recomendación, se juntaron las columnas de la figura, con el fin de medir el efecto completo de la recomendación de las ofertas del *Newsletter*. En la Ilustración 4.34, se pueden observar los nuevos grupos.

Ilustración 4.30: Grupos de Factor Recomendación

Se desea determinar si estos modelos afectan en el comportamiento de los consumidores de acuerdo con las mismas cuatro métricas evaluadas en la personalización. La primera es la Tasa de Apertura, cuyos resultados se muestran a continuación.

i. Primer Envío

Grupo	Tasa de Apertura (%)
Control	21,48%
Reglas de Asociación con Productos Asociados	22,74% (+ 1,26%)*
Reglas de Asociación con Productos Frecuentes	23,79% (+ 2,31%)***
Filtros Colaborativos con Productos Asociados	23,47% (+ 1,99%)**
Filtros Colaborativos con Productos Frecuentes	22,35% (+ 0,87%)

Tabla 4.34: Tasa de Apertura en Primer Envío (Factor Recomendación)

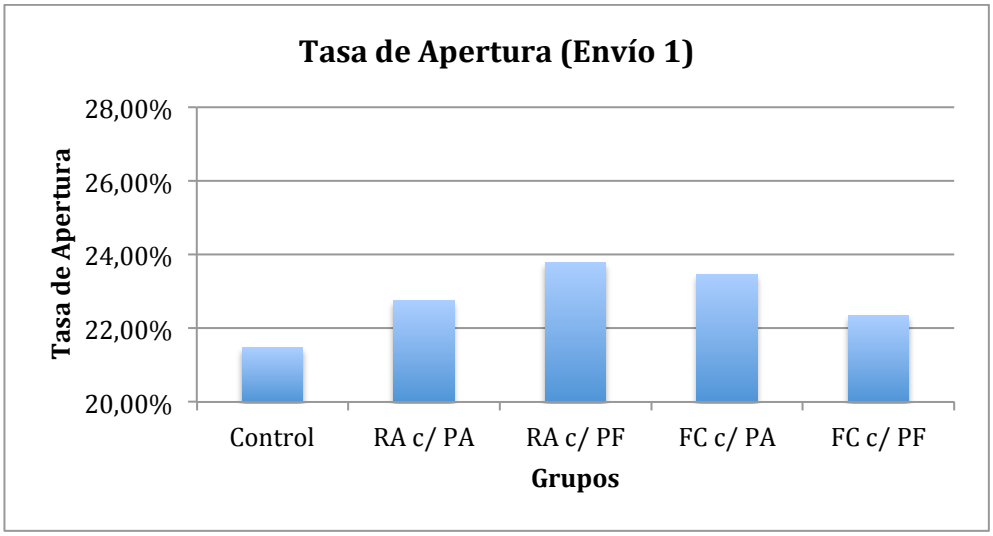


Ilustración 4.31: Tasa de Apertura en Primer Envío (Factor Recomendación)

ii. Segundo Envío

Grupo	Tasa de Apertura (%)
Control	25,31%
Reglas de Asociación con Productos Asociados	27,06% (+ 1,75%)*
Reglas de Asociación con Productos Frecuentes	25,84% (+ 0,53%)
Filtros Colaborativos con Productos Asociados	26,38% (+ 1,07%)
Filtros Colaborativos con Productos Frecuentes	25,72% (+ 0,41%)

Tabla 4.35: Tasa de Apertura en Segundo Envío (Factor Recomendación)

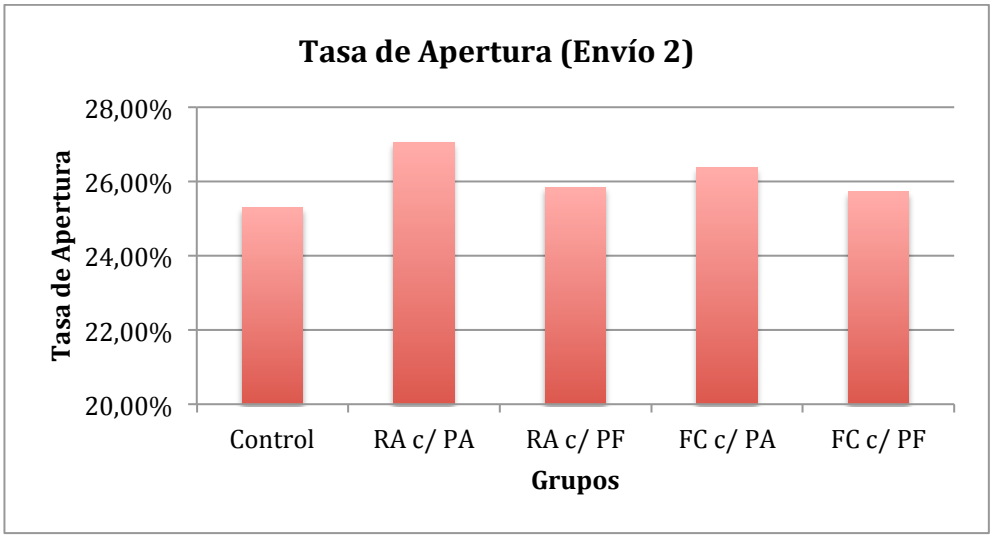


Ilustración 4.32: Tasa de Apertura en Segundo Envío (Factor Recomendación)

iii. Tercer Envío

Grupo	Tasa de Apertura (%)
Control	25,51%
Reglas de Asociación con Productos Asociados	25,09% (- 0,42%)
Reglas de Asociación con Productos Frecuentes	26,52% (+ 1,01%)
Filtros Colaborativos con Productos Asociados	25,91% (+ 0,4%)
Filtros Colaborativos con Productos Frecuentes	25,60% (+ 0,09%)

Tabla 4.36: Tasa de Apertura en Tercer Envío (Factor Recomendación)

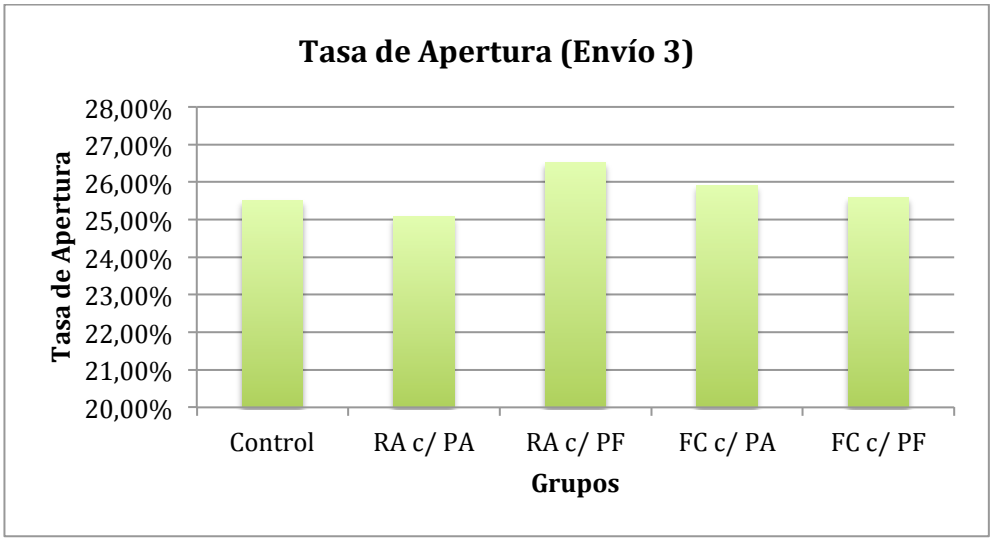


Ilustración 4.33: Tasa de Apertura en Tercer Envío (Factor Recomendación)

Como se puede observar en los distintos envíos del experimento, las Tasas de Apertura son muy variables a lo largo del tiempo. La tasa promedio se puede ver en la tabla y el gráfico que se presentan a continuación.

Grupo	Tasa de Apertura Promedio (%)	Desviación Estándar (%)
Control	24,10%	2,27%
Reglas de Asociación con Productos Asociados	24,96% (+ 0,86%)	2,16%
Reglas de Asociación con Productos Frecuentes	25,38% (+ 1,28%)*	1,42%
Filtros Colaborativos con Productos Asociados	25,25% (+ 1,15%)	1,56%
Filtros Colaborativos con Productos Frecuentes	24,56% (+ 0,46%)	1,91%

Tabla 4.37: Tasa de Apertura Promedio (Factor Recomendación)

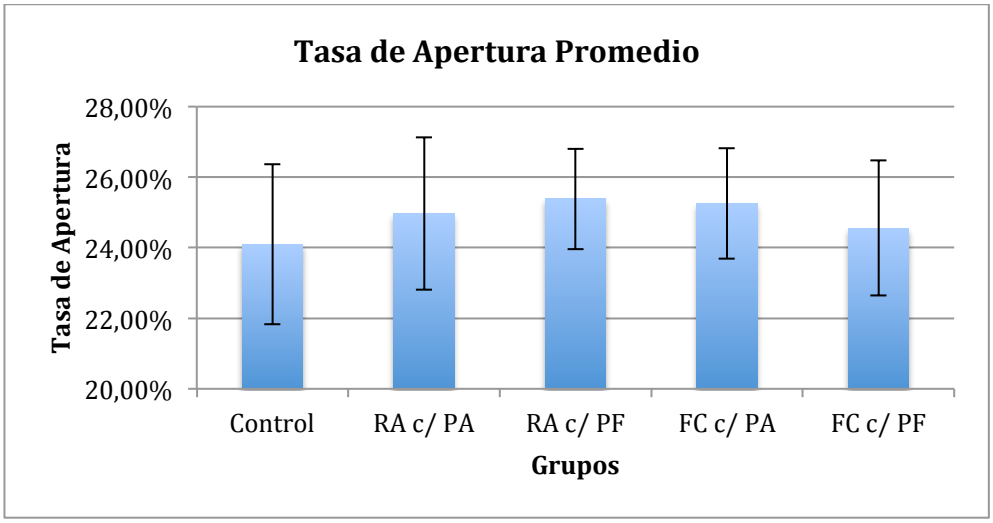


Ilustración 4.34: Gráfico de Tasas de Apertura Promedio (Factor Recomendación)

Para medir si estos resultados son significantes, se hizo un Test de Proporciones de Población y resultó que, con un 95% de confianza, no es significativo. Sin embargo, este resultado es esperable debido a que las recomendaciones afectan sólo en el contenido del correo y no en el *Asunto* de este. Por lo que los clientes no pueden ver nada de este antes de abrir el *Newsletter*.

