

# ¿Cómo **reformular** la estrategia de comercio electrónico de una organización mediante el análisis de datos?

Arturo Palomino Gayete

PID\_00242822



Director de la colección: Lluís Pastor



El encargo y la creación de este material ha sido coordinado por el profesor: Josep Cobarsí (2017)

Primera edición: septiembre 2017

© Arturo Palomino Gayete  
Todos los derechos reservados  
© de esta edición, FUOC, 2017  
Av. Tibidabo, 39-43, 08035 Barcelona

Realización editorial: Oberta UOC Publishing, SL

Depósito legal: B-21.147-2017

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del *copyright*.

# Índice

**4      Cómo usar un modelo H2PAC**

**7      El reto**

**15     El conocimiento imprescindible**

17     1. Canal de distribución, el sector  
         del *retail* nacional

31     2. Análisis estratégico de información  
         en *retail*

**47     Las soluciones**

92     Bibliografía



# **Cómo usar un modelo H2PAC**

**Este modelo plantea resolver propuestas clave a partir de  
ACTIVIDADES. A continuación os explicamos cómo manejarlos  
por el material y sacarle partido a través de tres fases.**

1

## El reto

En las páginas iniciales encontrarás el **reto** que te plantea este material.

2

## El conocimiento imprescindible

En las páginas centrales encontrarás la **teoría imprescindible** que te ayudará a entender los conceptos clave y poder obtener las respuestas al reto.

3

## Las soluciones

En las páginas finales encontrarás el **solucionario** para resolver correctamente el reto propuesto.





**El reto**





Con este material se pretende poner al estudiante en la piel de un *data scientist* que quiere potenciar las acciones y optimizar las decisiones de una empresa de *e-commerce*, similares a Amazon, con el objetivo de incrementar su retorno de inversión (ROI) y protegerse de información desfavorable, empleando diferentes técnicas estadísticas y de minería de datos.

En el apartado «Conocimiento» definiremos los conceptos básicos que atañen al sector del *e-commerce* y el área en que se engloba, el sector del *retail* (supermercados, hipermercados, *discount*, especialistas, etc.). Veremos los agentes que están presentes en el sector tradicional y estrategias que han llevado al éxito a algunos de los principales representantes del sector –los *major players*– como Mercadona. Seguiremos con la importancia de la elección del producto y el plan estratégico de empresa. A continuación analizaremos más en detalle el presente y el futuro del *e-commerce*, con las nuevas tendencias que están revolucionando este sector tan incipiente. Seguidamente definiremos los conceptos del *data scientist*, los **sistemas de información**, la **inteligencia de negocio**, el **big data**. Finalmente, para completar el apartado de «Conocimiento», ampliaremos la información con un pequeño recopilatorio de los métodos más utilizados que están al alcance del investigador en el campo de la **estadística** y el **aprendizaje automático**, para extraer conocimiento de los datos. Estas técnicas son utilizadas por Amazon y otras empresas en línea: las **empresas .com**.

A lo largo del material, resumiremos el contenido teórico y seguiremos con dos casos prácticos en el apartado «Las soluciones», el primero basado en un sistema de **recomendación de productos** con el software libre R, con una fuente de datos de productos simulada equiparable a un fichero log de transacciones de compras reales en línea. Con el caso, el alumno experimentará con este tipo de algoritmos, que las empresas de *e-commerce* utilizan para recomendar productos a usuarios. Estos procesos analizan los hábitos y patrones de compra de los usuarios en un rango temporal amplio con el fin de extraer conocimiento y predecir acciones del consumidor. Defini-

remos el algoritmo *k-nearest neighbor* y el algoritmo *apriori*, y detallaremos los pasos que seguir para cumplir el objetivo de poner en marcha un sistema de recomendación de productos en tiempo real. Finalmente haremos un ejercicio de **minería de opinión** (*sentiment analysis*), mediante el cual analizaremos la opinión de los usuarios respecto a nuestra marca y nuestros productos, para que el alumno vea la importancia y el potencial de monitorizar todo aquello que está en la mente del consumidor respecto al posicionamiento de nuestra oferta y de nuestra imagen corporativa. El objetivo de este tipo de análisis está ligado a la comprensión de las decisiones de los usuarios con el fin de diseñar y mejorar un **plan de marketing** y generar mecanismos y planes de contingencia ante situaciones de crisis, por una falta de credibilidad o una opinión desfavorable generalizada del consumidor hacia la organización.

Para entender la importancia de este tipo de análisis debemos comprender el potencial en términos de volumen y valor en ventas que se prevé que generarán en un futuro próximo las empresas de distribución y venta de productos a través de Internet. Como veremos más adelante, el sector gran consumo engloba grandes categorías de bienes que el consumidor compra habitualmente en tiendas y supermercados; podemos encontrar su clasificación en el *Real Decreto 367/2005, de 8 de abril, por el que se desarrolla el artículo 17.3 de la Ley 7/1996, de 15 de enero, de ordenación del comercio minorista, y se definen los productos de alimentación frescos y perecederos y los productos de gran consumo*<sup>1</sup>. Este mercado mueve anualmente setenta y dos mil millones de euros en España, y aglutina productos de alimentación, bebidas, perfumería y productos para mascotas. El canal Internet genera más de quinientos millones de euros, tiene un crecimiento entre 2015 y 2016 del 26%, y supone un 1% del gasto total realizado en *retail*. El sector de *e-commerce* todavía es muy pequeño, pero los expertos señalan que, con este crecimiento, es evidente su enorme potencial para distribuidores y fabricantes.

---

1 [https://www.boe.es/diario\\_boe/txt.php?id=BOE-A-2005-6795](https://www.boe.es/diario_boe/txt.php?id=BOE-A-2005-6795)

El sector de *retail* hace años que tiene plataformas de venta en línea, pero los distribuidores que solo venden de este modo desempeñan un papel clave para este crecimiento en España. Internacionalmente, en Estados Unidos, por ejemplo, las ventas del canal en línea alcanzan una cuota de mercado superior al 15%. En Europa, los países que lideran el *ranking* son Reino Unido y Francia, ambos con cuotas por encima del 5%.

Las plataformas de venta en línea exclusiva son un sector con mucha evolución y continuas mejoras. Las empresas diversifican sus productos con el fin de atraer nuevos compradores y para reducir el riesgo. Recientemente, Amazon anunció el lanzamiento de su servicio *Prime Now*. Con este servicio, la empresa pasa a engrosar su oferta, ofreciendo productos envasados de gran consumo que suministra mediante envío a domicilio en una o dos horas, tras la realización de un pedido en su web, como si de una cesta de la compra se tratase. Paralelamente proliferan otras webs como Ulabox, Tudespensa.com, Deliberry y Comprea que generan fricciones en el sector y empujan a los grupos tradicionales (Carrefour, El Corte Inglés, Eroski, Dia, Lidl) a renovar y relanzar sus plataformas, así como a mejorar sus procesos operativos y logísticos.

La globalización y sobre todo Internet han potenciado en las últimas décadas, gracias a la difusión de las tecnologías de la información, la proliferación de estas webs y la aceptación por parte del consumidor del uso de este canal de compra, pese al recelo inicial. Recientemente y como se puede observar por las acciones de estos grandes grupos, el canal está en pleno crecimiento, y es evidente el auge experimentado por el comercio electrónico. Este cambio en los hábitos del consumidor no ha sido repentino; muy al contrario, el crecimiento ha sido gradual, lento y muy dependiente de las mejoras en el sector tecnológico, y más concretamente, en la evolución experimentada en Internet en los últimos años.

Internet facilita otro tipo de relaciones, no solo comerciales, sino también de fidelización y los sistemas de información, basados en

*Consumer Relationship management* (CRM), reciben también un renovado impulso con la creación de nuevas herramientas adaptadas a Internet y el análisis de información proveniente de las transacciones que se llevan a cabo en los servidores que alojan las webs. En este campo, aparecen sistemas de información como Google Analytics, Adobe Omniture o IBM Digital Analytics (Coremetrics) que ayudan a entender las preferencias, las decisiones y las motivaciones de los internautas en su navegación. Es de especial interés entender el proceso de compra de los visitantes en las plataformas de *e-commerce*.

Un sector tan dinámico como el *e-commerce* ha experimentado cambios estructurales empresarialmente, poco frecuentes hasta este momento en el *retail* tradicional. Amazon, por ejemplo, con los años se ha ido transformando: empezó como un canal de venta de libros en línea (*e-books*), pasando por la venta de CD de música, venta de productos informáticos, y recientemente venta de productos envasados de gran consumo o incluso proveedor de música en *streaming*. El alcance del negocio de Amazon fue de cien mil millones de dólares en 2015, prácticamente el doble de la facturación en gran consumo en España. Esta facturación solo es comparable con la de empresas como Facebook, Google o Apple, y junto con ellas forman el grupo que recibe el acrónimo de GAFA (**G**oogle, **A**ma-**z**on, **F**acebook y **A**pple).

Con el paso de los años, la tecnología ha sido uno de los puntales de éxito de la empresa de Jeff Bezos (creador de Amazon), innovando con diferentes técnicas de minería de datos, aprendizaje automatizado y algoritmos de optimización. Han creado sistemas de recomendación para sugerir productos, han usado minería de datos y reconocimiento de patrones, modelización y predicción para anticipar compras y optimizar el envío de productos a sus almacenes locales (*anticipatory shipping*) con el fin de minimizar los tiempos de espera. También, como el resto de las empresas del sector tecnológico, analiza las opiniones y monitoriza las redes sociales y foros para detectar tendencias en la venta de productos y la eficiencia de su oferta actual.

A lo largo del material esbozaremos una suerte de introducción al mundo del *e-commerce*, haciendo hincapié en la reciente noticia de la incursión de Amazon en el mercado español de gran consumo, así como la creación en Madrid del primer centro de operaciones Europeo (HUB) del gigante norteamericano. Veremos qué técnicas pueden emplear este tipo de empresas para hacerse un hueco importante en la distribución de productos de gran consumo en España y veremos algunas de estas metodologías.





**El conocimiento  
imprescindible**



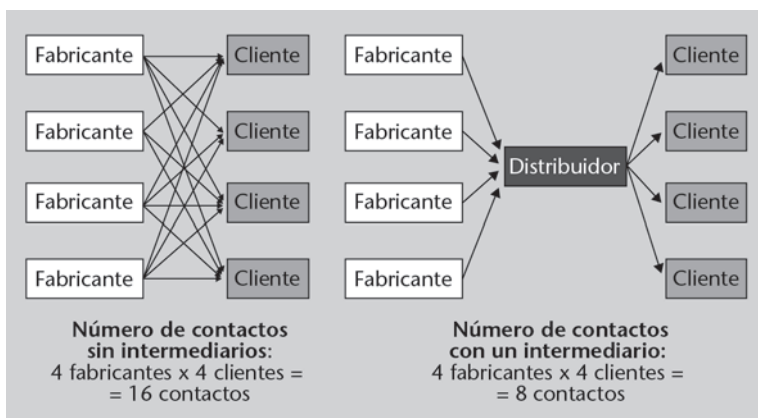


# 1. Canal de distribución, el sector del *retail* nacional

En este apartado veremos la utilidad de la figura del distribuidor en el proceso de venta de productos. Ya hemos definido el sector de gran consumo y los productos que lo componen; aquí definiremos los canales de distribución y en concreto el sector *retail*, que es aquel que se encarga de vender los productos de gran consumo.

Los canales de distribución surgen como consecuencia de la necesidad de separar los procesos meramente productivos de los de intercambio comercial entre cliente y fabricante. La importancia de esta separación radica en la reducción del número de intermediarios, de tal forma que varios fabricantes pueden vender sus productos a un único distribuidor, el cual los pone a disposición del consumidor final a cambio de un margen de beneficio en la transacción.

**Figura 1.** Los intermediarios minimizan los contactos



Fuente: Inma Rodríguez Ardura, Guillermo Maraver Tarifa y Francisco J. Martínez López (2005). *Canales de distribución* (pàg. 15). Barcelona: UOC.

Para el productor, el flujo comercial se reduce a un proceso entre dos únicos agentes; además, la especialización en las funciones productivas le permite optimizar sus productos y, en muchos casos, evita incurrir en el gasto adicional que supondría la necesidad de disponer de establecimientos propios para la venta directa. Como vemos en la figura 1, el distribuidor evita la necesidad que el fabricante tenga que establecer contactos con los clientes finales.

Entendemos por **longitud del canal** el número de intermediarios existentes entre el fabricante y el consumidor.