Por otro lado, se desea estudiar si hay algún cambio en el comportamiento de los clientes debido a las recomendaciones personalizadas de los correos, por lo que se mide el desempeño mediante la Tasa de Clic, cuyos resultados se muestran a continuación.

i. Primer Envío

Grupo	Tasa de Clics (%)
Control	17,39%
Reglas de Asociación con Productos Asociados	19,54% (+ 2,15%)***
Reglas de Asociación con Productos Frecuentes	18,05% (+ 0,66%)
Filtros Colaborativos con Productos Asociados	18,04% (+ 0,65%)
Filtros Colaborativos con Productos Frecuentes	18,49% (+ 1,1%)**

Tabla 4.38: Tasa de Clics en Primer Envío (Factor Recomendación)

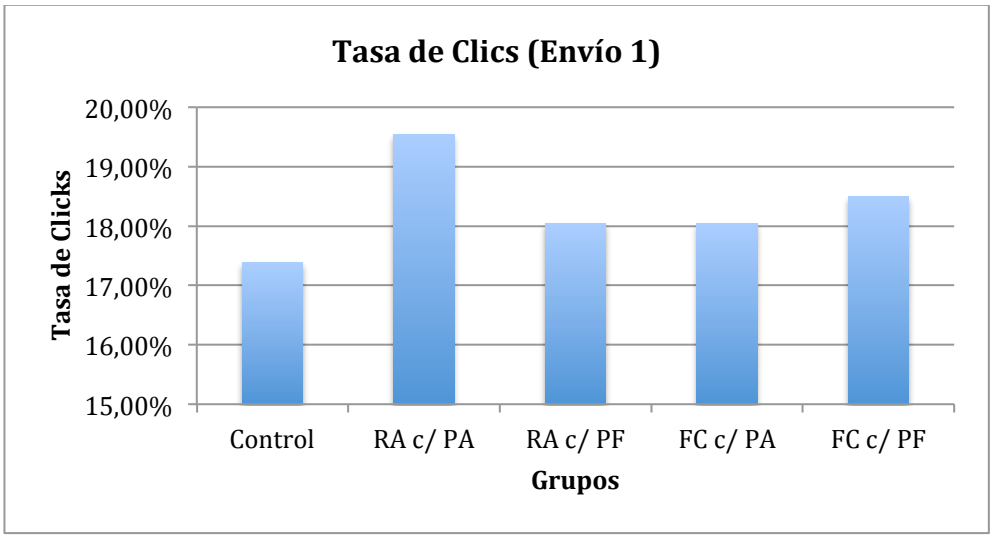


Ilustración 4.35: Tasa de Clics en Primer Envío (Factor Recomendación)

ii. Segundo Envío

Grupo	Tasa de Clics (%)
Control	16,75%
Reglas de Asociación con Productos Asociados	16,30% (- 0,45%)
Reglas de Asociación con Productos Frecuentes	18,73% (+ 1,98%)***
Filtros Colaborativos con Productos Asociados	17,48% (+ 0,73%)
Filtros Colaborativos con Productos Frecuentes	17,45% (+ 0,7%)

Tabla 4.39: Tasa de Clics en Segundo Envío (Factor Recomendación)

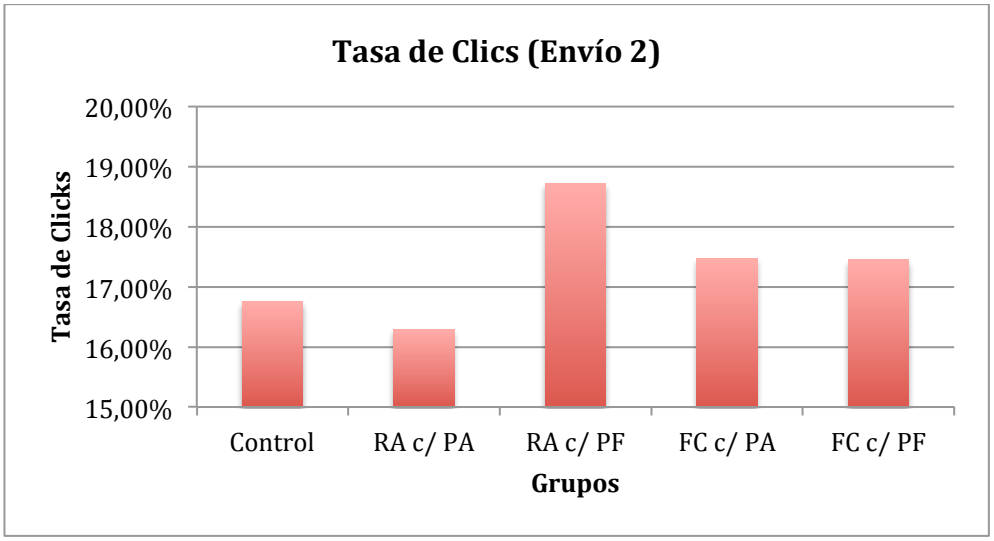


Ilustración 4.36: Tasa de Clics en Segundo Envío (Factor Recomendación)

iii. Tercer Envío

Grupo	Tasa de Clics (%)
Control	16,97%
Reglas de Asociación con Productos Asociados	17,11% (+ 0,14%)
Reglas de Asociación con Productos Frecuentes	17,46% (+ 0,49%)
Filtros Colaborativos con Productos Asociados	16,52% (- 0,45%)
Filtros Colaborativos con Productos Frecuentes	17,77% (+ 0,8%)*

Tabla 4.40: Tasa de Clics en Tercer Envío (Factor Recomendación)

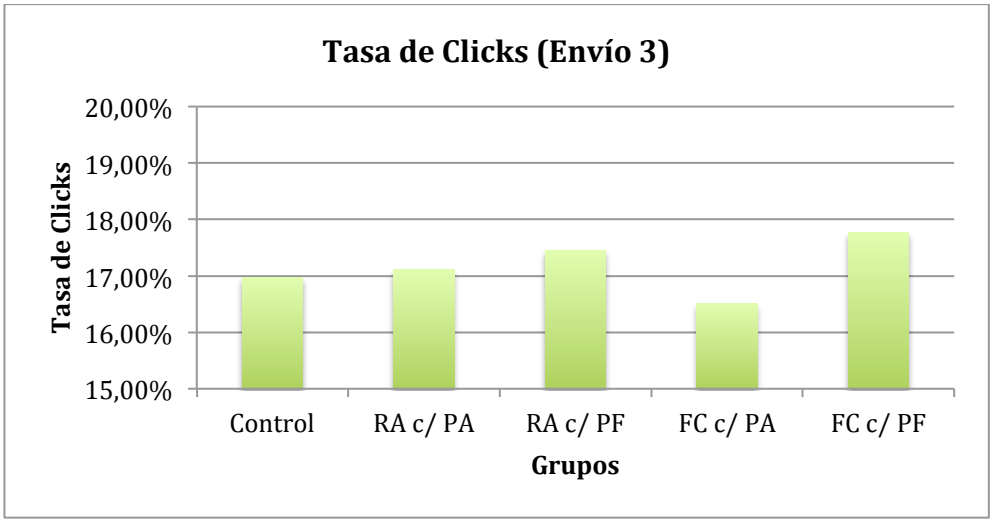


Ilustración 4.37: Tasa de Clics en Tercer Envío (Factor Recomendación)

Observando los resultados por cada uno de los envíos, se puede advertir que el modelo más constante en el tiempo, en términos de porcentaje por envío y por superioridad el grupo de control, es el Modelo de Reglas de Asociación con Productos Frecuentes. Para ver esto de mejor manera, se elaboró una tabla y un gráfico que contiene estos porcentajes en promedio.

Grupo	Tasa de Clics Promedio (%)	Desviación Estándar (%)
Control	17,04%	0,32%
Reglas de Asociación con Productos Asociados	17,65% (+ 0,61%)	1,69%
Reglas de Asociación con Productos Frecuentes	18,08% (+ 1,04%)**	0,64%
Filtros Colaborativos con Productos Asociados	17,35% (+ 0,31%)	0,77%
Filtros Colaborativos con Productos Frecuentes	17,90% (+ 0,86%)	0.54%

Tabla 4.41: Tasa de Clics Promedio (Factor Recomendación)

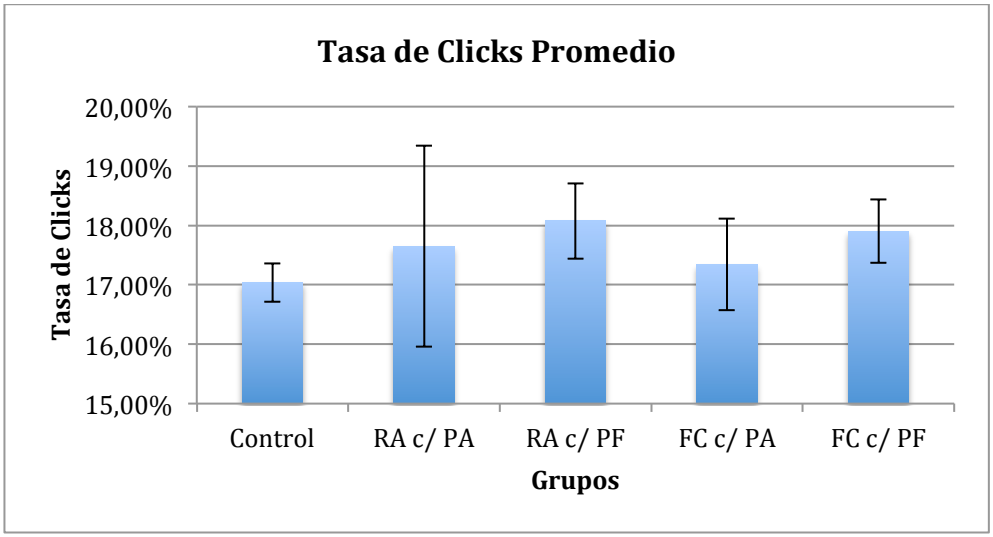


Ilustración 4.38: Gráfico de Tasa de Clics Promedio (Factor Recomendación)

Se puede observar que, en promedio, el modelo de Reglas de Asociación con Productos Frecuentes es el que tiene una mayor Tasa de Clics. No obstante lo anterior, para concluir si este resultado es significante o no, se utilizó un Test de Proporciones de Poblaciones, el cual arrojó que, con un 95% de confianza, este modelo supera a las ofertas populares.

Al superar por más de un 1% al grupo de control, se puede afirmar que, aplicando este modelo a toda la base de datos de la empresa, el sitio *web* tendría un aumento de aproximadamente 1.000 clientes por día, los cuales podrían hacer que la venta incremente.

Para medir también el efecto de las recomendaciones, se estudia la diferencia entre las cantidades de ventas que los clientes tuvieron de los distintos modelos. Estos resultados se muestran a continuación.

i. Primer Envío

Grupo	Cantidad de Venta
Control	6
Reglas de Asociación con Productos Asociados	3 (- 3)
Reglas de Asociación con Productos Frecuentes	3 (- 3)
Filtros Colaborativos con Productos Asociados	8 (+ 2)
Filtros Colaborativos con Productos Frecuentes	5 (- 1)

Tabla 4.42: Cantidad de Venta en Primer Envío (Factor Recomendación)

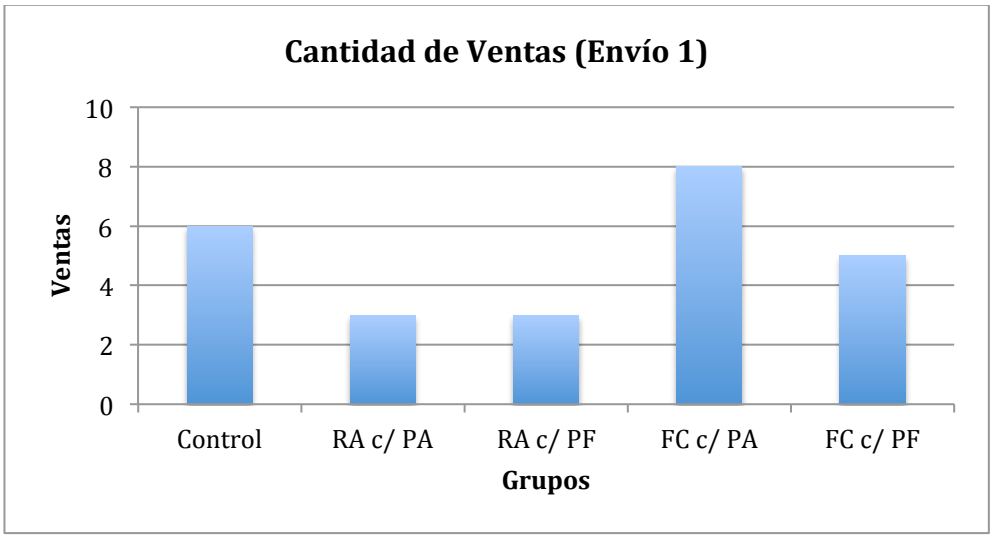


Ilustración 4.39: Cantidad de Venta en Primer Envío (Factor Recomendación)

ii. Segundo Envío

Grupo	Cantidad de Venta
Control	3
Reglas de Asociación con Productos Asociados	3 (+ 0)
Reglas de Asociación con Productos Frecuentes	12 (+ 9)
Filtros Colaborativos con Productos Asociados	8 (+ 5)
Filtros Colaborativos con Productos Frecuentes	5 (+ 2)

Tabla 4.43: Cantidad de Ventas en Segundo Envío (Factor Recomendación)

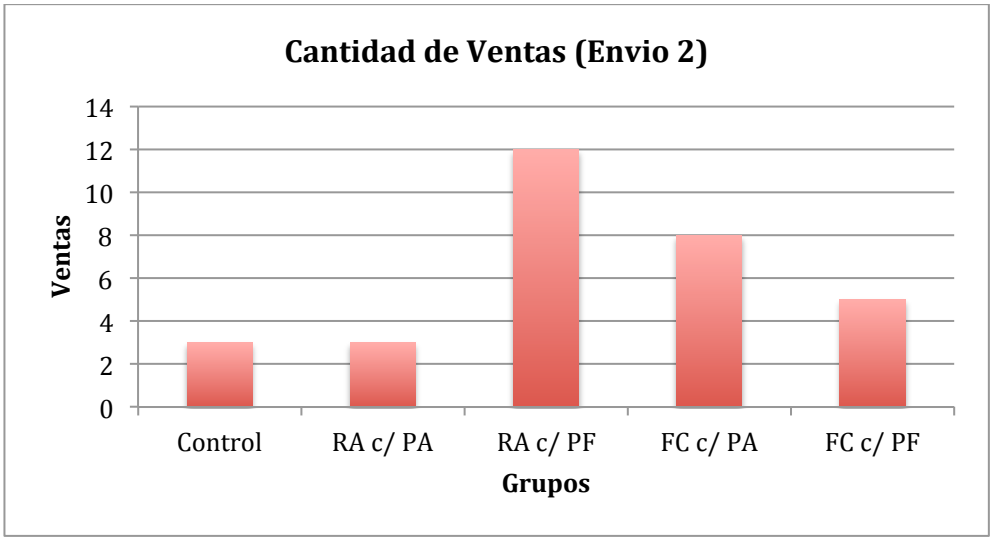


Ilustración 4.40: Cantidad de Ventas en Segundo Envío (Factor Recomendación)

iii. Tercer Envío

Grupo	Cantidad de Venta
Control	4
Reglas de Asociación con Productos Asociados	5 (+ 1)
Reglas de Asociación con Productos Frecuentes	7 (+ 3)
Filtros Colaborativos con Productos Asociados	4 (+ 0)
Filtros Colaborativos con Productos Frecuentes	5 (+ 1)

Tabla 4.44: Cantidad de Ventas en Tercer Envío (Factor Recomendación)

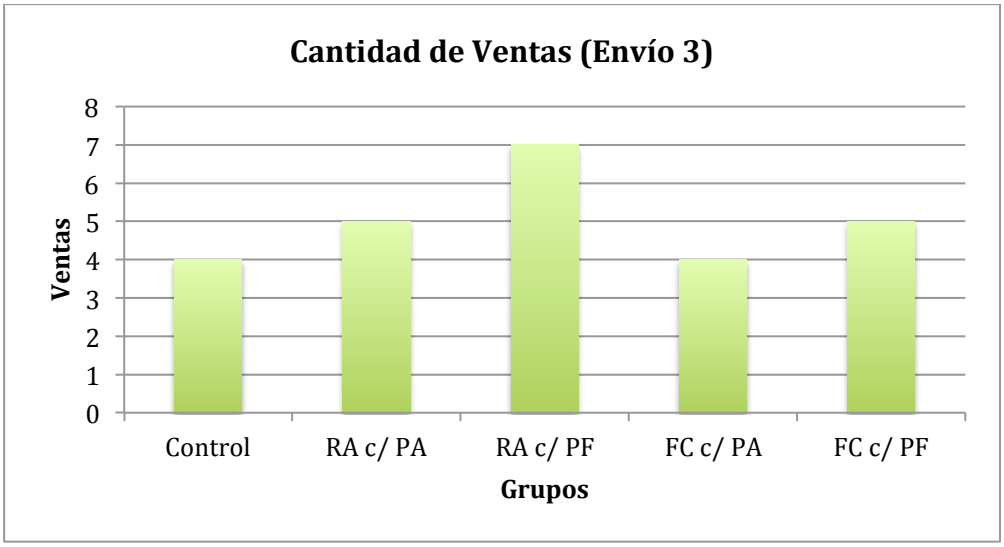


Ilustración 4.41: Cantidad de Ventas en Tercer Envío (Factor Recomendación)

En estos tres envíos, se puede apreciar que la variabilidad de las ventas de cada grupo es extremadamente voluble. Sin embargo, en los últimos dos envíos, se puede observar una clara superioridad de uno de los modelos, el cual coincide con el que tuvo una mayor Tasa de Clics, correspondiente a las Reglas de Asociación con Productos Frecuentes. Para ver cuales fueron los resultados en promedio se elaboró la siguiente tabla.

Grupo	Cantidad de Venta Promedio	Desviación Estándar
Control	4,33	1,53
Reglas de Asociación con Productos Asociados	3,66 (- 0,67)	1,15
Reglas de Asociación con Productos Frecuentes	7,33 (+ 3,0)	4,50
Filtros Colaborativos con Productos Asociados	6,66 (+ 2,33)	2,31
Filtros Colaborativos con Productos Frecuentes	5,00 (+ 0,67)	0,00

Tabla 4.45: Cantidad de Venta Promedio (Factor Recomendación)

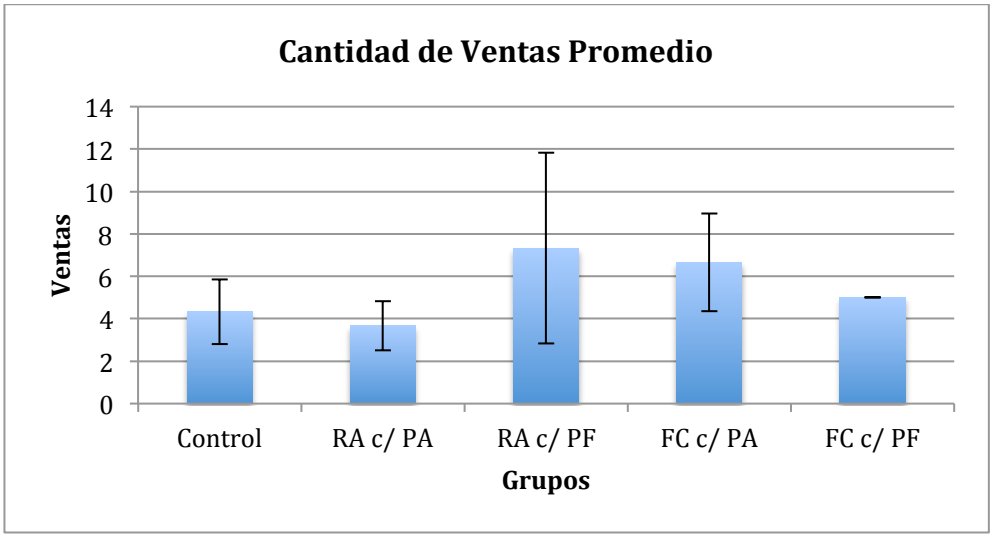


Ilustración 4.42: Gráfico de Cantidad de Venta Promedio (Factor Recomendación)

Como se puede observar en la tabla y en el gráfico anterior, las ventas del modelo descrito anteriormente son muy superiores en promedio, llegando a ser hasta casi un 70% mayor. Sin embargo, se midió la significancia de estos resultados, los que no alcanzaron un 95% de confianza. Esto se debe a que las ventas son en cantidades muy pequeñas y muy variables. Pero, observado la data descriptiva, y a pesar de que no sea significativo, la diferencia es muy grande, por lo que sí se puede concluir que este modelo de recomendación ayudaría a incrementar la venta en la empresa.

Finalmente, se quiere concluir si la recomendación personalizada para cada uno de los clientes afecta en la Tasa de Conversión de compras provenientes del correo. A continuación, se muestran los resultados de los experimentos respecto de esta métrica.

i. Primer Envío

Grupo	Tasa de Conversión (%)
Control	3,19%
Reglas de Asociación con Productos Asociados	1,26% (- 1,93%)
Reglas de Asociación con Productos Frecuentes	1,19% (+ 2,0%)
Filtros Colaborativos con Productos Asociados	3,48% (+ 0,29%)
Filtros Colaborativos con Productos Frecuentes	2,13% (- 1,06%)

Tabla 4.46: Tasa de Conversión en Primer Envío (Factor Recomendación)

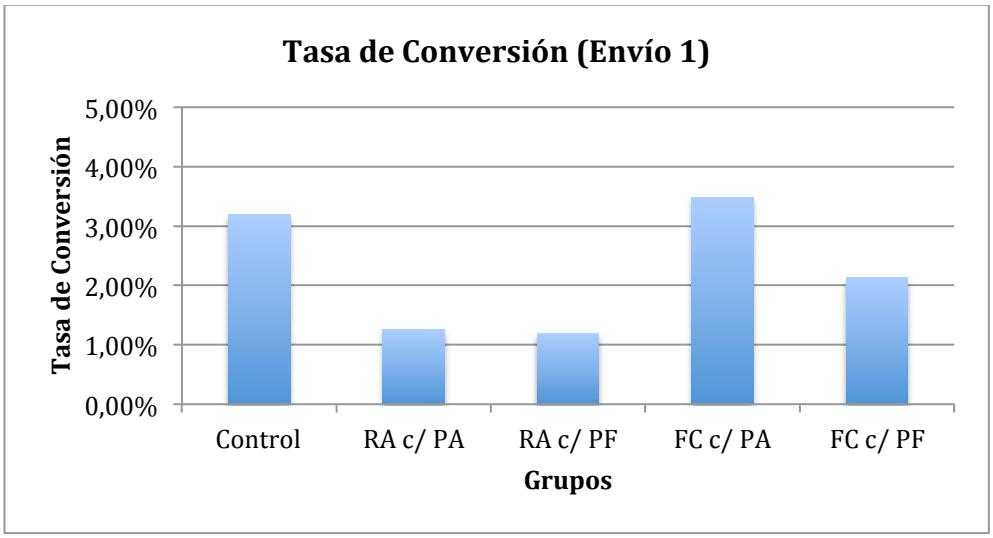


Ilustración 4.43: Tasa de Conversión en Primer Envío (Factor Recomendación)

ii. Segundo Envío

Grupo	Tasa de Conversión (%)
Control	1,25%
Reglas de Asociación con Productos Asociados	1,33% (+ 0,08%)
Reglas de Asociación con Productos Frecuentes	4,71% (+ 3,46%)***
Filtros Colaborativos con Productos Asociados	3,43% (+ 2,18%)***
Filtros Colaborativos con Productos Frecuentes	2,11% (+ 0,86%)***