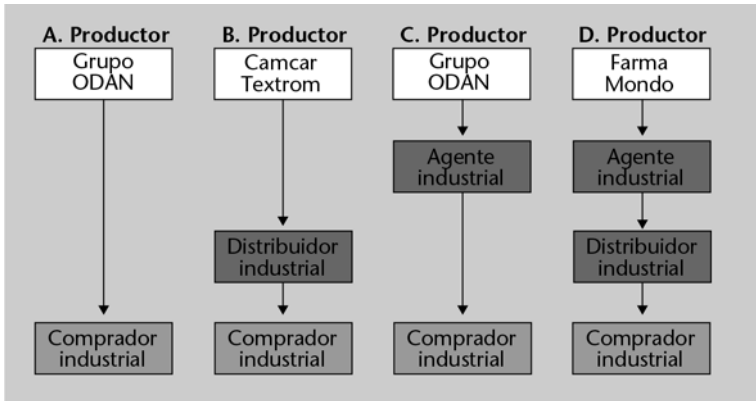
Los canales de distribución se pueden subdividir en función de la longitud del canal en:

- **Canales de distribución directos:** el flujo es directo entre fabricante y consumidor.
- **Canales de distribución indirectos cortos:** el flujo es indirecto entre el fabricante y el consumidor; aparece la figura del minorista y en algunos casos se suma la del mayorista.
- **Canales de distribución indirectos largos:** el flujo es indirecto entre fabricante y consumidor; además del minorista y el mayorista, aparece el agente comercial.

En la figura 2 vemos las diferentes estrategias de estructura de distribución por las que puede optar el fabricante.

En ocasiones, un fabricante opta inicialmente por una estrategia indirecta larga y, con el tiempo, decide integrar en sus funciones las de los agentes intermediarios con los que interactúa, en lo que conocemos como **integración vertical**. Esto suele darse cuando un mercado pasa a ser maduro y la empresa ha superado la etapa de internacionalización o de implantación con éxito. Llegado ese momento, la empresa tiene un mayor conocimiento de los diferentes factores regionales y asume las funciones del agente comercial; en algunos casos, puede llegar a conseguir integrar las funciones del distribuidor industrial modificando de esta forma los acuerdos del canal; a veces, sencillamente, se eliminan las funciones.

Figura 2. Tipos de canales de distribución



Fuente: Inma Rodríguez Ardura, Guillermo Maraver Tarifa y Francisco J. Martínez López (2005). *Canales de distribución* (pàg.38). Barcelona: UOC.

Por otro lado, también es habitual el paso contrario, que consiste en añadir un canal de distribución; esto suele suceder cuando alguna innovación tecnológica propicia la **creación de un nuevo canal**, con aceptación masiva por parte del consumidor. Un ejemplo sería la venta por teléfono de productos tales como: servicios de telefonía, seguros, cursos.

El uso de una nueva tecnología por parte del consumidor, es contemplado por el distribuidor como una oportunidad para generar nuevos ingresos. Si es capaz de adaptarse a los nuevos hábitos de consumo antes que sus competidores, conseguirá nuevos compradores y evitará la pérdida de clientes. El primer distribuidor que incorpora el nuevo canal y genera los procesos que facilitan la transacción con éxito genera una ventaja competitiva, que puede llegar a suponer un cambio disruptivo en el entorno competitivo.

Otro ejemplo claro de aparición de un nuevo canal adoptado masivamente por las empresas es Internet. Gracias a ella, aparece el **comercio electrónico** que permite a las empresas crear escaparates virtuales de sus productos (con fotos publicadas en su web), dar toda la información posible mediante correo electrónico o a través de descarga de documentos PDF, y realizar la transacción comercial

mediante terminales de punto de venta en línea para realizar pago con tarjeta o contrareembolso. Es decir, el fabricante puede fácilmente realizar las funciones de venta directa a escala internacional sin incurrir en grandes costes. También el distribuidor puede ahorrar costes de distribución haciéndose más competitivo en precios y, por tanto, atrayendo a más clientes que a través de los canales tradicionales. Surgen entonces empresas especializadas en la venta por Internet como Amazon, Ebay, Jet y muchas otras.

En el proceso de decisión de la elección del mejor diseño de sistema de distribución posible para un fabricante es importante plantear un **enfoque analítico por etapas** donde se debería valorar, en primer lugar, la documentación existente sobre los posibles canales de distribución. En paralelo es necesario analizar y comprender el sistema actual. En esta primera etapa, será importante organizar talleres y entrevistas sobre el canal de distribución existente, con el fin de valorar las opciones que mejor se adapten. Del mismo modo, se analizarán los canales de distribución de la competencia, mientras se está pendiente de nuevas oportunidades a corto plazo, desarrollando un plan de ataque de este modo, siempre que sea necesario. Así como es importante conocer las propias limitaciones y posibilidades, no hay que perder perspectiva respecto a lo que pide el cliente; en este sentido, es importante realizar un **análisis cualitativo** de las necesidades del usuario final por medio de reuniones de grupo y entrevistas personales, o bien realizar análisis cuantitativos mediante encuestas y otras técnicas. Más allá del perímetro de negocio, también hay que analizar y comprender las soluciones que ofrecen empresas de otras industrias cercanas. Tras haber analizado esta información, los directivos deberán desarrollar un **canal de marketing ideal**. Se comparará el canal ideal respecto al resultado del examen de la solución actual y los canales de la competencia para analizar las divergencias y, en función del resultado, identificar y desarrollar las opciones estratégicas que nos permitan diseñar el canal óptimo.

Si una empresa ha pensado en ganarse un puesto en un sector tan maduro debe tener muy bien definida su estrategia, ya que, sin duda, las demás empresas ya establecidas no estarán dispuestas a ponérselo fácil.

## 1.1. Agentes tradicionales del sector *retail* y estrategias de éxito

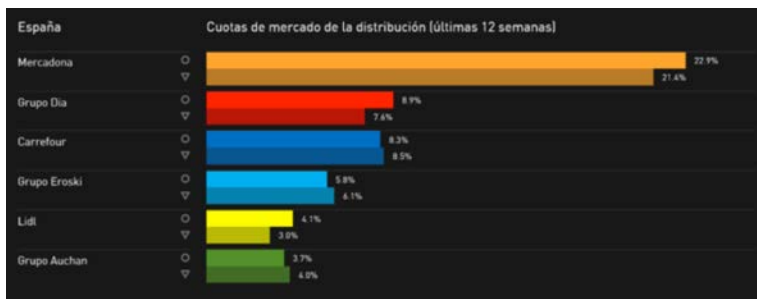
En este apartado veremos el esquema actual de la distribución en España y las acciones que han llevado a los *major players* a lograr el éxito. El sector *retail* está en continua evolución, el *ranking* de ventas lo controlan un número reducido de empresas que se reparten el mercado de la distribución.

En general, en alimentación, bebidas, droguería y perfumería, las empresas se suelen clasificar en función de diferentes criterios relativos a la estrategia de precios, tamaño y ubicación, en lo que entendemos por la clasificación del canal de distribución. En el lenguaje de estudios de mercado, la clasificación más habitual es la siguiente:

- **Híper:** en este grupo encontramos grandes distribuidores como Eroski, Carrefour, Alcampo o Caprabo. Se caracterizan por tener una presencia limitada, ubicados habitualmente en zonas de la periferia o áreas muy concurridas de la ciudad, con un mayor espacio de venta, con mayor surtido y variedad de productos.
- **Súper:** encontramos distribuidores habitualmente con mayor presencia en los barrios, de espacio reducido y con un nivel de surtido medio, que cubre una variedad de productos suficiente, para realizar compras de abastecimiento, de *stock* y de primera necesidad.
- **Discount:** es el grupo de distribución con elevada presencia en barrios de ciudad con espacios reducidos y nivel de surtido muy enfocado a productos básicos, de primera necesidad y de precios muy reducidos.
- **Especialistas:** engloban tiendas de barrio especializadas en tipos de productos muy concretos, entre las cuales encontraríamos charcuterías, fruterías, verdulerías.
- **En línea:** tiendas tradicionales que ofrecen sus productos a través de este medio y empresas .com dedicadas a la venta exclusiva por este canal.
- **Otros:** farmacias, parafarmacias, etc.

En España, el *ranking* de ventas por distribuidor del sector *retail* lo lidera Mercadona, empresa dirigida por Joan Roig, fundada en 1977<sup>1</sup> seguido por grupo Dia, Carrefour, Eroski, Lidl y grupo Auchan. En la figura 3, vemos un *benchmark* del sector en España del primer semestre de 2016.

**Figura 3.** *Benchmark* del mercado de distribución primer semestre de 2016



Font: aplicación de difusión abierta de Kantar worldpanel en <http://www.kantarworldpanel.com/es/grocery-market-share/spain>

Este escenario no siempre había sido así: en el pasado había uno muy diferente al actual en el que Mercadona ocupaba la cola del *ranking*<sup>2</sup>; sin embargo, un gran acierto en las diferentes estrategias consolidó la empresa de Joan Roig como el nuevo líder del sector en España. En la figura 4 se observa dicha evolución.

Previo a la llegada del distribuidor valenciano, dominaban el sector empresas como El Corte Inglés, Carrefour, Dia y Eroski. Las estrategias seguidas por Joan Roig, que han influido en gran medida en este liderazgo, podemos sintetizarlas en el siguiente esquema, reflejado en la figura 5.

1 [https://es.wikipedia.org/wiki/Juan\\_Roig\\_Alfonso](https://es.wikipedia.org/wiki/Juan_Roig_Alfonso)

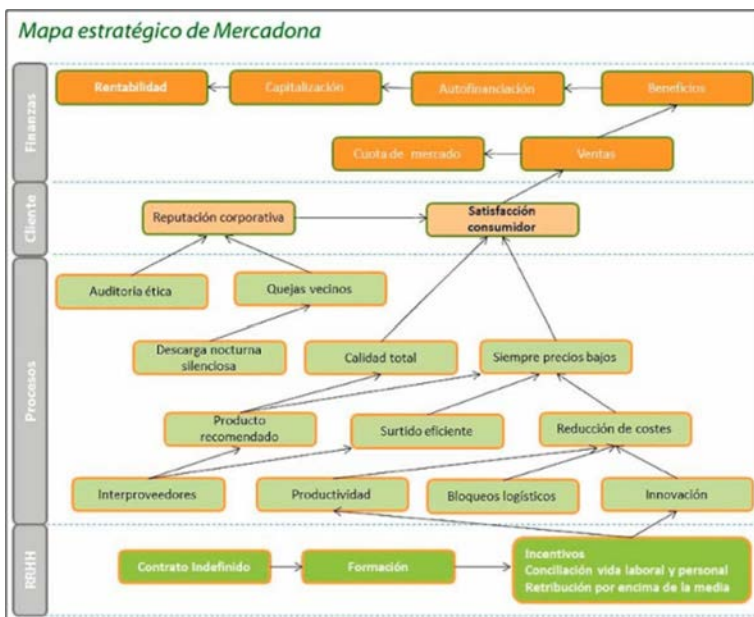
2 <http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modo-lo-de-exito-de-mercadona.html>

**Figura 4.** Evolución de ventas de Mercadona hasta 2012



Fuente: «El mundodelaempresa» en <http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html>

**Figura 5.** Mapa estratégico de Mercadona



Fuente: «El mundodelaempresa» en <http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html>

Como vemos, en la base se sitúa el personal, que actúa como motor de la productividad de la empresa y la innovación. Esta última, por tanto, es uno de los puntos diferenciales respecto a esquemas tradicionales. La innovación redundante en la reducción de costes. En paralelo, la gestión eficiente del surtido, proveniente de los interproveedores (proveedores de marca blanca), acciona la palanca de mayor valor estratégico del plan de acción de Mercadona, a la vez que su eslogan «Siempre precios bajos», puntal fundamental del éxito de la cadena.

Por tanto, podemos decir que, como es lógico, la productividad es un factor importante para el éxito; sin embargo, si una organización quiere destacar, además debe tener muy en cuenta el **surtido**, la **innovación** y los **precios**.

Estos son los puntos clave diferenciadores en la base de los procesos de la estrategia seguida por la ejecutiva de Mercadona.

**Figura 6.** «Siempre precios bajos»: eslogan de la cadena valenciana



Fuente: «El mundodelaempresa» en <http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html>



En el factor del surtido de productos, la relación de Mercadona con sus interproveedores es un caso de integración vertical, en que la cadena recurre a los fabricantes, que elaboran los productos de gran calidad que son integrados en el paraguas de marcas blancas de la cadena, Hacendado, Deliplus, Compys, Bosque Verde, Como Tú, 9.60, Dermik y Solcare, para atender las categorías de alimentación, cosmética, alimentación de mascotas, droguería, perfumería, dermoestética y protección solar.

Un nuevo competidor que desee entrar a competir en el mercado de la distribución en España deberá fijarse en las estrategias que mejor han funcionado regionalmente para tratar de replicarlas o mejorarlas.

## **1.2. Planes de empresa y la importancia de elección del producto**

A continuación veremos cómo se clasifican los productos y la importancia de conocer las categorías en las que vamos a entrar a competir para delimitar e identificar mejor a nuestros rivales. En el proceso de gestación de una empresa, la decisión de elección de combinación de factores productivos no es una tarea sencilla; a menudo el proceso que hemos definido no es el que va a obtener los mejores resultados, o bien porque no es lo que está buscando el consumidor, o bien porque la fórmula elegida ya existe con una combinación de costes y calidad más competitivos que los nuestros. En este sentido es importante definir un **plan de empresa**, que habitualmente se plasma en un documento donde se recoge la idea de proyecto que se pretende poner en marcha y que plasma la definición de la idea de negocio al mayor detalle posible.