Tabla 4.47: Tasa de Conversión en Segundo Envío (Factor Recomendación)

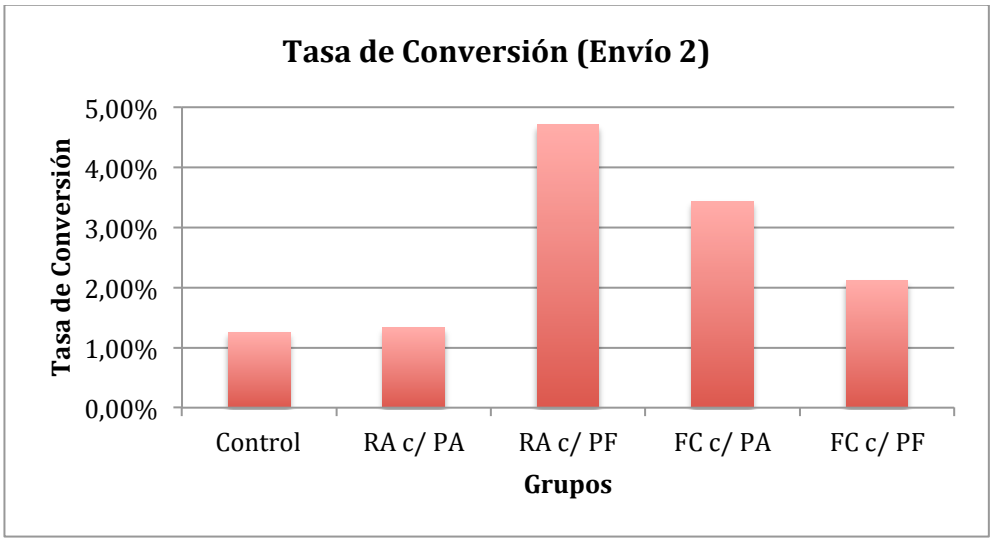


Ilustración 4.44: Tasa de Conversión en Segundo Envío (Factor Recomendación)

iii. Tercer Envío

Grupo	Tasa de Conversión (%)
Control	1,86%
Reglas de Asociación con Productos Asociados	2,25% (+ 0,39%)**
Reglas de Asociación con Productos Frecuentes	2,93% (+ 1,07%)***
Filtros Colaborativos con Productos Asociados	1,77% (- 1,09%)
Filtros Colaborativos con Productos Frecuentes	1,95% (+ 0,09%)

Tabla 4.48: Tasa de Conversión en Tercer Envío (Factor Recomendación)

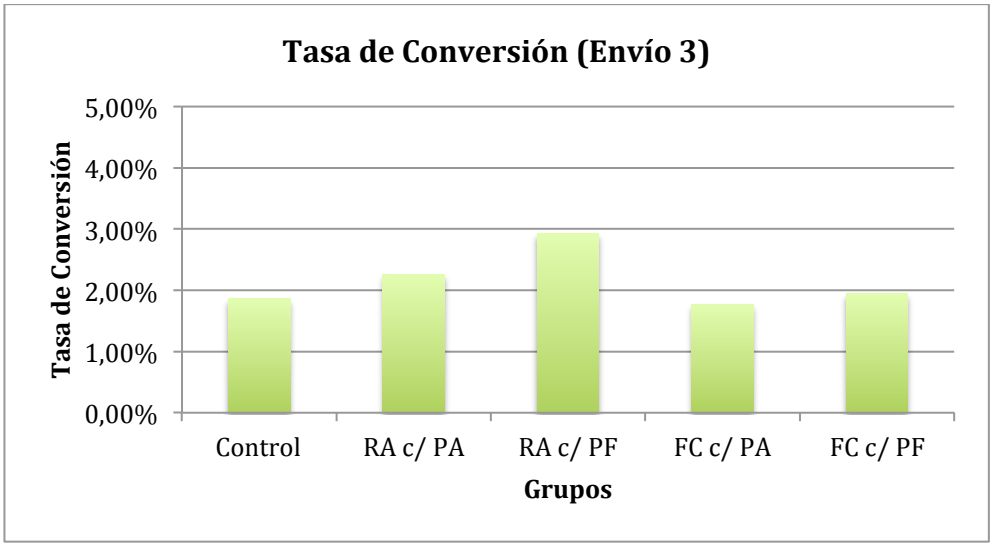


Ilustración 4.45: Tasa de Conversión en Tercer Envío (Factor Recomendación)

Estos resultados están ligados a las cantidades de ventas y a la tasa de clics, ya que es calculado según la cantidad de ventas por sobre la cantidad de clics. Al igual que como se realizó en las métricas anteriores, se elaboró una tabla y un gráfico con las tasas en promedio, los cuales se muestran a continuación.

Grupo	Tasa de Conversión Promedio (%)	Desviación Estándar (%)
Control	2,10%	0,99%
Reglas de Asociación con Productos Asociados	1,61% (- 0,49%)	0,55%
Reglas de Asociación con Productos Frecuentes	2,94% (+ 0,84%)***	1,76%
Filtros Colaborativos con Productos Asociados	2,89% (+ 0,79%)***	0,97%
Filtros Colaborativos con Productos Frecuentes	2,06% (- 0,04%)	0,10%

Tabla 4.49: Tasa de Conversión Promedio (Factor Recomendación)

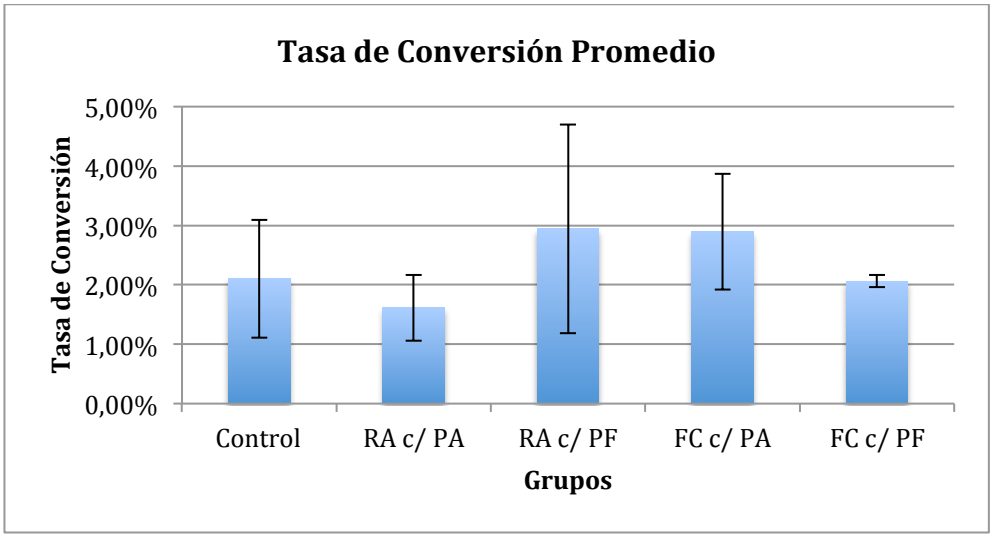


Ilustración 4.46: Gráfico de Tasa de Conversión Promedio (Factor Recomendación)

Nuevamente, se puede observar que el mejor modelo, y el que mayor Tasa de Conversión tiene en promedio, es el de Reglas de Asociación con Productos Frecuentes. Para ver si este resultado es significativo o no, se hizo un Test de Proporciones de Poblaciones, el cual arrojó que, con un 95% de confianza, este resultado es significativo.

Analizando este 0,84% de superioridad entre la Tasa de Conversión del modelo con el grupo de control, se puede concluir que la diferencia de los que entraron al sitio *web* y además compraron un producto se superó en un 40%, porcentaje importante para el aumento de las ventas.

4.5.4. Resultados Individuales por Grupo Experimental

Ya se pudo observar que, en términos generales, el Factor de personalización afecta de manera positiva y significativa para aumentar la Tasa de Apertura y la Tasa de Clics del

Newsletter, y además, que las recomendaciones mediante las Reglas de Asociación con Productos Frecuentes son las que más aumentan la Tasa de Clics y la Tasa de conversión.

Al conocer estos resultados generales, es necesario poder contrastarlos con los resultados individuales de cada uno de los grupos experimentales, para poder observar cómo se comportan los clientes con cada una de las combinaciones de factores, y para posteriormente concluir cómo es conveniente utilizar la personalización y también las recomendaciones.

En primer lugar, se muestran los resultados de la Tasa de Apertura para cada uno de los grupos experimentales y de control.

Grupo	Envío 1	Envío 2	Envío 3	Prom.	Desv. Est.
Asunto Control – Ofertas Control	20,74%	24,03%	24,29%	23,02%	1,98%
Asunto Control - Reglas de Asociación con Productos Asociados	22,34%	25,89%	24,97%	24,40%	1,84%
Asunto Control - Reglas de Asociación con Productos Frecuentes	22,98%	25,63%	25,97%	24,86%	1,63%
Asunto Control - Filtros Colaborativos con Productos Asociados	23,23%	25,48%	25,33%	24,68%	1,26%
Asunto Control - Filtros Colaborativos con Productos Frecuentes	21,64%	25,65%	24,87%	24,05%	2,13%
Asunto Peronalizado – Ofertas Control	22,21%	26,59%	26,72%	25,17%	2,57%
Asunto Peronalizado - Reglas de Asociación con Productos Asociados	23,14%	28,24%	25,20%	25,53%	2,56%
Asunto Peronalizado - Reglas de Asociación con Productos Frecuentes	24,57%	26,05%	27,06%	25,89%	1,25%
Asunto Peronalizado - Filtros Colaborativos con Productos Asociados	23,70%	27,27%	26,48%	25,82%	1,87%
Asunto Peronalizado - Filtros Colaborativos con Productos Frecuentes	23,06%	25,80%	26,34%	25,07%	1,76%

Tabla 4.50: Tasa de Apertura por Grupo Experimental

Como se explicó anteriormente, la Tasa de Apertura solo se puede ver afectada con la personalización del *Asunto* del correo, por lo que, como se puede observar en el cuadro

anterior, existe una clara diferencia entre los grupos con asunto personalizado y los con el asunto original.

A continuación, se muestran los resultados para la métrica de Tasa de Clics de cada uno de los grupos.