Esta herramienta no es un documento estándar, sino que suele seguir un esquema con unos conceptos comunes a la mayoría de los proyectos de negocio. Se suelen tener en cuenta tres grandes conceptos: la **idea de negocio**, el **producto** y la **comercialización**.

En general es un documento dinámico que suele variar en el tiempo a medida que se mejora el planteamiento. Cabe destacar que tener

un plan de empresa bien definido suele servir para obtener financiación, socios o colaboradores.

Cualquiera de los puntos del documento son esenciales –financiación, selección de personal, forma jurídica, etc.–, pero sin duda el producto merece ser analizado en detalle.

Los productos son esenciales en el flujo circular de la renta del mercado de bienes y servicios.

Las empresas generan ingresos ofreciendo un producto atractivo a los consumidores; estos, a su vez, deciden dedicar una parte de su renta a la compra de bienes y servicios.

La primera gran división de actividad de la empresa está basada en el tipo de oferta, que puede ser de dos tipos: el primer grupo es el de las empresas de bienes; el segundo es el de las empresas que ofrecen servicios; es decir, el tipo de oferta según su tangencia. Dentro de la **oferta en bienes** podemos clasificar los productos según durabilidad: en **bienes perecederos** y **bienes imperecederos**. En el primer grupo tendríamos los productos de consumo relativamente rápido, como el café, el jabón, las frutas; en el segundo grupo tendríamos productos que pueden usarse varias veces, como los electrodomésticos, los coches o los ordenadores.

Según los principales institutos de estudios de mercado, la clasificación de bienes perecederos incluye productos de alimentación y bebidas, droguería, perfumería familiar, alimentación para mascotas y productos para el bebé; es decir, los productos que habitualmente encontramos en los distribuidores de gran consumo<sup>3</sup>.

---

3 <http://www.promonegocios.net/producto/tipos-productos.html>,  
<http://www.kantarworldpanel.com/es/Noticias/El-Gran-Consumo-se-estabiliza-en-el-segundo-trimestre-cuotas-distribucion-junio>

Una empresa de *e-commerce* que venda productos de informática o electrodomésticos obviamente no necesitará fijarse en Mercadona, Eroski o Lidl, ya que su entorno competitivo es muy distinto; en este sentido, será muy importante tener muy claro la categoría de productos que vamos a ofrecer.

### 1.3. El *e-commerce* y las nuevas oportunidades

En este apartado veremos el sector del e-commerce con más detalle, haciendo énfasis en las nuevas oportunidades que vienen acompañadas de nuevas mejoras tecnológicas y en los hábitos de los consumidores. En los últimos años ha habido una evidente evolución tecnológica que ha sido motor de impulso de crecimiento del sector que ha derivado en un nuevo esquema de distribución que conocemos como *e-commerce*<sup>4</sup>. Su característica principal es el medio a través del cual se realiza el contacto con el cliente, la oferta, el pedido y el medio de pago: Internet en cualquiera de sus formas, mediante PC, *Tablet* o *Smartphone*.

Entre las empresas que utilizan este canal encontramos distribuidores tradicionales que se adaptan a las nuevas tecnologías y crean su propia web de *e-commerce* (Carrefour en línea, Eroski en línea, DIA en línea...), pero, por otro lado, también encontramos empresas *.com* de distribución en línea, originadas en Internet, para venta de productos a bajo precio, que readaptan sus surtidos y empiezan a ofrecer productos de gran consumo, mediante estrategias de diversificación.

Un claro ejemplo es el de Amazon, que se gestó originalmente como una empresa de venta de libros en línea y que en los últimos tiempos está ofreciendo todo tipo de productos, desde electrodomésticos y productos electrónicos, pasando por servicios *cloud computing*, hasta más recientemente incluso venta de bienes perecederos.

---

4 Worldpanel, K. (2014). *Accelerating the Growth of E-commerce in FMCG*. Kantar Worldpanel

En el otro extremo encontramos grandes superficies como el gigante norteamericano Walmart, que adoptan el nuevo canal con la compra de webs de *e-commerce* practicando la estrategia de integración vertical; por ejemplo, con la reciente adquisición de Jet.com para hacer frente a Amazon.

Con el fin de analizar las amenazas y las oportunidades, los agentes deben estudiar el entorno para planificar la entrada en el mercado, marcarse unos objetivos de crecimiento o identificar riesgos y amenazas. Un primer enfoque pasa por hacer un esquema del **grado de estabilidad del entorno** para evaluar la conveniencia de adoptar las diferentes estrategias o incluso valorar la posibilidad de abandonar el proyecto.

Las oportunidades evidentes para este tipo de empresas son claras por las cifras que hemos mencionado en el apartado 1. Con ellas es obvio que merece la pena hacerse un hueco y entrar a competir. La gran fortaleza que posee este tipo de empresa es el hecho de ser una marca ya establecida y conocida con un aparato logístico optimizado y con el enorme poder de negociación de una compañía que domina el comercio en línea internacionalmente. Sin duda, estos dos elementos del DAFO superan con creces los elementos en contra de debilidad y amenazas, ya que entran a competir en un tipo de canal muy incipiente en el que la peor amenaza son ellos mismos y donde carecen de las debilidades de otros negocios menos diversificados y regionales como los ya establecidos.

Como vemos, una empresa como Amazon tiene grandes argumentos a favor, y además cuenta también con la inercia de las nuevas tecnologías de las que son máximo exponente; un campo en el que los demás competidores lo tienen realmente complicado para alcanzar su nivel.

## **1.4. Nuevas tendencias y evolución del sector**

Hemos hablado de ejemplos de empresas que adoptan estrategias de diversificación de productos, estrategias de integración horizontal o

vertical; por otro lado, también es habitual el paso transversal, que consiste en añadir un nuevo canal de distribución. Esto suele pasar cuando alguna innovación tecnológica propicia la creación de un nuevo canal, con aceptación masiva por parte del consumidor. Un ejemplo sería la venta por teléfono de productos como servicios de telefonía, seguros, cursos. Otro ejemplo más reciente es la venta a través de Internet, que recibe el nombre de *e-commerce*. Una variante de este último canal es el de *m-commerce* o *e-commerce* a través del móvil<sup>5</sup>.

El comercio electrónico a través del móvil cobra especial interés en los años más recientes. Actualmente, los usuarios tienden, cada vez más, a realizar mediante dispositivos móviles inteligentes las funciones que habitualmente se llevaban a cabo mediante PC. Los *smartphones* y las *tablets* cada vez incorporan más funciones. En origen su gestación fue la fusión de dos tecnologías con funciones separadas, las *personal data assistant* o PDA (*pocket PC*) y los teléfonos móviles. Los primeros permitían gestionar datos, como agendas, contactos, pequeñas bases de datos e incluso dibujar notas a mano alzada con lápices ópticos. Poco a poco, los segundos fueron incorporando las funciones de los primeros hasta el punto de que no se concebía un teléfono móvil sin funciones de PDA ni un PDA sin funciones de telefonía móvil. Esta, a su vez, dotó a los dispositivos de la posibilidad de enviar y recibir datos por Internet abriendo un abanico enorme de posibilidades: abarcando ocio, trabajo y comercio, entre otros.

El *m-commerce* creció un 69% en 2015 en España respecto el año anterior: un crecimiento solo superado por Brasil según «Zanox Mobile Performance Barometer 2015 1.er semestre», que mide la evolución

---

5 <http://www.xataka.com/moviles/htc-una-historia-de-poco-ruido-y-muchas-nueces>  
<http://www.distribucionactualidad.com/el-m-commerce-crece-un-60-con-un-gasto-por-carrito-de-95-euros/>  
[https://www.emarketer.com/public\\_media/docs/eMarketer\\_Mobile\\_Commerce\\_Roundup\\_2016.pdf](https://www.emarketer.com/public_media/docs/eMarketer_Mobile_Commerce_Roundup_2016.pdf)  
<http://blog.zanox.com/en/zanox/2016/03/22/zanox-mobile-performance-barometer-2016/>

de este canal basándose en el análisis de más de 4.300 anunciantes, en once territorios. Globalmente, este crecimiento está por encima de un 140% de crecimiento; y el gasto total en PC y móviles ha crecido un 9% respecto el año anterior. Los *smartphones* son los dispositivos con mayor crecimiento en número de transacciones, respecto a *tablets* y los PC. Pero estos últimos continúan teniendo un *share* superior en el global de transacciones, aunque cediendo terreno a las nuevas tecnologías. Los gastos con un valor más elevado se realizan mediante *tablet*: posiblemente, estos aparatos den mayor confianza al consumidor en aquellas transacciones que comportan un gasto más elevado.

Las nuevas tecnologías suponen también un cambio en las estrategias de *marketing*. Las empresas de *marketing* tradicional acostumbradas a realizar campañas a través de los *mass media* tradicionales han tenido que adaptar sus procesos y productos a las nuevas tecnologías. Por su parte, el entorno digital ha propiciado que la oferta, las campañas y los análisis del retorno de inversión se hayan adaptado incorporando técnicas más complejas para:

- Mejorar la experiencia de compra.
- Analizar los efectos de las campañas.
- Optimizar los procesos de subasta de espacios publicitarios.
- Generar sistemas de recomendación para fidelizar clientes.
- Optimizar los tiempos de envío, *stock*.

Además de muchos otros procesos que años atrás se apoyaban principalmente en el *know how* de expertos en *marketing* y que actualmente están siendo articulados y reinventados en departamentos de inteligencia de negocio, de IT y de *data science*. Entre las técnicas utilizadas por estos departamentos cobran especial interés las de *big data management*, *bussiness intelligence*, *machine learning*, estadística y la *web analytics*.

## 2. Análisis estratégico de información en retail

### 2.1. Sistema de información, inteligencia de negocio y datos masivos

En este apartado veremos las herramientas de que dispone un *data scientist* para traducir los datos a conocimiento con el que dotar de elementos clave a la toma de decisiones de la organización. Entre los recursos que afectan a una empresa podemos encontrar factores tradicionalmente establecidos como los **humanos, financieros, materiales**. Recientemente, gracias a los avances tecnológicos, encontraríamos los intangibles, como el **valor de la marca, la investigación y la información**. Esta última es de especial importancia para la supervivencia de una empresa. Gracias a la información, la empresa es capaz de establecer un rumbo que le permita optimizar sus ventas y sortear los obstáculos que aparezcan en el entorno competitivo.

La empresa recibe información de su entorno por medio de agentes externos, pero también emite al exterior y genera información interna mediante métricas derivadas de sus operaciones diarias tanto de producción como de los propios recursos. Así, la empresa puede recoger información de fuentes de *open data* o de institutos de estudios de mercado si hablamos de fuentes externas; paralelamente, puede generar información contable, según el marco legal, y hacerla pública, o bien recoger información de las transacciones, gastos y medición de productividad, para trabajarla internamente, con el fin de detectar puntos débiles y fuertes.

La información se genera a partir de los datos, que se recogen a través de diversas fuentes; por ejemplo por medio de sensores en las máquinas productivas, o bien mediante personal de admi-

nistración con introducción manual en sistemas, entre muchos otros. Esos datos son procesados y convertidos en información. Algunos son descartados o guardados, pero no se emplean de forma inmediata (*data exhaust*). Para que un dato se considere ‘información’, debe tener un valor informativo. Por ejemplo, el número de habitantes por país o el número de ventas de una marca son información. En cambio una serie de datos aleatorios no tiene valor informativo. De esta forma, los datos ininteligibles se procesan y se convierten en información inteligible, comprensible y útil. Finalmente, de esa información inteligible y ordenada, disponible para analizarla, se obtiene conocimiento en forma conclusiones, previsiones y recomendaciones útiles para la toma de decisiones.

De este modo, los **sistemas de información** almacenan y permiten extraer datos relevantes para la toma de decisiones.

Por tanto, se trata de repositorios y herramientas en los que previamente a su gestión se han llevado a cabo procesamiento y preprocesamientos de datos mediante procesos por lotes (*batch*) o cargas (*bulk*) que han llevado a cabo a su vez una criba de información relevante de todas las posibles fuentes de datos existentes en las organizaciones (*data lake*). Estos sistemas de información los podemos clasificar de la siguiente forma:

- ***Transactions processing systems, TPS.*** Son aquellos sistemas encargados de gestionar la ejecución de procesos indivisibles en las organizaciones conocidas como transacciones.
- ***Management information systems, MIS.*** Estos sistemas permiten la toma de decisiones en cuanto a gestión empresarial.
- ***Decision support system, DSS.*** Los DSS son sistemas expertos que permiten a la alta dirección obtener un apoyo en el proceso de toma de decisiones.



- ***Executive information systems, EIS.*** Los *executive information systems* son un tipo específico de decisión *support system*, pero con la finalidad de ayudar en la toma de decisiones a un nivel de ejecutivo sénior. Permite a estos obtener información sobre ventas, gastos y evolución de la empresa y sus departamentos.
- ***Office automatization systems, OAS.*** Son aquellos sistemas englobados tanto en el software como en el hardware que permiten la gestión automatizada de información de la empresa en lo administrativo, en tareas tanto de comunicación como de digitalización, almacenamiento y procesamiento de tareas rutinarias y estandarizadas en las empresas.
- ***Expert systems, SE.*** Es un software o la combinación de software más hardware capaz de emular un comportamiento humano que permite ejecutar tareas de cierto grado de complejidad de modo similar a como lo haría un ser humano.
- ***Enterprise resource planning, ERP.*** Son sistemas capaces de aglutinar en un único sistema de información todos los procesos de la operativa de una empresa en un único componente, recopilando información a todos los niveles en la organización tanto en ámbitos de logística, compras, contabilidad, finanzas, *stock*, pedidos, recursos humanos, producción, proyectos y tantos otros.
- ***Customer relationship management, CRM.*** Los sistemas de gestión de relación con el cliente son un tipo de software con orientación específica hacia el cuidado y el seguimiento de acciones con el cliente; su objeto es gestionar los contactos, las ventas y la oferta/demanda con los clientes.
- ***Supply chain management, SCM.*** Los sistemas de gestión de cadenas de suministro recopilan información y procesos en un sistema de información que integran con datos relativos a operaciones, logística, aprovisionamiento, almacenamiento, distribución, postventa y compras.

Hoy en día, el concepto de sistemas de información viene fuertemente ligado al de tecnologías de la información y las comunicaciones (TIC)<sup>6</sup>. Estas permiten automatizar todo el ciclo por el que circula la información en las empresas, desde su recogida mediante sensores o introducción de datos manual pasando por el modelado de datos y hasta la toma de decisiones. Gracias a esa automatización, las empresas son capaces de dar respuesta casi inmediata a cualquier operación, optimizando sus procesos internos y sus servicios a los clientes. Las empresas se han dado cuenta de que aplicar estas tecnologías les permite posicionarse frente a su competencia como una empresa más eficiente, a la vez que se ahorran costes y se mejoran los beneficios.

Las TIC tienen la ventaja de ofrecer alta disponibilidad de la información gracias al almacenamiento; por otro lado, la continua innovación en estas tecnologías ha facilitado la implementación de metodologías de estadística y *business intelligence*, de mucha importancia para la toma de decisiones y de gran valor para las empresas.

Las herramientas de *business intelligence* (BI) se han implantado y establecido en la mayoría de las empresas. Las grandes corporaciones disponen de un departamento en el que se invierten recursos de BI, conocedores del valor añadido que suponen para hacer frente a los competidores y para conocer el entorno de mercado. Estos departamentos se han convertido en el centro neurálgico de las organizaciones, donde se analizan y modelan los datos, se realizan predicciones o se contrastan teorías, entre otras muchas funciones.

Estas herramientas de BI aportan valor a todos los procesos de las organizaciones, permitiendo el acceso a todos los niveles de infor-

---

6 <https://www.techopedia.com/definition/770/decision-support-system-dss>

mación para analizarla y proveer de respuestas a las diferentes problemáticas de empresas.

De esta forma, los departamentos de BI responden a preguntas sobre el pasado, mediante estadísticas descriptivas con datos históricos. Sobre este pasado permiten responder a por qué se han dado determinados sucesos, gracias a análisis de determinación de causalidad con series temporales. Asimismo, permiten generar escenarios con modelos de multivariantes, para averiguar cuáles serían las consecuencias de determinadas decisiones estratégicas. Estos mismos modelos también pueden utilizarse para predecir resultados en el futuro y anticipar los efectos posteriores a una acción.

Es de gran importancia tener una idea de cuál será la mejor distribución de información. Para que esta sea útil a la hora de tomar decisiones, son necesarias herramientas para distribuir el conocimiento generado, en el momento adecuado y a los usuarios que puedan sacar provecho de él.

Los principales sistemas de presentación de información de los que podemos disponer son los informes estándar y los sistemas de consulta. Los **informes estándar** o *reports* son las formas más tradicionales de representar información. Se basan en una presentación en dos dimensiones donde cada una de ellas puede tener una estructura jerárquica. Presentan resultados de una forma predefinida y, por lo tanto, no se pueden utilizar en un análisis dinámico de estos resultados. Es habitual la automatización de este tipo de presentación. Las herramientas que se suelen emplear son numerosas, desde Qlik o Tableau, pasando por SAS, hasta herramientas más especializadas como Report Builder.

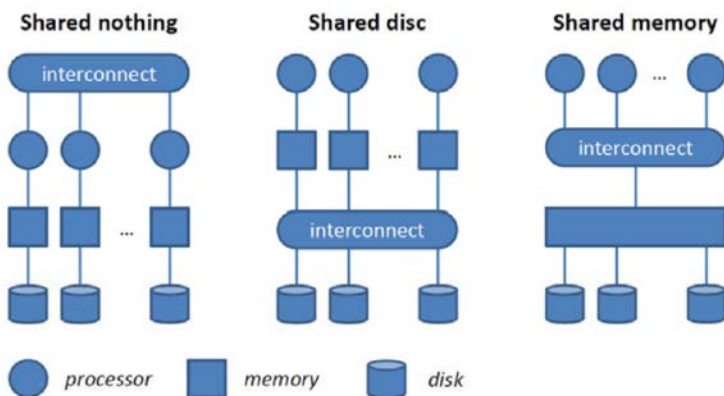
Los **sistemas de consulta** o herramientas de análisis *ad hoc* permiten alcanzar un gran nivel de detalle (*drill down*) o modificar los ejes de análisis. El nivel más básico de estas herramientas corresponde a trabajar con datos extraídos en una *pivot table* de una hoja de cálculo. Las herramientas más sofisticadas permiten gestionar dinámicamente la consulta en el *data warehouse* previamente a su presentación.

Estos sistemas requieren que sus usuarios, además del conocimiento técnico de la herramienta, puedan utilizar ayudas a la consulta que acceden, a las definiciones de la semántica de los datos y, por lo tanto, al contenido del *data warehouse*.

No podemos entender el concepto de *big data* sin entender antes el de «base de datos distribuida». Este tipo de BDD es aquel donde la gestión de datos se distribuye en diversos nodos de una red. Cada nodo es una base de datos por sí misma, con cierta heterogeneidad. En todo momento, los nodos pueden comunicarse mediante la red. Este esquema permite la computación paralelizada, que dota a la arquitectura de una gran potencia de cálculo (en paralelo) con tanta velocidad y con tanta capacidad de procesamiento de datos como queramos o seamos capaces de disponer (**escalabilidad**).

Este tipo de arquitecturas (*parallel distributed databases*) pueden optar por una serie de combinaciones de recursos a varios niveles: de procesador, de memoria o de disco. Y el objetivo de cualquiera de ellas permite dotar a la infraestructura de una mejora sustancial en varios o en todos sus niveles, en función del esquema elegido. En la figura 7 vemos los diferentes esquemas por los que podemos optar.

**Figura 7.** Esquemas de distribución según elementos de infraestructura



Fuente: D. DeWitt, J. Gray (1992). *Parallel Database Systems: The future of High Performance Database Processing*.

El **big data**<sup>7</sup> es un cambio de paradigma en el procesamiento y modelado de datos; es una evolución y confluencia de varias técnicas complementarias y tecnologías que permiten abarcar un **volumen**, **variedad** y **velocidad** de procesamiento de datos poco habituales, lo que habitualmente se conoce como las 3 V del **big data**, aunque algunos autores incluyen adicionales, como la **veracidad** y el **valor**.

Dentro del **big data** se entiende que para ser considerado como tal un proceso y un conjunto de datos se engloban dentro de este paradigma si cumple una serie de condiciones de volumen, velocidad, variedad, veracidad y valor. Las áreas relacionadas con estos conceptos abarcan diferentes disciplinas y funcionalidades.

En el concepto de volumen y velocidad encontramos las áreas de consultas declarativas y la optimización de consultas. En el de variedad y variabilidad están las áreas de calidad de datos, integración de datos, *webmining*, *textmining* y recuperación de datos desestructurados. En el de veracidad, se encuentran la consistencia de datos, el razonamiento estadístico, la incertidumbre, la conexión y la fusión de datos. En el concepto de valor, se encuentran el *data analytics*. Y, dentro de este, el *data mining*, la simulación, la algoritmia y el aprendizaje automático.

En el apartado «Análisis de opiniones para la gestión del conocimiento», veremos un ejemplo de análisis de texto (*text mining*) a partir de datos desestructurados provenientes de Twitter, que nos permitirá comprobar la importancia de este tipo de datos. Este tipo de análisis tiene sentido si se realizan en tiempo real (*real time streamming*): la componente velocidad de las 3 V del **big data**. Por otro lado, es necesario disponer de técnicas para procesar datos desestructurados: la variedad; y finalmente resulta evidente la importancia de disponer de una estructura de base de datos distribuidos si queremos abarcar el procesamiento de tuits de una red de más de quinientos millones de usuarios generando sesenta y cinco millones de tuits diarios: el volumen.

---

7 [www.gartner.com/it-glossary/big-data/](http://www.gartner.com/it-glossary/big-data/)

## 2.2. Estadística<sup>8</sup>

A continuación veremos las herramientas estadísticas que el *data scientist* tiene a su disposición para extraer conocimiento de los datos, interpretar sus resultados y traducirlos a acciones para la toma de decisiones en la organización.

La estadística tiene su origen en el siglo XVIII. La raíz de la palabra «estadística» es Estado. No es casualidad, ya que originalmente la estadística tenía su principal ocupación o fuente de análisis en los datos de los Estados y todos los esfuerzos iban encaminados a resolver grandes asuntos de los Gobiernos, pero posteriormente el perímetro de estudio se fue ampliando, llegando a abarcar dominios tan dispares como la economía, la biología o el *marketing*.

La materia prima de la estadística son los datos. Podríamos definirla como una ciencia cuyo objeto es el análisis, la interpretación, la representación y la organización de datos.

Los datos que se recogen son de varios tipos. El primer gran grupo es el de los **datos cuantitativos**; por ejemplo: la recogida de estaturas de una clase de alumnos. Otro importante grupo es el de los datos cualitativos; por ejemplo: las localidades donde residen los alumnos. Los **datos cualitativos** pueden ser **ordinales**, como los que recogemos con el concepto tamaño (pequeño, mediano, grande) o pueden ser **nominales**, como los colores (azul, verde, amarillo...).

Con los datos, la estadística es capaz de extraer conocimiento, sintetizar la información, representarla o incluso hacer predicciones. Entre las distintas técnicas clásicas destacamos las siguientes:

---

<sup>8</sup> En la bibliografía encontraréis referencias específicas de este apartado.

- **Análisis de la varianza (ANOVA):** se utiliza para examinar diferencias entre varios grupos. Si bien para contrastar diferencias entre dos grupos podríamos aplicar el contraste de hipótesis de la diferencia de medias, en el caso de que queramos ver si hay diferencias entre más de dos grupos podemos considerar el ANOVA. Esta técnica se enmarca dentro de los modelos bivariantes, dado que se trabaja sobre muestras de una variable y sobre diversas poblaciones.
- **Contraste de hipótesis:** se hacen a partir de la construcción de intervalos de confianza con los datos de una determinada muestra. En los contrastes partimos de una hipótesis que deberemos confirmar o rechazar en base a los valores de un paquete estadístico y su función de probabilidad. Una hipótesis será rechazada o aceptada siempre con un grado de significación estadística.
- **Regresión lineal múltiple:** una regresión lineal simple nos permite relacionar dos variables: explicativa y explicada. ¿Qué sucede cuando una variable es explicada por más de una variable independiente? En general, nos encontraremos que la mayoría de los fenómenos que se estudian dependerán de más de una variable. Para hacer estos estudios, usaremos la regresión lineal múltiple. La relación que se establece entre las variables independientes (o explicativas) y la variable dependiente (o explicada) es una relación lineal, que habitualmente expresaremos de la siguiente manera:

$$I = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k$$

Aquí los diferentes  $\beta_i$  representan los diversos parámetros del modelo. Conocer la relación que existe entre las variables es conocer cuál es el valor de estos parámetros. La estimación mínima cuadrática de los parámetros servirá para saber cuáles son estos valores. La significación individual y global de los parámetros, así como el coeficiente de determinación nos ayudarán a sacar conclusiones sobre la relación final que se establezca por medio del modelo.

- **Análisis de conglomerados jerárquico:** es una de las técnicas más usadas en modelizaciones multivariantes para agrupar elementos que tienen propiedades similares. Un elemento central de este análisis es el concepto de distancia, que es empleado para saber cómo es de próximo un elemento respecto de otro mediante la comparación de sus características.