Grupo	Envío 1	Envío 2	Envío 3	Prom.	Desv. Est.
Asunto Control – Ofertas Control	17,21%	15,52%	15,95%	16,22%	0,88%
Asunto Control - Reglas de Asociación con Productos Asociados	17,95%	15,73%	19,11%	17,59%	1,72%
Asunto Control - Reglas de Asociación con Productos Frecuentes	19,82%	16,90%	17,72%	18,15%	1,50%
Asunto Control - Filtros Colaborativos con Productos Asociados	18,22%	12,40%	15,91%	15,51%	2,93%
Asunto Control - Filtros Colaborativos con Productos Frecuentes	17,10%	14,71%	16,18%	16,00%	1,20%
Asunto Peronalizado – Ofertas Control	17,55%	17,86%	17,90%	17,77%	0,19%
Asunto Peronalizado - Reglas de Asociación con Productos Asociados	21,08%	16,82%	15,13%	17,67%	3,07%
Asunto Peronalizado - Reglas de Asociación con Productos Frecuentes	16,43%	20,47%	17,21%	18,03%	2,14%
Asunto Peronalizado - Filtros Colaborativos con Productos Asociados	17,87%	22,12%	17,09%	19,03%	2,71%
Asunto Peronalizado - Filtros Colaborativos con Productos Frecuentes	19,82%	20,20%	19,29%	19,77%	0,46%

Tabla 4.51: Tasa de Clics por Grupo Experimental

Acá se puede observar si existen diferencias entre los distintos elementos de los factores. En el factor de recomendación, se concluyó que el modelo de Reglas de Asociación con Productos Frecuentes es el más consistente y con una mayor diferencia significativa con respecto al grupo de control, sin embargo, al observar esta data descriptiva, se puede concluir que la mejor combinación de factores corresponde a la Personalización del *Asunto* en conjunto con el modelo de Filtros Colaborativos con Productos Frecuentes, ya que es el que tiene una mayor Tasa de Clics en promedio, y con una desviación estándar muy pequeña, lo que lo hace ser muy consistente.

Una vez analizada la tasa de clics de los clientes que abren los correos en cada grupo, es necesario saber cuantos de ellos efectivamente compraron. Es por esto que, a continuación, se muestra cuantas ventas se realizaron en los distintos grupos.

Grupo	Envío 1	Envío 2	Envío 3	Prom.	Desv. Est.
Asunto Control – Ofertas Control	4	3	2	3,00	1,00
Asunto Control - Reglas de Asociación con Productos Asociados	1	2	4	2,33	1,53
Asunto Control - Reglas de Asociación con Productos Frecuentes	3	7	5	5,00	2,00
Asunto Control - Filtros Colaborativos con Productos Asociados	5	2	2	3,00	1,73
Asunto Control - Filtros Colaborativos con Productos Frecuentes	1	1	5	2,33	2,31
Asunto Peronalizado – Ofertas Control	2	0	2	1,33	1,15
Asunto Peronalizado - Reglas de Asociación con Productos Asociados	2	1	1	1,33	0,58
Asunto Peronalizado - Reglas de Asociación con Productos Frecuentes	0	5	2	2,33	2,52
Asunto Peronalizado - Filtros Colaborativos con Productos Asociados	3	6	2	3,67	2,08
Asunto Peronalizado - Filtros Colaborativos con Productos Frecuentes	4	4	0	2,67	2,31

Tabla 4.52: Cantidad de Ventas por Grupo Experimental

Al calcular la significancia de los cambios de las ventas en los distintos factores, se concluyó que estos no eran significativos, ya que estas cantidades son muy pequeñas y se pueden deber a la suerte. Sin embargo, en la tabla se puede apreciar el desempeño de cada uno de los grupos, donde el modelo de Reglas de Asociación con Productos Frecuentes en conjunto con el asunto original del correo, son la mejor combinación para aumentar las ventas. Esto solo se puede concluir observando data descriptiva y no mediante un test de significancia.

Luego de analizar las cantidades de ventas de cada uno de los grupos, se puede establecer la Tasa de Conversión de cada uno de ellos, considerando la cantidad de clics con las ventas. Estos resultados se exponen a continuación.

Grupo	Envío 1	Envío 2	Envío 3	Prom.	Desv. Est.
Asunto Control – Ofertas Control	4,55%	2,63%	2,11%	3,09%	1,28%
Asunto Control - Reglas de Asociación con Productos Asociados	0,93%	2,08%	3,15%	2,06%	1,11%
Asunto Control - Reglas de Asociación con Productos Frecuentes	2,21%	5,79%	4,27%	4,09%	1,80%
Asunto Control - Filtros Colaborativos con Productos Asociados	4,50%	2,53%	2,04%	3,03%	1,30%
Asunto Control - Filtros Colaborativos con Productos Frecuentes	0,97%	1,08%	3,88%	1,97%	1,65%
Asunto Peronalizado – Ofertas Control	2,00%	0,00%	1,67%	1,22%	1,07%
Asunto Peronalizado - Reglas de Asociación con Productos Asociados	1,53%	0,77%	1,05%	1,12%	0,38%
Asunto Peronalizado - Reglas de Asociación con Productos Frecuentes	0,00%	3,73%	1,64%	1,79%	1,87%
Asunto Peronalizado - Filtros Colaborativos con Productos Asociados	2,52%	3,90%	1,56%	2,66%	1,17%
Asunto Peronalizado - Filtros Colaborativos con Productos Frecuentes	3,03%	2,78%	0,00%	1,94%	1,68%

Tabla 4.53: Tasa de Conversión por Grupo Experimental

Observando la tabla, se puede concluir que la mejor combinación de factores para aumentar la Tasa de Conversión es utilizar el asunto original del correo en conjunto con el modelo de Reglas de Asociación con Productos Frecuentes, donde es superada por un 1%, equivalente a aproximadamente un 33% mayor al grupo que le sigue.

5. CONCLUSIONES

Este trabajo fue formulado, en un principio, para poder mejorar el desempeño actual del *Newsletter* diario que envía la empresa a sus clientes promocionando las ofertas más populares o las que se cree que más venderán. Es por esto, que se desarrollaron dos distintos modelos de recomendaciones y, además, se creó un factor de personalización, cumpliendo así con el objetivo general del proyecto.

Para llevar a cabo estos modelos de recomendación, se elaboró una metodología que contenía todos los pasos para lograr obtener dos modelos que mejor se ajustaran a los datos relevantes de la empresa. De esta forma, se propuso y se generó un modelo en base a Filtros Colaborativos sustentado en el Usuario y otro basado en Reglas de Asociación. Como la empresa no guarda actualmente qué productos o servicios son recomendables para promocionarlos una vez que un cliente ya lo compró, se elaboraron dos variantes a estos modelos, uno repitiendo productos que los clientes ya compraron (Productos Frecuentes) y otro sin repetirlos (solo Productos Asociados).

Es así como se lograron extraer recomendaciones desde dos distintos modelos con dos distintas variables, por lo que se obtuvieron distintos productos asociados a distintos clientes desde dos algoritmos distintos, para después probar cuál de estos es el mejor.

Para poder verificar la calidad de las recomendaciones de los distintos modelos, se desarrolló una experimentación que consistía en probar en conjunto las recomendaciones con la personalización del correo. Es así como se formaron 10 distintos grupos que combinaban el contenido y el asunto original del correo (control) y las recomendaciones generadas por los distintos modelos (tratamiento).

Se hicieron tres envíos exitosos para probar la efectividad de cada una de las combinaciones en cada uno de los grupos, donde se pudieron sacar las siguientes conclusiones principales:

- En general, la personalización del *Asunto* del correo aumenta la Tasa de Apertura y la Tasa de Clicks con un 95% de confianza, logrando así aumentar el tráfico del correo y además el tráfico en el sitio *web* de la empresa.

- Con un 95% de confianza, la Tasa de Clics y la Tasa de Conversión se puede aumentar mediante la implementación de un modelo de recomendación, específicamente con el modelo de *Reglas de Asociación con Productos Frecuentes*.

- En promedio, las ventas se pueden incrementar al implementar un modelo de recomendación a pesar de que los resultados no sean significativos. Observando la data descriptiva de los resultados, se puede advertir un claro patrón en el cambio de las ventas entre los grupos de control y las recomendaciones, por lo que sí se puede sacar esta conclusión. Al ser cantidades muy pequeñas, los resultados de la significancia no son los esperados.

Finalmente, se concluye que la personalización del *Asunto* del correo aumenta en general la Tasa de Apertura del *Newsletter* diario, aumentando así el tráfico del correo electrónico. Aplicando las mismas cifras a todos los clientes que reciben el correo, se podría aumentar el tráfico de este en hasta 4.500 personas al día, pudiendo así aumentar las probabilidades de que otros clientes realicen una compra en el sitio *web*.

Por otro lado, el mejor modelo de recomendación corresponde al creado por *Reglas de Asociación con Productos Frecuentes*, por sobre todos los demás. Al observar la evidencia de los resultados obtenidos en los distintos experimentos, se pudo constatar que este modelo, aplicando las mismas cifras a toda la base de datos que recibe el correo, aumentaría la Tasa de Clics en cerca de un 1%, lo que equivale a alrededor de 1.000 clientes adicionales diariamente que ingresan al sitio *web* de la empresa. Además, la Tasa de Conversión se aumentó en un 50% (de un 2% a un 3%), lo que quiere decir que se podría aumentar en un 50% las personas que ingresan a la página y realizan una compra mediante el correo. Por último, se pudo observar que la diferencia de compras entre este modelo y el grupo de control es de un 70%, por lo que si se aplica esta cifra a toda la base de datos, la empresa podría aumentar considerablemente sus ventas.

6. RECOMENDACIONES Y TRABAJOS FUTUROS

Una vez obtenidos los resultados de la experimentación realizada con los modelos de recomendación, y en particular con las Reglas de Asociación con Productos Frecuentes, es imprescindible poder implementar a toda la base de datos este modelo para poder lograr mayor tráfico en el sitio web y poder aumentar las ventas. Sin embargo, este modelo está limitado a clientes con 3 compras o intentos de compra cada uno, ya que es la única forma de poder tener un historial de transacciones adecuado para generar recomendaciones. Para solucionar esto, se le recomienda a la empresa empezar a guardar las visitas de las ofertas que cada uno de los clientes realiza al día, para así poder tener otras referencias y otra información con la que se pueda trabajar para hacer recomendaciones. Con esto, se le podría generar recomendaciones a clientes con incluso ninguna compra.

Como se mostró en el informe, la cantidad de recomendaciones por persona no superan las 7 ofertas en promedio dentro de las 33 que se envían diariamente, es por esto que se recomienda utilizar este modelo para generar un correo semanal o mensual que contenga sólo recomendaciones para cada cliente, además del correo diario.

Por otro lado, al indicar una clara postura por asociaciones de categorías en las distintas recomendaciones, es posible utilizar las *Reglas de Asociación con Productor Frecuentes* para recomendar productos y servicios cuando se ingresan a ver distintas ofertas.

Es también recomendable ordenar el sitio *web* de la empresa sobre la base de las recomendaciones de cada uno de los clientes, es decir, cuando una persona ingrese con su usuario, automáticamente se le debe ordenar la página desde las ofertas más recomendadas a las menos.

7. BIBLIOGRAFÍA

[1] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, 2003.

[2] J. J. Levandoski, A. Eldawy, M. D. Ekstrand, M. F. Mokbel, M. J. Ludwig, and J. T. Riedl, "RecBench: Benchmarks for Evaluating Performance of Recommender System Architectures," *Pvldb*, pp. 911–920, 2013.

[3] L. Candillier, F. Meyer, and M. Boullé, "Comparing state-of-the-art collaborative filtering systems," *Lect. Notes Comput. Sci.*, vol. 4571, p. 548, 2007.

[4] M. D. Ekstrand, "Collaborative Filtering Recommender Systems," *Found. Trends® Human–Computer Interact.*, vol. 4, no. 2, pp. 81–173, 2010.

[5] R. Burke, "Integrating Knowledge-based and Collaborative-filtering Recommender Systems," pp. 69–72, 1999.

[6] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Proc. 14th Conf. Uncertain. Artif. Intell.*, vol. 461, no. 8, pp. 43–52, 1998.