Por ejemplo, si queremos saber si dos países,  $X$  e  $Y$ , son similares en cuanto a su desarrollo en telecomunicaciones, deberemos considerar diferentes indicadores en telecomunicaciones (por ejemplo, líneas de teléfono, número de usuarios de Internet, etc.) y compararlos con una distancia (posiblemente la distancia euclídea es la más utilizada a la hora de hacer este análisis de clasificación):

$$d_e = \sqrt{\sum_i (X_i - Y_i)^2}$$

- **Análisis de conglomerados no jerárquico:** tiene como objetivo la agrupación de las variables analizadas en grupos que cuentan con características similares y, a la vez, permite reducir el número de casuísticas. En el caso del análisis de conglomerados no jerárquico, los grupos se definen previamente, ya sea a partir de los criterios considerados a la hora de definir las distancias, ya sea para que cada una de las variables que consideramos se agrupa con el vecino más cercano. En este caso, las necesidades de datos son menores.
- **Análisis de correspondencias múltiple:** en el análisis de correspondencias simple se ve cómo podemos analizar relaciones entre variables, que pueden ser no métricas. En concreto, se ve la forma de relacionar una lista de atributos con diferentes individuos (empresas, en aquel caso) a partir de la valoración de los usuarios. Ahora, en el análisis de correspondencias múltiple no tendremos una lista única de atributos, sino que habitualmente habrá diferentes variables cualitativas con diferentes categorías entre las que queramos estudiar relaciones de dependencia e independencia. La base de este análisis es disponer de una tabla de contingencia  $Z$  en el que se relacionen individuos con categorías.



Si hacemos una lista de todas las categorías de manera consecutiva, podremos escribir:

$$z_{ij} = \begin{cases} 1, & \text{si el individuo } i \text{ escoge la categoría } j \\ 0, & \text{si el individuo } i \text{ no escoge la categoría } j \end{cases}$$

- **Análisis de series temporales:** se emplea si queremos estudiar la evolución de una variable (por ejemplo, del valor de un determinado fondo de inversión) a lo largo del tiempo. Este valor, que vamos observando a medida que pasan los días, constituye un caso de lo que se denomina serie temporal. El estudio de las series temporales es básico en el ámbito económico y empresarial, dado que se constituyen en el método cuantitativo más utilizado a la hora de hacer previsiones o de crear expectativas futuras.

Una serie temporal está formada por cuatro componentes: la tendencia, el ciclo, el componente estacional y el componente errático. El análisis clásico de series temporales estudia el valor que toma la serie temporal en cada momento del tiempo en función de todos o de algunos de estos cuatro componentes según un esquema aditivo o multiplicativo. Por lo tanto, atendiendo al número de componentes que figuren en una serie temporal, se utilizarán unos métodos de estimación u otros. Para detectar cuáles de estos componentes contienen una serie temporal podemos utilizar los métodos gráficos o recurrir a los contrastes estadísticos; el contraste de Daniel nos ayudará a averiguar si una serie tiene tendencia; el de Kruskal-Wallis nos determinará si la serie tiene componente estacional.

El coeficiente de correlación entre dos variables no contemporáneas ( $X_t$  y  $Y_{t-k}$ ) se denomina correlación serial; mientras que la correlación de una variable consigo misma diferida  $k$  periodos ( $Y_t$  y  $Y_{t-k}$ ) recibe el nombre de autocorrelación. Calcular estas correlaciones mediante la fórmula de Pearson es como si calculáramos un coeficiente de correlación.

Lo que se busca encontrar con la correlación serial es, por ejemplo, qué relación existe entre el valor de una variable en un determinado periodo con el valor de otra variable en el periodo

inmediatamente siguiente. En cambio, con la autocorrelación se pretende encontrar la relación entre el valor de una variable en un periodo concreto y el valor de esta variable en el periodo anterior o posterior.

- **Análisis factorial:** aquí se engloban todas aquellas técnicas que tienen como objetivo reducir el número de variables con las que trabajamos y, por tanto, la dimensión de la matriz de datos inicial. A partir de una base de datos se obtiene un número reducido de factores o nuevas variables que sintetizan la información de partida y que permiten estudiar las relaciones existentes entre las variables iniciales. En estas técnicas es de gran ayuda la representación gráfica simplificada de las filas y columnas, ya que muchas veces resulta clave para la interpretación de los resultados.

Aunque todas las técnicas de análisis factorial tienen una estructura común, según sea la naturaleza de los datos, se originan diferentes métodos: El análisis factorial de componentes principales analiza tablas de variables cuantitativas o métricas (individuos x variables métricas). El análisis factorial de correspondencias simples se centra en tablas de contingencia y, en general, cualquier tabla de números positivos, siempre que la suma de una fila y una columna tengan sentido y se puedan interpretar. El análisis factorial de correspondencias múltiples estudia tablas de variables cualitativas (individuos x variables cualitativas).

## 2.3. Aprendizaje automático

El **aprendizaje automático** o *machine learning* (ML) es el campo de la informática que estudia métodos automáticos para hacer predicciones basados en experiencias pasadas de un sistema.

Una de las finalidades de este campo es producir «buenos» modelos o introducir mejoras en los modelos tradicionales.

Los modelos son descripciones compactas de una muestra de datos que permite predecir escenarios y generar simulaciones. El ML se ocupa de diferentes áreas, pero las más habituales son los métodos de clasificación y los modelos de predicción. El ML es una de las ramas de la inteligencia artificial; por tanto, una de sus finalidades es dotar a las computadoras de la capacidad de aprender.

Es un campo que bebe de diversas fuentes. La más destacable es, sin duda, la estadística con la que comparte algunas áreas de estudio. En ocasiones encontramos metodologías estadísticas englobadas dentro del abanico de técnicas que se enseñan habitualmente en los cursos de *machine learning* y *analytics*. Entre los métodos más conocidos encontramos los siguientes:

- **Redes neuronales:** una red neural está formada por «neuronas» en forma de nodos de un grafo y unas conexiones o capas que las comunican que reciben una ponderación. Las redes formadas representan aquello que queremos analizar, que quedan reflejadas de forma analítica en forma de grafos. Se utilizan símbolos para representarlas y resulta difícil extraer partes del conjunto, ya que su forma queda sujeta a una visión global del planteamiento. Con las *Neural networks* se realiza la inferencia robusta del problema que se analiza y mediante un proceso de aprendizaje y adaptación son capaces de hacer una representación del sujeto de análisis y de hacer predicción de comportamiento y de estructuras o clases mediante modificaciones adaptativas en sus pesos.
- **Support vector machines (SVM):** se basan en el principio de minimización del riesgo estructural de la teoría del aprendizaje computacional. En los problemas de clasificación tratan de encontrar la solución al problema de minimización del riesgo estructural, buscando una forma o hiperplano que mejor separa las clases. Los SVM son muy multidisciplinarios. Podemos utilizar funciones de umbral lineales para encontrar la solución al problema de clasificación, pero sin duda su potencial radica en mejorar esta linealidad mediante otras funciones más adecuadas (kernel) que pueden ser polinomiales, radiales (RBF), sinusoidales y de muchas otras formas. Además de utilizarse en problemas de

clasificación también se utilizan para la predicción y para problemas de segmentación.

- **Redes bayesianas:** un tipo de sistemas intencionales son las llamadas redes causales probabilísticas. Una red causal probabilística, o red bayesiana, se define sobre un grafo dirigido acíclico (GDA) donde los nodos representan variables, y los arcos del grafo describen relaciones de dependencia, tipo causa-efecto, entre las variables. Si un nodo con una variable A es padre de un nodo (predecesor inmediato) con una variable B, entonces se considera que A es una causa directa de B, o bien que B es un efecto directo de A. Estas relaciones de dependencia se cuantifican en cada nodo con la distribución de probabilidades condicionada de la variable asociada a ese nodo respecto a las variables en los nodos padre. Por lo tanto, en una red causal tenemos a la vez una componente cualitativa y una componente numérica de representación del conocimiento: la cualitativa describe las relaciones de dependencia/independencia entre las variables que intervienen en la descripción del problema, y la cuantitativa o numérica cuantifica estas relaciones mediante probabilidades (o, más en general, podría ser mediante medidas de incertidumbre).
- **Árboles de regresión y clasificación (C&RT):** en términos generales, el propósito de los análisis a través de algoritmos de construcción de árboles es determinar un conjunto de condiciones lógicas (divisorias) que permite una predicción o clasificación precisa de los casos. Los problemas de tipo de regresión son generalmente aquellos en los que se intenta predecir los valores de una variable continua a partir de una o más variables predictoras continuas y/o categóricas. Los problemas de tipo de clasificación son generalmente aquellos en los que intentamos predecir los valores de una variable dependiente categórica (clase, pertenencia a un grupo, etc.) a partir de una o más variables predictoras continuas y/o categóricas.
- **Algoritmos genéticos:** son algoritmos de búsqueda estocásticos inspirados en los fenómenos naturales de herencia genética y

en la idea de que el mejor es el que sobrevive (supervivencia de las especies). En una población de individuos, las nuevas generaciones están más adaptadas al medio que las precedentes y, en promedio, las nuevas poblaciones serán más rápidas, con un grado de mimetismo más grande, etc., que las anteriores, porque eso es lo que les permite sobrevivir. La analogía con el proceso biológico de la evolución dirige todos los pasos de los algoritmos genéticos. Así, consideraremos poblaciones de individuos representados por sus cromosomas (cromosomas formados por genes). Y, dado un cromosoma, consideraremos su cruce con otro cromosoma o su mutación. Los cruces y las mutaciones llevarán a nuevas generaciones de poblaciones. Como todos los métodos de búsqueda, los algoritmos genéticos permiten encontrar una solución a un problema dado. Sin embargo, debido a que son algoritmos estocásticos, normalmente no encuentran la mejor solución del problema, sino que dan con una que aproxima la solución óptima.





**Las soluciones**





# 1. Sistemas de recomendación de productos

A continuación veremos como un *data scientist* puede enfrentarse al diseño e implementación de un sistema de recomendación de productos con el objeto de maximizar las ventas en el canal en línea de su empresa.

Los sistemas de recomendación de productos tienen una finalidad intrínseca más allá de mejorar para los usuarios la experiencia de compra, y esta no es otra que incrementar las ventas para el distribuidor. Mediante algoritmos que analizan los patrones de compra, se extraen comportamientos y reglas que se repiten a lo largo del histórico de datos y que sugieren relaciones entre productos y/o compradores que van más allá de la casualidad.

Tenemos tres grandes grupos de sistemas de recomendación:

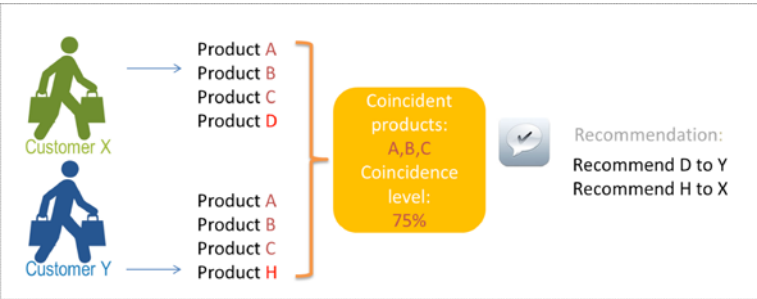
- 1) **Basados en los compradores que adquieren productos similares:** buscan compradores similares en sus actos de compra. Un sistema de este tipo analiza las compras de los consumidores, busca pruebas de similitud entre parejos y deduce cuáles son los productos en los que no ha habido coincidencia. Finalmente recomienda estos productos no coincidentes a los compradores parecidos basándose en la similitud en sus preferencias y bajo la hipótesis de que el producto no coincide debido a que el comprador desconoce su existencia. Pero podría ser de su agrado, dado que otros compradores similares han decidido comprarlo.
- 2) **Basados en los productos que compran compradores similares:** estos analizan los patrones desde la perspectiva del producto, buscan conjuntos de productos muy similares, según los compra-

dores que los compran. Cuando un producto ha sido comprado por tipos de compradores muy similares se puede recomendar otro producto complementario que también compran mucho los consumidores que adquieren el primero.

- 3) **Basados en productos que se compran con otros productos habitualmente:** este tipo de sistema recorre el histórico de compras buscando patrones que se repiten de forma frecuente. Busca productos que habitualmente han sido comprados conjuntamente en un gran número de cestas o de ocasiones de compra. Deduce qué producto se compra muy habitualmente ante la compra de otro u otros productos.

Entre los tipos de algoritmos basados en similitud de compradores y similitud de productos encontraríamos aquellos basados en distancias, como el *k-nearest neighbor*, que analiza los hábitos de compra de una base de datos de clientes, busca las semejanzas entre todos los clientes en función de su histórico de compras, y localiza aquellos clientes que más se asemejan, de tal manera que es capaz de recomendar productos que el cliente objetivo aún no ha comprado, pero que sí ha comprado un cliente que ha realizado un tipo de compras muy parecidas a las suyas.

**Figura 8.** ejemplo de recomendación por compradores similares (*alikes*)



Fuente: elaboración propia.

Por ejemplo, imaginemos un caso como el de la figura 8, en el que tenemos un cliente X de Amazon que ha comprado los productos A, B, C y D; por otro lado, encontramos otro cliente Y que ha comprado A, B, C y

H. El algoritmo encontraría similitudes entre ambos, en tanto que los dos han comprado los productos A, B y C. A su vez, sería capaz de recomendar a X la compra de H, puesto que Y es un tipo de cliente muy parecido que además de tener gustos muy parecidos a X también ha comprado el producto H, el cual X todavía no ha probado.

Otro tipo de algoritmos con finalidades muy parecidas serían los englobados dentro del campo de reglas de asociación, como el *apriori*. Este algoritmo busca entre la base de datos de compras, patrones y reglas de asociación. Así, es capaz de detectar qué productos tiene más propensión a comprar un determinado cliente si también ha comprado otro producto o una combinación de productos en una cesta determinada.

Figura 9. ejemplo de recomendación por patrones de compra



Fuente: elaboración propia.

La forma en que el algoritmo genera recomendaciones es el siguiente: imaginemos un caso como el de la figura 9, en el que tenemos una base de datos donde se repite la cesta con los productos A, B, C un número elevado de veces; por ejemplo, noventa de cada cien veces: en tal caso el algoritmo encontrará que noventa de cada cien veces que un cliente ha comprado A y B también ha comprado C. De igual modo, indicará que noventa de cada cien veces que un cliente ha comprado A y C también ha comprado B. Y, finalmente, también detectará que noventa de cada cien veces que un cliente ha comprado B y C también ha comprado A. Por tanto, si el algoritmo se propone recomendar un producto a un comprador que ha realizado una compra de A y C, podría recomendar para la siguiente compra, o en el mismo acto de compra, el producto B.

En este caso práctico veremos como una empresa como Amazon podría beneficiarse de una base de datos con histórico de productos perecederos de un panel de consumo, de una fuente proveniente de una empresa privada de estudios de mercado. En el ejemplo usaremos datos simulados de compra en línea; el objetivo es que el estudiante sepa comprender la utilidad del método e interpretar los resultados con el fin de obtener conocimiento de este algoritmo del campo de la minería de datos.

Con el material de los puntos 1, 2 y 3, el estudiante ha visto conceptos relevantes sobre el sector *retail* del mercado español, también ha entendido la importancia de la elección del producto y del canal, para finalmente ver cómo exprimir al máximo los métodos que ofrecen los campos de *business Intelligence*, la minería de datos y la estadística.

El caso práctico pone al estudiante en la piel de un data scientist, que tiene como objetivo generar un sistema de recomendación en tiempo real de productos en función de las compras realizadas en un *retailer* en línea que ofrece productos de alimentación, similar a Amazon Prime Now. Sin duda, una empresa como Amazon parte de una situación de líder de mercado en el *e-commerce* y de un potencial en operaciones y logística dentro de este ámbito que difícilmente las otras empresas puedan alcanzar. El punto en el que Amazon tiene menos capacidad es en el de conocimiento de mercado del sector gran consumo. Si una empresa de venta solo a través de Internet quiere hacerse un hueco y entrar a competir en este sector tan maduro, necesitará recurrir a fuentes de datos de terceros y al *know how* en metodologías aplicadas al *retail* Español. Por otro lado, tendrá que adaptar sus procesos y metodologías a las de la nueva categoría, utilizando herramientas de minería de datos como las que se verán en esta solución.

Anteriormente en este apartado, hemos definido dos métodos que pueden ser útiles para crear un sistema de recomendación de productos. Nos centraremos en el segundo, el *apriori*, que pertenece al campo de algoritmos de **reglas de asociación** (*Association Rules Field*). El problema que analizan se conoce como **análisis**

**de cestas de compra (ACC).** El ACC asume que el mercado se compone de un conjunto elevado de elementos (ítems), como el pan, el agua, yogur, etc., que pueden combinarse en las cestas de los compradores. El objetivo que se persigue es conocer qué elementos aparecen en la misma cesta un mayor número de veces y a la vez usar esta información para obtener algún tipo de beneficio, habitualmente económico. Las conclusiones que se extraen de estos patrones pueden servir para conocer cómo combinar promociones aprovechando las sinergias entre productos, o cómo ubicar en los lineales virtuales de compra estos productos que habitualmente se compran juntos.

En el contexto de gran consumo, tenemos cestas y tenemos productos, aunque podríamos utilizar esta estrategia en otros contextos muy diferentes con finalidades similares. Por eso definimos el concepto de **transacciones**, que son todo el conjunto de elementos que suceden en un momento determinado; en nuestro caso productos de cada cesta, una a una. Y por **elementos** entendemos sucesos que se combinan para formar una transacción; en nuestro caso productos que forman cada cesta, uno a uno.

De todas las posibles combinaciones de elementos, estamos interesados en particular en las **combinaciones de elementos más frecuentes (CEF)**; es decir, aquellas combinaciones de elementos que se repiten más a menudo en las cestas.

En un segundo estadio analizaremos las reglas de asociación (AR); es decir, una vez conozcamos las combinaciones más frecuentes de elementos podremos generar recomendaciones en función de cómo se asocian los diferentes elementos.

Entenderemos por **regla** a un conjunto de elementos que generan un patrón o recomendación.

Por ejemplo, en la siguiente figura se observa la siguiente regla:

Figura 10. Ejemplo con diez cestas

**Historical Database**

Basket 0 U K C T A B  
Basket 1 A U B N C R  
Basket 2 B E A U C R  
Basket 3 C B A S R T  
Basket 4 C A V B S R { A , C } => B  
Basket 5 B N C A R T  
Basket 6 N A C B F R  
Basket 7 A H B C S R  
Basket 8 B N C R A N  
Basket 9 N C A E Y U

Fuente: elaboración propia.

Para entender todo el proceso, también necesitamos definir previamente el concepto de soporte, frecuencia, confianza y elevación. **Soporte** es el número de veces que se repite una combinación de productos. El soporte mínimo será un número que deseamos considerar como mínimo. Es decir, un elemento será frecuente siempre que su soporte sea superior al valor de soporte mínimo. Dicho soporte viene expresado como el coeficiente entre el número de transacciones en que aparece cierta combinación de elementos respecto el total de transacciones posibles; es decir, como un porcentaje respecto al número total de transacciones.

En nuestro ejemplo de la figura 10 tenemos que el soporte de {A, B, C} es 9/10, ya que encontramos esta combinación nueve veces y tenemos en total diez cestas.

La **frecuencia** es el número de veces que un conjunto de elementos expresados como una regla se repite a lo largo de todas las cestas posibles; o lo que es lo mismo, el número de cestas en las que aparece dicha combinación de elementos en todo el conjunto de transacciones.

Por ejemplo, el soporte de {A, C} es 1,0, ya que está presente en todas las cestas. La regla {A, C} => B tiene una frecuencia que corresponde al soporte de  $\{A, C\} \cup B$ ; es decir, 0,9 porque los dos conjuntos (los tres elementos) aparecen combinados nueve veces de diez.

Finalmente, la **confianza** es el cociente entre el soporte de la regla y el soporte de los elementos de la izquierda de la regla.

En nuestro caso, el soporte de los elementos de la izquierda de la regla  $\{A, C\}$  es 1,0, y la frecuencia de la regla  $\{A, C\} \Rightarrow B$  es de 0,9 y, por tanto, la confianza es de 0,9.

Nótese que esta confianza coincide con la de la regla  $\{A, B\} \Rightarrow C$  formada por los mismos elementos, ya que el soporte de la parte izquierda  $\{A, B\}$  es de 0,9, y la frecuencia de la regla es de 0,9; por tanto la confianza es de 1,0 (de 0,9/0,9).

Finalmente, la **elevación** es un concepto que intenta extraer reglas relevantes y que se realice una búsqueda inteligente de las reglas relevantes. Parte del supuesto de que, si el elemento de la derecha de la regla es muy frecuente entonces, es muy probable que cualquier regla aparezca con una confianza y soporte elevados solo por el hecho de incluir este elemento. Con este nuevo valor eliminamos este sesgo. Para calcularlo haremos el cociente entre la confianza de la regla y el soporte del elemento de la derecha.

En el ejemplo  $\{A, C\} \Rightarrow B$  tenemos una confianza de 0,9 y un soporte para B de 0,9; por tanto, una elevación de 1.

Para trabajar con esta resolución, el alumno deberá instalarse el software gratuito R.<sup>1</sup>

A continuación, se suministra, mediante el siguiente enlace, una base de datos de compra en línea con datos simulados: Link de la base de datos de cestas simuladas.<sup>2</sup>

---

1 Es recomendable leer previamente los primeros tres capítulos de Castillo, A. J. S. «Métodos Estadísticos con R y R Commander». [Recurso electrónico gratuito] <<http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>>.

2 Para el análisis descrito a continuación, el consultor proporcionará enlace a un fichero a través del aula.

Con esta base de datos, aplicaremos la técnica *apriori* y se nos pedirá jugar con la tabla de datos intentando encontrar un conjunto de diez reglas de asociación relevantes. Pero previamente deberemos decidir si conservar o eliminar los valores NA (valores faltantes) y otros procesos relacionados con asegurar la calidad de datos. Esta fase previa es la que se conoce como *pre-processing*.

Partimos de un *dataframe*, al que hemos llamado *df* con la estructura que muestra la figura 11. Tenemos tantas columnas como el número máximo de productos encontrados en una cesta; y tantas filas como número de cestas realizadas por los consumidores.

Como se puede observar, el número máximo de elementos que hemos encontrado en una cesta es de treinta y siete. Es decir, la cesta más grande encontrada tiene treinta y siete productos distintos. Por otro lado, tenemos un total de 6.472 cestas en la base de datos simulada.

El hecho de tener cestas más pequeñas hace que aparezcan valores NA en algunas celdas. Debemos plantearnos si tiene sentido en esta etapa del proceso. La siguiente etapa consistirá en pasar cada fila de nuestro *dataframe* a una lista. Entender qué hacemos en esta etapa nos permitirá ver cómo plantear este problema.

Para pasar nuestra tabla a una lista contamos con varias opciones; como nuestro conjunto de datos es pequeño, podemos hacerlo con un simple bucle. De la siguiente forma, ejecutaremos estas líneas de *script*.

```
a_list<-list()

a_list<-list()
for (i in 1:dim(df)[1]){
  if( length(df[i,which(df[i,]!="")]) >= 2 )
    a_list<-append(a_list, list( as.character(df[i,which(d
f[i,]!="")]) ))
}
```



Figura 11. Tabla cruzada de cesta por producto

BE	Element11	Element12	Element13	Element14	Element15	Element16	Element17
Basket1	ACEITUNAS	CERVEZA CON ALCOHOL	FRUTA	JUDIAS VERDES	LECHE CONDENSADA	POLLO ASADO	---
Basket2	ACEITE	ACEITUNAS	PAN MOLDE	PATATAS FRITAS	PAVO	VERDURA	---
Basket3	ALCACHOFAS	CERVEZA CON ALCOHOL	FLAN	FRUTA	HUEVOS	PAN TOSTADO	---
Basket4	ACEITUNAS	CERVEZA CON ALCOHOL	FRUTA	PAN	PATATAS FRITAS	PLATO PREPARADO	---
Basket5	COLORACION	FRUTA	MEMBRILLO	PAN	VERDURA	YOGUR	---
Basket6	LECHE CONDENSADA	LEGUMBRES	VERDURA				---
Basket7	CERVEZA CON ALCOHOL	CONSERVAS DE ATUN	CONSERVAS MEJILLONES	CONSERVAS SARDINAS	ESPECIAS	FRUTA	---
Basket8	CERVEZA CON ALCOHOL	FRUTA	LECHE CONDENSADA	POLLO ASADO	VERDURA		---
Basket9	ACEITUNAS	CERVEZA SIN ALCOHOL	FRUTA	PAN MOLDE	SOPAS	VERDURA	---
Basket10	ACEITUNAS	CACAO SOLUBLE	CAFE	CERVEZA CON ALCOHOL	CONSERVA DE PESCADO	CONSERVAS MEJILLONES	---
Basket11	CALDO	FLAN	FRUTA	PAN	POLLO CONGELADO	VERDURA	---
Basket12	ACEITUNAS	CERVEZA SIN ALCOHOL	HUEVOS	LEJIA	QUESO	SOPAS	---
---	---	---	---	---	---	---	---
---	---	---	---	---	---	---	---
Basket 6472	CAFE	CERVEZA CON ALCOHOL	CHORIZO	FRUTA	FUET	LECHE	---

Fuente: elaboración propia.