[7] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in Action*. 2011.

[8] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Comput. Linguist.*, vol. 19, pp. 61–74, 1993.

[9] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Item-based collaborative filtering recommendation algorithms."

[10] F. Faltin, and R. Kenett, "Bayesian Networks," *Encycl. Stat. Qual. Reliab.*, vol. 1, no. 1, p. 4, 2007.

[11] D. De Ciencias and D. Computaci, *Universidad de Granada Modelos de Recomendaci ´ on Basados en Redes Bayesianas*. 2011.

[12] Y. Chen and E. I. George, "A bayesian model for collaborative filtering," *Direct*, no. 1, 1999.

[13] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *Organization*, vol. 1215, no. 1558601538, pp. 487–499, 1994.

[14] J. Han, J. Pei, and Y. Yin, "Mining F requen tP atterns without Candidate Generation," *Sigmod*, vol. 79, no. 3, pp. 1–14, 2000.

[15] R. Van Meteren and M. Van Someren, “Using Content-Based Filtering for Recommendation,” *ECML/MLNET Work. Mach. Learn. New Inf. Age*, pp. 47–56, 2000.

[16] N. Tintarev and J. Masthoff, *Recommender Systems Handbook*, vol. 54. 2011.

[17] R. Burke, “Hybrid web recommender systems,” *Adapt. web*, pp. 377–408, 2007.

[18] G. Shani and A. Gunawardana, “Evaluating recommendation systems,” *Recomm. Syst. Handb.*, pp. 257–298, 2011.

[19] J. D. M. Rennie, “Derivation of the F-Measure,” *In other words*, vol. 11, no. 1, p. 4, 2004.

[20] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” *Ann. Phys. (N. Y).*, vol. 54, p. 770, 2006.

[21] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and a. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[22] D. Pelleg and A. Moore, “X-means: Extending K-Means with Efficient Estimation of the Number of Clusters,” *Morgan Kaufmann Publ. Inc. San Fr. CA, USA*, no. ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, p. Pages 727–734, 2000.

[23] S. Ghosh and S. Dubey, “Comparative Analysis of K-Means and Fuzzy C-Means Algorithms,” *Ijacsa*, vol. 4, no. 4, pp. 35–39, 2013.

[24] N. Alldrin, A. Smith, and D. Turnbull, “Clustering with EM and K-means,” *Univ. San Diego, California, …*, 2003.

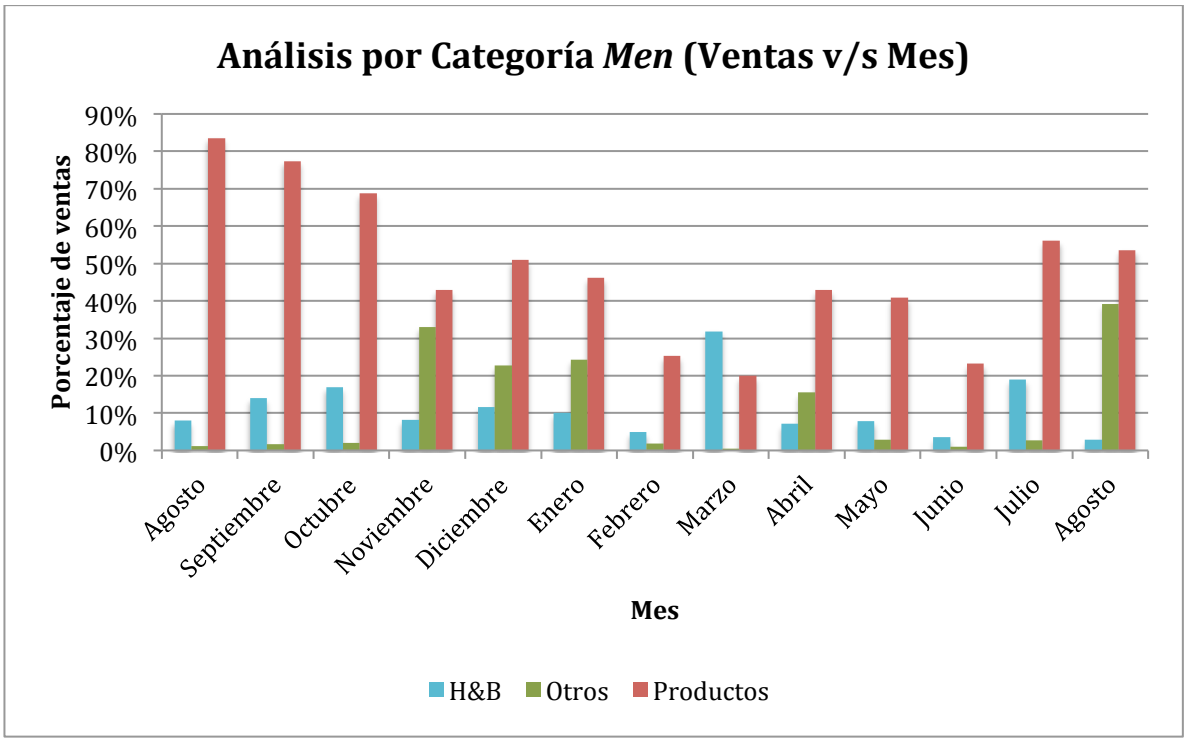
[25] M. Bosch, A. Musalem, "Análisis de Interrelaciones en las Canastas de Compras en un Supermercado", *Revista Ingeniería de Sistemas,* Vol. XV, Número 1, 2001.

8. ANEXOS

Anexo 1: Análisis por Categoría *Men* y *Default*

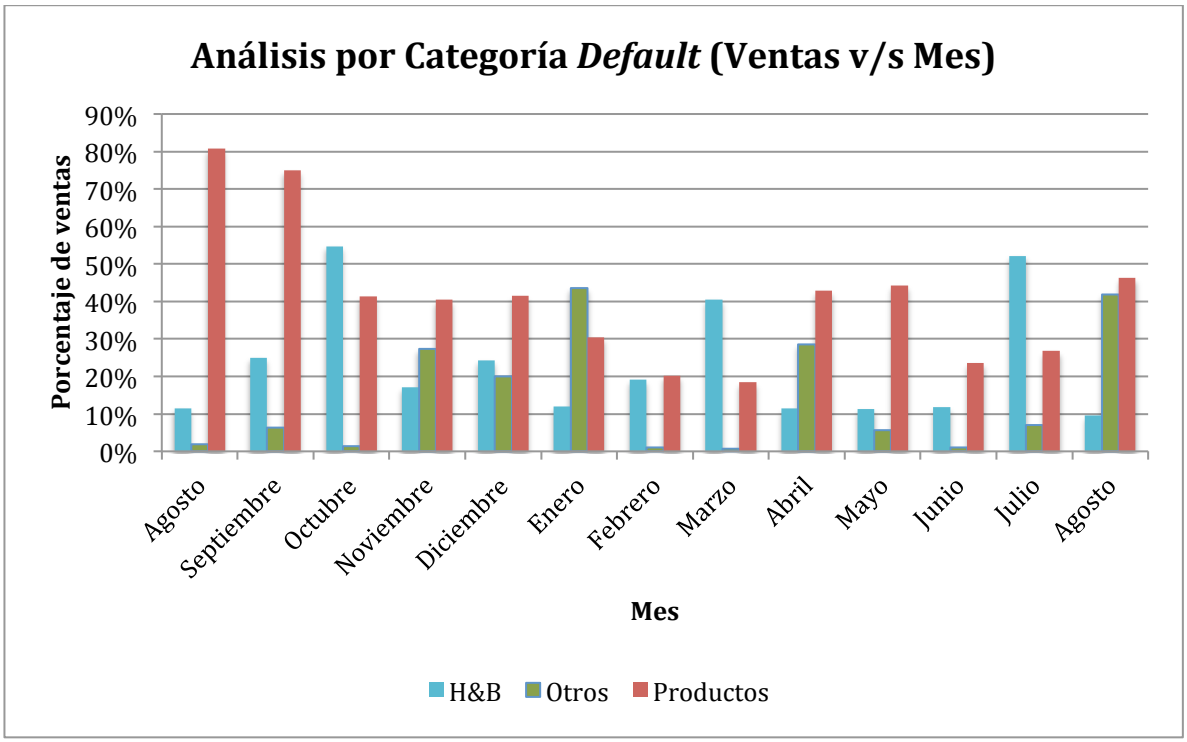
Análisis por Categoría *Men*

Se puede observar que *Productos* es la categoría con mayores porcentajes de venta. El porcentaje predeterminado del *Newsletter* de *Men* expuesto en la Tabla 1.1, es de un 25%, que comparado con el promedio de los porcentajes de ventas de esta categoría, 49%, se está exhibiendo casi la mitad de productos de lo que se debería. Al igual que en el caso de *Women*, las categorías tienen una alta desviación estándar, un 19,7% para la categoría *Productos*. Sin embargo, la categoría *Health & Beauty* tiene una desviación estándar menor a las demás, llegando a un 7,84%, donde se podría determinar más precisamente cuánto deberían comprar. El promedio del porcentaje de ventas en esta categoría es de un 11%, frente a un 18% que es lo que actualmente se exhibe.



Análisis por Categoría *Men*

Esta ilustración se puede comparar con la Ilustración 1.1, donde el comportamiento de los clientes es similar. A esto se le puede atribuir que la mayoría de los clientes suscritos a esta lista son mujeres que nunca declararon su género. Sin embargo, al tener esta distribución de ventas, el *Newsletter* destinado a esta lista tampoco debería tener un formato predeterminado.



Anexo 2: Distancias y Similitudes

i. Distancia Euclidiana:

La Distancia Euclidiana es la métrica más simple que existe para medir similitudes. Esta es conocida por medir la distancia entre dos puntos en un plano mediante la siguiente fórmula:

$$d(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Como ejemplo, se muestra la siguiente tabla:

	Producto 1	Producto 2	Producto 3	Producto 4	Producto 5
Usuario 1	5	4	1	2	1
Usuario 2	3	4	1	3	4

En este ejemplo, se tiene que la Distancia Euclidiana se calcula como sigue:

$$d(U1, U2) = \sqrt{(5-3)^2 + (4-4)^2 + (1-1)^2 + (2-3)^2 + (1-4)^2}$$

$$= \sqrt{14}$$

$$= 3{,}74$$

La distancia entre el Usuario 1 y el Usuario 2 es de 3,74. Esto se repite para cada uno de los usuarios con cada uno de los demás. Por ejemplo, si existen 100 usuarios, cada uno medirá su distancia con los 99 restantes. Los usuarios que tengan una distancia más cercana a 0, serán los más similares.

ii. Distancia Euclidiana Cuadrática:

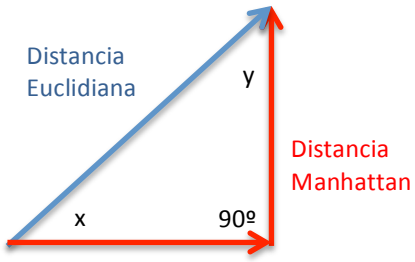
La Distancia Euclidiana Cuadrática corresponde, como lo dice su nombre, al cuadrado del valor de la Distancia Euclidiana. Su valor es calculado mediante la siguiente fórmula:

$$d(X, Y) = \sum_{i=1}^{N} (x_i - y_i)^2$$

En el ejemplo anterior, el valor de la Distancia Euclidiana Cuadrática es de $\sqrt{14}^2 = 14$. Al igual que en la Distancia Euclidiana, los usuarios que tengan una distancia más cercana a 0, serán los más similares.

iii. Distancia Manhattan:

La Distancia Manhattan también es conocida por medir dos puntos en un plano, pero no calculando la distancia diagonal entre estos (Distancia Euclidiana), sino que en distancias horizontales o verticales. Como ejemplo se muestra la Ilustración 2.1.



Esta distancia es calculada mediante la siguiente fórmula:

$$d(X, Y) = \sum_{i=1}^{N} |x_i - y_i|$$

En el mismo ejemplo propuesto anteriormente, la Distancia Manhattan corresponde a 6. Los usuarios más similares son los que la distancia es más cercana a 0.

iv. Distancia Coseno:

Para utilizar la Distancia Coseno, es necesario pensar en cada calificación como un punto en un plano, y trazar un vector desde el origen hasta dicho punto, con el fin de formar un ángulo $\boldsymbol{\theta}$ entre ambos vectores, como se muestra en la Ilustración 2.2.

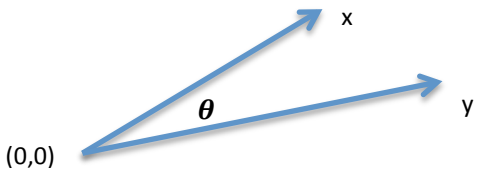


Ilustración 8.1: Ejemplo de Distancia Coseno

Para calcular la Distancia Coseno se utiliza la siguiente fórmula:

$$d(X,Y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}} = \frac{X \times Y}{||X|| \cdot ||Y||}$$

Utilizando el mismo ejemplo de la tabla anterior, el valor de similitud entre ambos usuarios es de 0,86. En las medidas anteriores, los valores podían variar desde 0 hasta el infinito, pero en este caso, sólo puede ir desde 0 hasta 1, donde 1 corresponde a una similitud perfecta, debido a que el vector va exactamente en la misma dirección que el otro y 0 a una disimilitud perfecta, ya que ambos vectores forman un ángulo de 90º.

	Producto 1	Producto 2	Producto 3	Producto 4	Producto 5
Usuario 1	10	8	6	6	8
Usuario 2	5	4	3	3	4

$$d(U1,U2) = \frac{10 \cdot 5 + 8 \cdot 4 + 6 \cdot 3 + 6 \cdot 3 + 8 \cdot 4}{\sqrt{(10^2 \cdot 8^2 \cdot 6^2 \cdot 6^2 \cdot 8^2) \cdot (5^2 \cdot 4^2 \cdot 3^2 \cdot 3^2 \cdot 4^2)}}$$

$$= \frac{150}{\sqrt{300 \cdot 75}}$$

$$= \frac{150}{\sqrt{22.500}}$$

$$= \frac{150}{150}$$

$$= 1$$

Considerando esta tabla, se puede observar que las calificaciones son distintas, pero que la Distancia Coseno calza perfectamente con cada uno de éstos. Esto sirve para poder

enlazar usuarios que se mueven dentro de pocos rangos de calificaciones con otros que no. Por lo tanto, es recomendable utilizar esta medida cuando se quieren encontrar similitudes entre usuarios regularizando las calificaciones.

v. Distancia Jaccard:

A diferencia de las distancias mencionadas anteriormente, que capturan la distancia o el ángulo entre los distintos usuarios, el Coeficiente de Tanimoto, o también conocida como la Distancia Jaccard, considera la distancia y el ángulo en el cálculo de la similitud.

Se calcula mediante la siguiente expresión:

$$T = \frac{\sum_{i=1}^{N}(x_i \wedge y_i)}{\sum_{i=1}^{N}(x_i \vee y_i)}$$

Para mostrar esta expresión de una forma más explicativa, se muestra la siguiente tabla:

	Producto 1	Producto 2	Producto 3	Producto 4	Producto 5
Usuario 1	X	X	X	X	-
Usuario 2	-	X	X	-	-

Para determinar la Distancia Jaccard, se calcula lo siguiente:

$$\sum_{i=1}^{N}(x_i \wedge y_i) = 2$$

$$\sum_{i=1}^{N}(x_i \vee y_i) = 4$$

$$T = \frac{2}{4} = 0{,}5$$

Considerando este ejemplo, la distancia entre el Usuario 1 y el Usuario 2 es de 0,5. Esta distancia va en un rango entre 0 y 1. Si el valor es cercano a 1, los usuarios son similares, y cuando es cercano a 0, son disímiles.

vi. Coeficiente de Correlación de Pearson:

El Coeficiente de Correlación de Pearson mide la tendencia de dos series de números que se muevan en la misma dirección. En otras palabras, si dos usuarios tienen calificaciones similares o simplemente similares proporcionalmente (al igual que la Distancia Coseno),

serán similares según esta métrica.

Para calcular este coeficiente, se utiliza la siguiente expresión:

$$d(X,Y) = \frac{N\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)\,(\sum_{i=1}^{N} y_i)}{\sqrt{(N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2)(N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2)}}$$

Utilizando el ejemplo de las métricas anteriores, el valor de el coeficiente es de 1. Este coeficiente puede estar entre los valores -1 y 1, donde -1 corresponde a una perfecta disimilitud y 1 a una perfecta similitud.

vii. Correlación de Spearman:

Esta medida de similitud es una variante del Coeficiente de Correlación de Pearson, ya que se utiliza la misma expresión, pero cambiando las calificaciones de los usuarios. En primer lugar, se reemplaza por un 1 la calificación del último producto comprado, por un 2 el penúltimo producto comprado y así sucesivamente.

Al haber cambiado todas las calificaciones, se podrá observar una base de datos como la que se muestra en la tabla siguiente:

	Producto 1	Producto 2	Producto 3	Producto 4
Usuario 1	1	3	2	-
Usuario 2	2	3	1	4
Usuario 3	3	2	4	1
Usuario 4	-	2	1	3

Luego de concluir esta nueva base de datos, se calcula el Coeficiente de Correlación de Pearson, dando como resultado las similitudes entre los usuarios.

Anexo 3: Tipos de Segmentación

i. K-Means:

Este método es el más utilizado para hacer segmentaciones y consiste en hacer *k* segmentos según *k* centroides [21]. En general, se utiliza para encontrar *k* medias, haciendo que cada elemento se una a su media más cercana, para así dividirlos en *k* distintos segmentos. En el caso de un Sistema de Recomendación, una media o un centroide puede ser representado por un producto, una categoría o un usuario, donde se pueden encontrar grupos de estos, para luego generar un modelo de recomendación para cada uno de los segmentos.

Para determinar qué tan cerca está cada elemento de su media más cercana, se pueden utilizar las distintas distancias mencionadas en los Filtros Colaborativos en base a la memoria.

ii. X-Means:

X-Means [22] es una extensión del K-Means. Funciona de la misma manera, uniendo cada elemento con su media más cercana. Sin embargo, difieren en el *k* del K-Means. Este método calcula las medias sin exigirle *k* centroides, por lo que el algoritmo por sí solo genera *X* segmentos de acuerdo a la información de los elementos, lo que lo hace más conveniente en algunos casos.

iii. Fuzzy C-Means:

El Fuzzy C-Means [23] es un algoritmo de segmentación que, a diferencia de los dos descritos anteriormente, logra que un elemento pueda estar en uno o más segmentos, dependiendo de la información que se disponga. A este algoritmo se le debe asignar un número *k* de segmentos a dividir, al igual que el K-Means.

iv. Expectation-Maximization (EM):

El algoritmo EM [24] crea segmentos combinando distintas funciones Gaussianas a partir de una base de datos, donde cada una de estas tiene una media esperada y una matriz de covarianza. La probabilidad de cada función está determinada por la fracción del total de los elementos utilizados por esa misma función.

Este método sólo encuentra un óptimo local, lo que no siempre determina el óptimo global, por lo que hay que probar con distintos parámetros iniciales.

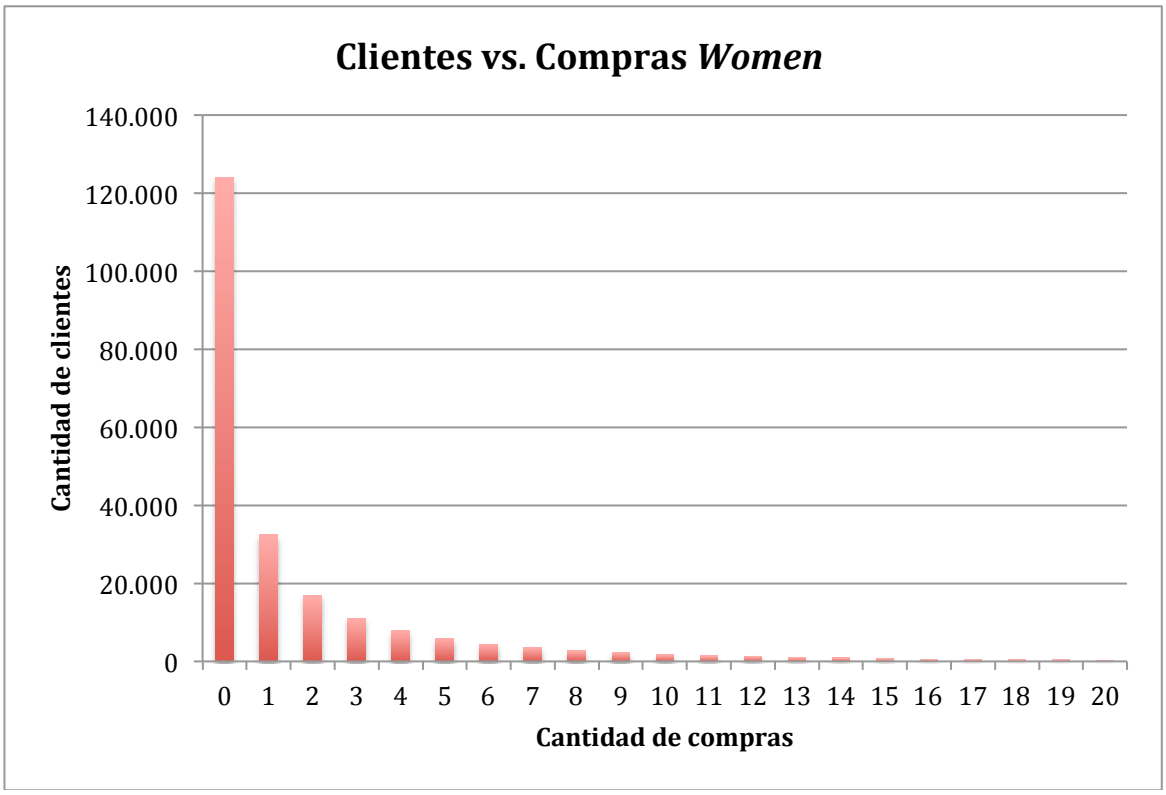
Anexo 4: Clientes vs. Compras *Women* y *Default*

Clientes vs. Compras *Women*

La lista *Women*, posee 223.004 suscritos y tienen un promedio de compras de 2,01 cada uno. En la tabla siguiente, se puede observar los datos actuales de cada uno de las métricas que se utilizaron para medir la eficiencia de los modelos generados.

Indicador	Porcentaje
Open Rate	15,21%
Unique Open Rate	11,7%
CTR (enviados)	2,1%
CTR (abiertos)	18,5%
TCTR	19,5%
Tasa de Conversión	2,2%

A continuación, se muestran la distribución de compras de los clientes asociados a la lista *Women*.