El *pre-processing* que hemos necesitado se concentra en la siguiente condición, que filtra cestas de un solo elemento.

```
length(df[i, which(df[i, ] != "")]) >= 2
```

Esto tiene sentido y se basa en el hecho de que queremos obtener recomendaciones de productos en función de la compra de otros productos, y una cesta con un único elemento no nos proporciona la información que necesitamos.

A continuación, a nuestros elementos dentro de nuestras listas de cestas deberemos ponerles un nombre y número para identificarlos. Utilizaremos la siguiente sentencia:

```
names(a_list) <- paste("Basket", c(1:dim(df)[2]), sep = "")
```

Vamos a ver qué aspecto tiene nuestra lista para entender bien que acción hemos llevado a cabo; nótese que para cada cesta tenemos el sufijo «basket» seguido de un valor o secuencia autonumérica (Basket1, Basket2, Basket3...). Usaremos la función **head()** y para ver el resultado:

```
head(a_list)
$Basket1
[1] "ACEITUNAS"          "CERVEZA CON ALCOHOL" "FRUTA"
[4] "JUDIAS VERDES"      "LECHE CONDENSADA"    "POLLO ASADO"
[7] "SOPAS"              "VERDURA"

$Basket2
[1] "ACEITE"          "ACEITUNAS"          "PAN MOLDE"          "PA-
TATAS FRITAS"
[5] "PAVO"            "VERDURA"

$Basket3
[1] "ALCACHOFAS"        "CERVEZA CON ALCOHOL" "FLAN"
[4] "FRUTA"              "HUEVOS"              "PAN TOSTADO"
[7] "QUESO"              "VERDURA"            "ZUMO"
```

```
$Basket4
[1] "ACEITUNAS" "CERVEZA CON ALCOHOL" "FRUTA"
[4] "PAN" "PATATAS FRITAS" "PLATO PRE-
PARADO"
[7] "SALCHICHAS" "VERDURA" "VINAGRE"
[10] "VINO"
```

```
$Basket5
[1] "COLORACION" "FRUTA" "MEMBRILLO" "PAN" "VER-
DURA"
[6] "YOGUR" "ZUMO"
```

```
$Basket6
[1] "LECHE CONDENSADA" "LEGUMBRES" "VERDURA"
```

De esta forma, vemos la composición de productos para las primeras tres cestas.

El paso más importante para usar el algoritmo *apriori* pasa por instalar y cargar la librería «arules»; para ello utilizaremos las funciones **install.packages** y **library** como se muestra a continuación.

```
install.packages("arules")
library(arules)
```

Lo siguiente es pasar nuestra lista al formato de transacciones necesario para que la función entienda los datos que le estamos proporcionando.

```
trans <- as(a_list, "transactions")
```

Podemos ver si todo ha funcionado extrayendo un resumen del resultado con la función **summary()**.

```
summary(trans)
transactions as itemMatrix in sparse format with
5802 rows (elements/itemsets/transactions) and
197 columns (items) and a density of 0.03582346
```

most frequent items:

VERDURA	FRUTA	LECHE	PAN	QUESO	(Other)
2273	2077	1491	1398	1357	32350

element (itemset/transaction) length distribution:

sizes

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
17	18	19	20											
666	830	740	666	554	421	353	276	228	170	150	123	114	103	63
53	43	44	32											
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
36	37													
30	29	20	11	19	8	9	7	4	7	5	7	5	3	5
3	1													

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	3.000	5.000	7.057	9.000	37.000

includes extended item information - examples:

labels

1	ACEITE
2	ACEITUNAS
3	ACONDICIONADOR

includes extended transaction information - examples:

transactionID

1	Basket1
2	Basket2
3	Basket3

El *output* nos informa primero acerca de las dimensiones de nuestro conjunto de transacciones. Vemos que tenemos 5.802 cestas, que hemos reducido respecto a las 6.472 transacciones originales tras eliminar aquellas cestas de un único producto (compra esporádica). A continuación vemos que los ítems que aparecen en mayor número de cestas son: **verdura**, **fruta**, **leche**, **queso**, **pan** y el número de cestas en las que aparece cada producto. El resultado también nos muestra la distribución de cestas por número de elementos; así, por ejemplo,

podemos ver que tenemos 666 cestas con dos productos; 830 con tres productos, y en el extremo superior vemos una única cesta con treinta y siete elementos (cesta de carga o stock). También vemos algunos estadísticos descriptivos para esta distribución, con el mínimo, la mediana, el promedio, rango intercuartílico y máximo. En las dos últimas salidas de resultados, la función nos muestra un ejemplo de elementos (productos) y un ejemplo de transacciones (cestas).

El siguiente paso es lanzar el algoritmo. Ya hemos definido el concepto de soporte y de soporte mínimo; recordad que este es el número mínimo de soporte que estamos admitiendo para una regla; es decir, no contemplaremos ninguna regla que tenga un soporte por debajo de este valor. También es necesario usar el valor mínimo de confianza que nos permitirá filtrar las reglas más relevantes. Utilizaremos para ello la función `apriori()`.

```
rules<-apriori(trans,parameter=list(supp=.01,      conf=.1,
target="rules"))
```

Para inspeccionar las reglas tenemos precisamente la función `Inspect`; nos mostrará un resumen de las principales reglas. El número de reglas para visualizar las especificaremos con el parámetro «n».

```
inspect(head(sort(rules,by="lift"),n=5))
```

Podemos extraer ya algunas conclusiones: por ejemplo, de los conjuntos de dos elementos de nuestros datos simulados podemos ver que los consumidores que compran **flan** también suelen incluir en la misma cesta **natillas**. Para tres elementos vemos que los consumidores que incluyen **aguas** y **verduras** también suelen incluir **jamón curado**. Los coeficientes de elevación de estas reglas son muy elevados; por tanto, podemos decir que estas reglas, aunque cuentan con soporte y confianza bajos, tienen mucha relevancia.

A continuación puede interesarnos conocer cómo son las cestas que incluyen **colas**. Para este fin, la función nos deja escoger filtrar el conjunto de reglas por aquellas que tienen el elemento **colas** en el lado derecho con el parámetro «rhs». Veamos un ejemplo.

```

> rulesRCola <- subset(rules, subset = rhs %in% "COLA" &
lift > 1.5)
> inspect(rulesRCola)

```

	lhs		rhs	support	confidence
1	{BEBIDA NARANJA CON GAS}	=>	{COLA}	0.01206481	0.3465347
	3.367829				
2	{CERVEZA SIN ALCOHOL}	=>	{COLA}	0.01016891	0.2543103
	2.471539				
3	{BEBIDA SIN GAS}	=>	{COLA}	0.01172010	0.2605364
	2.532047				
4	{SALCHICHAS}	=>	{COLA}	0.01223716	0.2500000
	2.429648				
5	{CERDO}	=>	{COLA}	0.01034126	0.1617251
	1.571740				
6	{EMBUTIDO}	=>	{COLA}	0.01465012	0.2492669
	2.422523				
7	{LIMPIADOR HOGAR}	=>	{COLA}	0.01361599	0.2393939
	2.326572				
8	{TOMATE FRITO}	=>	{COLA}	0.01051362	0.1626667
	1.580891				
9	{POLLO CONGELADO}	=>	{COLA}	0.01275422	0.1859296
	1.806975				
10	{CERVEZA CON ALCOHOL}	=>	{COLA}	0.01895898	0.2135922
	2.075816				
11	{PASTAS BOLLERIA}	=>	{COLA}	0.01430541	0.1736402
	1.687538				
12	{FRUTOS SECOS}	=>	{COLA}	0.01620131	0.1649123
	1.602715				
13	{AGUAS}	=>	{COLA}	0.02016546	0.1700581
	1.652726				
14	{CAFE}	=>	{COLA}	0.01723544	0.1639344
	1.593212				
15	{PATATAS FRITAS}	=>	{COLA}	0.02223371	0.1804196
	1.753424				
16	{AGUAS,				
	VERDURA}	=>	{COLA}	0.01068597	0.2357414
	2.291075				

```

17 {LECHE,
    PATATAS FRITAS}          => {COLA} 0.01051362 0.2618026
2.544353
18 {FRUTA,
    PATATAS FRITAS}          => {COLA} 0.01034126 0.2222222
2.159687
19 {PATATAS FRITAS,
    VERDURA}                 => {COLA} 0.01103068 0.2064516
2.006419
20 {FRUTA,
    PASTEL BOLLERIA}          => {COLA} 0.01051362 0.1799410
1.748773
21 {PASTEL BOLLERIA,
    VERDURA}                 => {COLA} 0.01275422 0.2078652
2.020157
22 {DULCES,
    VERDURA}                 => {COLA} 0.01034126 0.1666667
1.619765
23 {LECHE,
    QUESO}                    => {COLA} 0.01085832 0.1567164
1.523063
24 {LECHE,
    VERDURA}                 => {COLA} 0.01723544 0.1618123
1.572588

```

Se puede observar que en las cestas en que el consumidor compra **bebida de naranja con gas** suele aparecer también el producto **cola**. A partir de la regla dieciséis vemos reglas con tres elementos, y podemos destacar que en las cestas en que aparecen los productos **aguas** y **verdura**, o en las cestas que aparecen **leche** y **patatas fritas**, también aparece **cola**.

Para ver las reglas en las que aparece cola en la parte izquierda, lo hacemos de forma análoga, pero utilizando el parámetro «lhs». Veamos otro ejemplo.

```

> rulesInLCola <- subset(rules, subset = lhs %in% "COLA" &
lift > 1.5)

```

```

> inspect(rulesInLCola)

```

	lhs	rhs	support
confidence	lift		
1	{COLA}	=> {BEBIDA NARANJA CON GAS}	0.01206481
0.1172529	3.367829		
2	{COLA}	=> {BEBIDA SIN GAS}	0.01172010
0.1139028	2.532047		
3	{COLA}	=> {SALCHICHAS}	0.01223716
0.1189280	2.429648		
4	{COLA}	=> {CERDO}	0.01034126
0.1005025	1.571740		
5	{COLA}	=> {EMBUTIDO}	0.01465012
0.1423786	2.422523		
6	{COLA}	=> {LIMPIADOR HOGAR}	0.01361599
0.1323283	2.326572		
7	{COLA}	=> {TOMATE FRITO}	0.01051362
0.1021776	1.580891		
8	{COLA}	=> {POLLO CONGELADO}	0.01275422
0.1239531	1.806975		
9	{COLA}	=> {CERVEZA CON ALCOHOL}	0.01895898
0.1842546	2.075816		
10	{COLA}	=> {PASTAS BOLLERIA}	0.01430541
0.1390285	1.687538		
11	{COLA}	=> {FRUTOS SECOS}	0.01620131
0.1574539	1.602715		
12	{COLA}	=> {AGUAS}	0.02016546
0.1959799	1.652726		
13	{COLA}	=> {CAFE}	0.01723544
0.1675042	1.593212		
14	{COLA}	=> {PATATAS FRITAS}	0.02223371
0.2160804	1.753424		
15	{COLA, VERDURA}	=> {AGUAS}	0.01068597
0.2540984	2.142847		
16	{COLA, PATATAS FRITAS}	=> {LECHE}	0.01051362
0.4728682	1.840095		
17	{COLA,		



0.2946860 2.391284	18 {COLA,	LECHE}	=> {PATATAS FRITAS}	0.01051362
0.2857143 2.318482	19 {COLA,	FRUTA}	=> {PATATAS FRITAS}	0.01034126
0.2622951 2.128442	20 {COLA,	VERDURA}	=> {PATATAS FRITAS}	0.01103068
0.2904762 1.961982	21 {COLA,	FRUTA}	=> {PASTEL BOLLERIA}	0.01051362
0.3032787 2.048455	22 {COLA,	VERDURA}	=> {PASTEL BOLLERIA}	0.01275422
0.2459016 1.668680	23 {COLA,	VERDURA}	=> {DULCES}	0.01034126
0.5916667 1.510273	24 {COLA,	PAN}	=> {VERDURA}	0.01223716
0.4225352 1.806595	25 {COLA,	YOGUR}	=> {QUESO}	0.01034126
0.3680982 1.778273	26 {COLA,	QUESO}	=> {YOGUR}	0.01034126
0.3865031 1.504018	27 {COLA,	QUESO}	=> {LECHE}	0.01085832
0.3904762 1.519479	28 {COLA,	FRUTA}	=> {LECHE}	0.01413306
0.4098361 1.594815		VERDURA}	=> {LECHE}	0.01723544

Vemos que los resultados son muy parecidos. Existe una simetría prácticamente exacta entre las reglas según la posición en la regla

del producto elegido. Sin embargo, nótese que no es del todo así, en el anterior resultado veíamos la regla {CERVEZA SIN ALCOHOL} => COLA, pero si nos fijamos en esta salida de resultados no encontramos la regla simétrica {COLA} => CERVEZA SIN ALCOHOL. Esto indica que, casi siempre que se compra **cerveza sin alcohol**, el cliente también se lleva **cola**, pero no es habitual que cuando un cliente compre **cola** también incluya en la cesta **cerveza sin alcohol**.