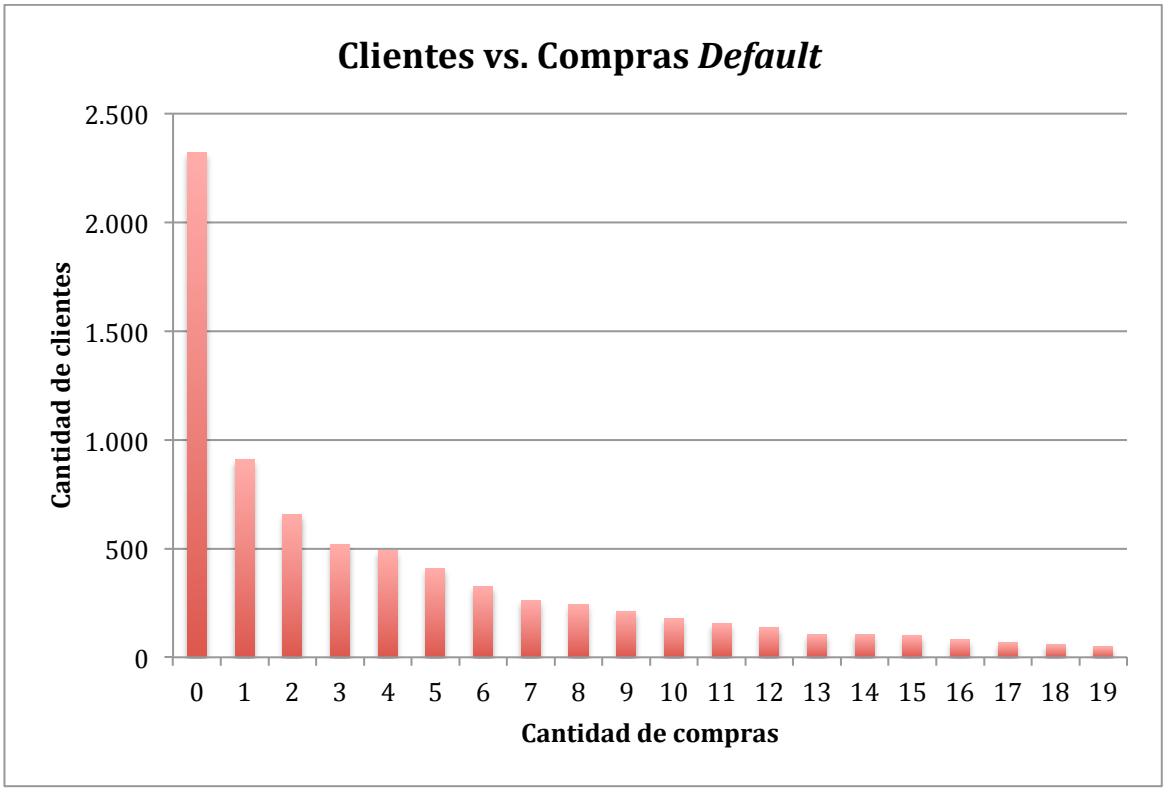


Clientes vs. Compras *Default*

La lista *Default* posee 7.853 suscritos y tiene un promedio de compras de 5,62 cada uno, donde, a pesar de ser pocos clientes, tienen cerca de un 200% más de compras en promedio del resto de las listas. Su desempeño actual es el siguiente:

Indicador	Porcentaje
Open Rate	18,6%
Unique Open Rate	15,3%
CTR (enviados)	1,85%
CTR (abiertos)	12,1%
TCTR	12,8%
Tasa de Conversión	2,2%

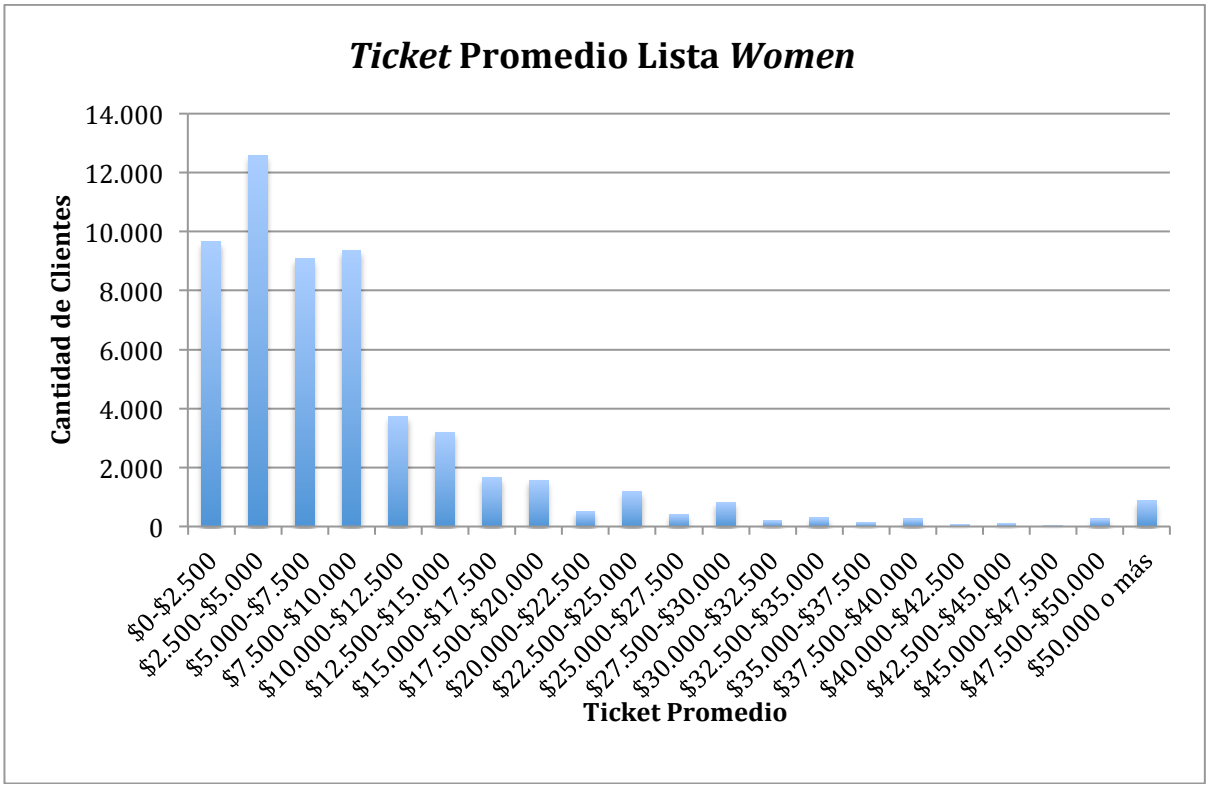
A continuación se muestra la distribución de compras de la lista *Default*, donde ocurre lo mismo que en las listas anteriores.



Anexo 5: Ticket Promedio de listas *Women* y *Default*

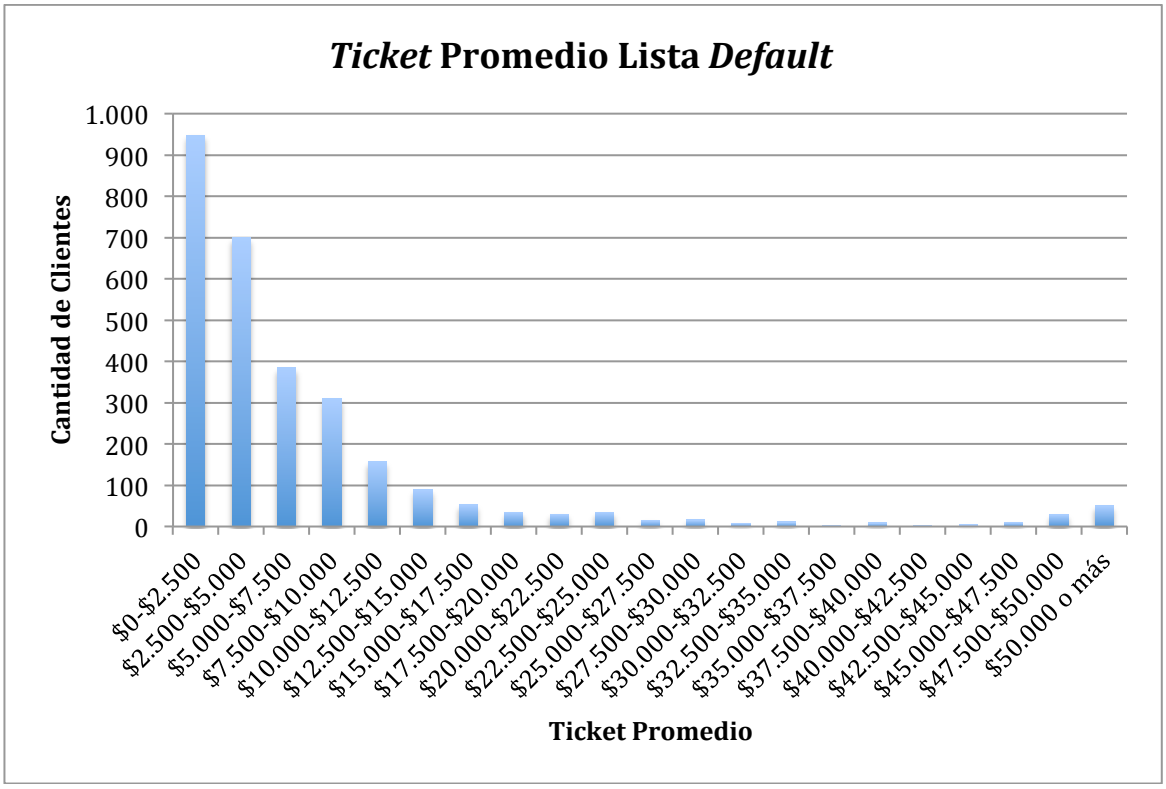
Ticket Promedio *Women*

Analizando la lista *Women*, que posee 223.004 clientes, el 25,1% de ellos ha realizado una o más compras. En el gráfico se puede observar que el comportamiento es similar al de la lista masculina, donde el 72,67% de ellos está dispuesto a comprar un cupón de hasta $10.000. Esto apoya la idea expuesta anteriormente, que se propone personalizar las ofertas según el comportamiento histórico de ellos, donde cada cliente recibe productos acorde a sus gustos y necesidades.



Ticket Promedio *Default*

La lista *Default*, que posee solo 7.853 clientes, correspondiente al 2,28% del total de activos, presenta un comportamiento distinto a las demás, debido a que los suscritos corresponden a clientes más nuevos que los otros y que nunca se han registrado completamente. El 36,2% de ellos ha realizado una compra alguna vez, y como se puede observar en el gráfico, están dispuestos a comprar cupones al menor precio posible, donde la curva es continuamente descendiente.



Anexo 6: Reglas de Asociación en *R*

Código de Reglas de Asociación en *R*

```
library(arules)
library(arulesViz)
dataset_categorias <- read.transactions("dataset_categorias_id.csv", sep = ";")
itemFrequencyPlot(dataset_categorias, topN = 20, type = "absolute")
rules = apriori(dataset_categorias, parameter = list(supp = 0.001, conf = 0.8, minlen
        = 2, maxlen = 2))
options(digits = 2)
inspect(sort(rules, by = "lift")[1:5])
plot(rules, method = "grouped")
```