Veamos otro ejemplo. Ahora filtraremos por el producto **dulces**, que incluye el concepto de galletas dulces. En esta ocasión, utilizaremos la función **Label()**, que da como resultado un *output* más resumido. En este hemos incluido las reglas con el producto a ambos lados de la regla, para mayor comodidad.

```
> rulesInNATA <- subset(rules, subset = lhs %in% "DULCES" &
lift > 2)
> labels(rulesInNATA)
[1] "{DULCES} => {AZUCAR}"
[2] "{DULCES} => {TABLETA CHOCOLATE}"
[3] "{AZUCAR,DULCES} => {LECHE}"
[4] "{DULCES,LECHE} => {AZUCAR}"
[5] "{DULCES,VERDURA} => {JAMON CURADO}"
[6] "{DULCES,LECHE} => {TABLETA CHOCOLATE}"
[7] "{DULCES,FRUTA} => {TABLETA CHOCOLATE}"
[8] "{DULCES,VERDURA} => {TABLETA CHOCOLATE}"
[9] "{DULCES,JAMON YORK} => {QUESO}"
[10] "{DULCES,QUESO} => {JAMON YORK}"
[11] "{DULCES,VERDURA} => {JAMON YORK}"
[12] "{DULCES,PASTAS BOLLERIA} => {LECHE}"
[13] "{DULCES,LECHE} => {PASTAS BOLLERIA}"
[14] "{DULCES,LECHE} => {CAFE}"
[15] "{DULCES,QUESO} => {CAFE}"
[16] "{DULCES,FRUTA} => {CAFE}"
[17] "{DULCES,VERDURA} => {CAFE}"
[18] "{DULCES,LECHE} => {CONSERVAS DE ATUN}"
[19] "{DULCES,YOGUR} => {PASTEL BOLLERIA}"
[20] "{DULCES,VERDURA,YOGUR} => {QUESO}"
[21] "{DULCES,QUESO,VERDURA} => {YOGUR}"
```

```

> rulesInNATA <- subset(rules, subset = rhs %in% "DULCES" &
lift > 2)
> labels(rulesInNATA)
[1] "{AZUCAR} => {DULCES}"
[2] "{TABLETA CHOCOLATE} => {DULCES}"
[3] "{AZUCAR,LECHE} => {DULCES}"
[4] "{LECHE, TABLETA CHOCOLATE} => {DULCES}"
[5] "{FRUTA, TABLETA CHOCOLATE} => {DULCES}"
[6] "{TABLETA CHOCOLATE, VERDURA} => {DULCES}"
[7] "{LECHE, PASTAS BOLLERIA} => {DULCES}"
[8] "{CAFE, LECHE} => {DULCES}"
[9] "{CAFE, QUESO} => {DULCES}"
[10] "{CAFE, VERDURA} => {DULCES}"
[11] "{CONSERVAS DE ATUN, LECHE} => {DULCES}"
[12] "{PASTEL BOLLERIA, YOGUR} => {DULCES}"
[13] "{QUESO, VERDURA, YOGUR} => {DULCES}"

```

Vemos un patrón curioso en estas reglas, respecto a la simetría que hemos mencionado, pues en las reglas que incluyen **dulces** en la parte izquierda aparecen dos referencias con **jamón york** y una con **jamón curado**; por tanto, parece que hay una pauta entre los compradores de galletas **dulces** en cuanto que también suelen comprar **jamón york** y **curado**. Sin embargo, no ocurre tan a menudo que cuando alguien compre estos dos tipos de productos también incluyan galletas **dulces**, ya que entre las reglas relevantes con **dulces** en el lado derecho de la regla en ninguna observamos **jamón york** ni **jamón curado**.

Con estas reglas podríamos elaborar, por tanto, un sistema de recomendación, en que, a medida que el usuario fuera añadiendo productos a su cesta, fueran apareciendo en la parte inferior o lateral de la pantalla productos de la categoría que aparece a la derecha de nuestras reglas de asociación.

Por ejemplo, si un usuario de nuestra web de *e-commerce* añadiese a su cesta un flan, podríamos recomendarle natillas; si un usuario añadiese una bebida de naranja con gas, podríamos recomendarle comprar a continuación una bebida de Cola; y, finalmente, si un comprador hubiese

añadido dulces y verdura, podríamos a continuación recomendarle jamón york. De esta forma, facilitaríamos al usuario la localización de los productos que probablemente estén en su lista de compra; además, la experiencia de compra en línea resultaría más agradable y ágil para el consumidor.

## 2. Análisis de opiniones para gestión del conocimiento

A continuación veremos como el *data scientist* de este caso puede analizar las opiniones que los clientes vierten en las redes sociales sobre la marca y los productos que nuestro distribuidor va a poner en su surtido en línea.

El *text mining* (minería de textos) es una de las ramas de la lingüística computacional que trata de obtener información y conocimiento a partir de conjuntos de datos que en principio no tienen un orden o no están dispuestos en origen para transmitir esa información.

La minería de textos comprende tres actividades fundamentales:

- 1) Recuperación de información; es decir, seleccionar los textos pertinentes.
- 2) Extracción de la información incluida en esos textos: hechos, acontecimientos, datos clave, relaciones entre ellos, etc.
- 3) Realización de minería de datos para encontrar asociaciones entre esos datos claves previamente obtenidos de entre los textos.

El *text mining* se apoya en otras técnicas como la categorización de texto, la recuperación de información y el procesamiento de lenguaje natural y el aprendizaje automatizado. Es la herramienta que nos faltaba para poder enlazar el conocimiento cuantitativo con el cualitativo. El objetivo, por ejemplo, es poder desglosar qué ha querido decir un cliente cuando nos ha dejado unas observaciones anotadas en una hoja de reclamaciones, y cómo afecta esto al negocio que nos genera dicho cliente y otros similares.

En el ámbito comercial, resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Este sistema de recuperación de información es capaz no solo de devolver objetos (como palabras clave, imágenes, etc.) relevantes para una consulta, sino también de inferir las actitudes de los emisores al mencionar los objetos en los resultados de búsqueda. Esas actitudes valorativas nos dan una información altamente relevante del sentimiento del emisor.

En empresas orientadas al cliente, esta técnica permite tener una información mayor basada en datos implícitos tras la recolección por medio de *feedback* del cliente: adelantarnos a las necesidades de los clientes, tener una información más completa sobre la valoración del servicio y mejora de la fidelización.

Dado que existe gran cantidad de información textual, esto nos permite encontrar conocimiento a partir de datos textuales sin estructurar.

La minería de textos constituye una herramienta de gran utilidad, ya que alrededor de un ochenta por ciento de la información de las organizaciones está almacenada en forma de texto no estructurado.

La minería de datos desestructurados se nutre de la información contenida en ficheros de texto, en Internet, ya que, gran parte de la información es desestructurada.

Las redes sociales son, por supuesto, una fuente inagotable de obtención de datos; de diferentes fuentes: libros, tuits, opiniones sobre artículos, etc.

La limpieza de datos también es conocida como la **fase de preprocesamiento**. El preprocesamiento resulta muy importante en la minería de textos, procesamiento de lenguaje natural y recuperación de la in-

formación. Básicamente consiste en limpiar o filtrar la información, eliminando aquellos elementos que no nos aportan información de interés; se dejan solo los que serán valiosos en la etapa posterior, para que esta fase sea lo más productiva, eficaz y eficiente posible.

En el caso de las redes sociales, cuyas fuentes de texto son comentarios, pueden provenir de Twitter. Estos son los pasos que se siguen para «limpiar» la información:

- 1) Los *hashtags* (#) son útiles en la búsqueda de los tuits que se desean analizar, por lo que el primer filtro se quedaría solo con aquellos *hashtags* definidos en las categorías de información por analizar.
- 2) La **tokenización** y detección de palabras derivadas, separando el texto por palabras o *tokens*.
- 3) Se eliminan signos de puntuación que no aportan información extra.
- 4) Se eliminan las *stop-words*: palabras que tampoco aportan información relevante, artículos, preposiciones, pronombres, artículos. Esto contribuirá a un mejor rendimiento de los algoritmos de procesamiento.
- 5) Detección de palabras derivadas (*stemming*): analiza la raíz de las palabras. Y si existen palabras repetidas con la misma raíz, se eliminan. Y se sustituyen por la palabra que mejor se adapta al significado de la raíz o lexema.
- 6) Transformación de mayúsculas a minúsculas.
- 7) Todos los tuits preprocesados de este modo se convierten en un único corpus; un conjunto grande y estructurado de textos, que será objeto de procesamiento en los análisis de opinión.

El análisis de opinión consiste en clasificar los comentarios de los clientes en tres grupos en función de su significado y la implicación que tiene en el análisis de resultados. Son tres las categorías en las que se agrupan las polaridades de las opiniones: comentarios positivos, negativos y neutros. Las **opiniones positivas** son las que usamos para describir algo que deseamos con agrado, un estado de felicidad o alegría, mientras que las **negativas** las utilizamos en ámbitos o escenarios circunstanciales que no deseamos.

Basado en si una palabra determinada evoca un sentimiento positivo, negativo o neutro, se construye un diccionario lexicográfico o **léxico**. Puede contener todas las palabras que se nos ocurran (existen léxicos ya contruidos) clasificadas según la polaridad que le asignemos. Cuando el corpus ya preprocesado se analiza y se «compara» con el léxico que usamos como base, mediante diferentes técnicas como la puntuación positiva o negativa, y diversos algoritmos, podemos obtener un valor que nos indique si el comentario expresa una opinión positiva, negativa o neutra. El resultado del análisis nos dará la información necesaria para que las empresas tomen las decisiones más acertadas para conseguir la mejor relación y valoración de sus clientes.

Cuando el algoritmo compara datos respecto a bases de datos de palabras con polaridad, utiliza lo que conocemos como **ontologías**. Existen dos grandes grupos de ontologías: las genéricas y las específicas de cada industria.

- **Generic Ontology**: representa conceptos generales que no son específicos de un dominio. Por ejemplo, ontologías sobre el tiempo, ontologías de conducta, de causalidad, etc. Pueden reutilizarse a través de diferentes dominios. Estas ontologías recogen palabras de todo el vocabulario sin diferenciar por categoría y se agrupan en palabras positivas y palabras negativas.
- **Industry Ontology**: las industrias que tienen como base la explotación de la información y del conocimiento. Este tipo de ontologías explicita los conceptos, las propiedades y las relaciones existentes propias en el dominio industrial.

Por ejemplo, una ontología sobre teléfonos móviles tendrá una recopilación de términos donde el concepto «reducido» puede tener un sentido positivo si el término se refiere a atributos del diseño, ya que en la industria de teléfonos móviles un diseño compacto es algo positivo. Sin embargo, el mismo término puede tener una connotación negativa si se refiere a los atributos de un diamante (un diamante de tamaño reducido implica algo malo).

La aplicación sobre el *marketing* de productos y servicios es evidente. Podemos conocer el grado de satisfacción de nuestros clientes



(para así saber si tenemos que mejorar o mantenernos en la misma línea); también qué es lo que estos desearían que una organización concreta les ofreciera; esto es realizar análisis predictivos sobre tendencias y consumos. Las ventajas, por lo tanto, son a todas luces obvias.

Previo a la definición del término *crawling* debemos delimitar el concepto **API**, que es una funcionalidad de una aplicación que permite a terceros desarrolladores usar algunas de sus características mediante librerías, bibliotecas o paquetes de funciones desarrolladas con esta finalidad.

Para hacer un análisis de opiniones, es necesario emplear alguna herramienta de *crawling*, para recoger los tuits de los usuarios de forma rápida y eficiente. Twitter pone a disposición de los usuarios su servicio de **Twitter Search** y también una **API** que envía los tuits que luego pueden ser recogidos por librerías de java, R y otros lenguajes de programación y *scripting*.

A continuación trataremos el caso de **TALC (Talcum Powder de Johnson & Johnson)**, un producto cosmético con efectos muy perjudiciales para la salud del consumidor que provocaba, hasta el momento de su retirada, desde irritaciones de piel hasta cáncer de óvulo (figura 12).

El producto generó una alarma social a partir de un juicio que enfrentó a la compañía y a la familia de una joven que murió, según la acusación, debido al uso del producto. J&J tuvo que pagar una compensación económica de setenta y dos millones de dólares. Dieciséis estudios realizados en 2003 demostraron que el uso del producto incrementaba el riesgo de padecer cáncer de ovario unas tres veces por encima de la media.

Figura 12. Talcum Powder



En el análisis recogeremos datos mediante R y la API de Twitter, y analizaremos el sentimiento general sobre el producto. Este análisis puede automatizarse y realizarse para todo el portfolio de referencias de una empresa de distribución de *e-commerce*, parecida a Amazon, con el fin de detectar productos impopulares que los *product manager* deberían eliminar rápidamente del surtido para desvincular la imagen de la empresa de productos perjudiciales para la salud y con mala aceptación.

Para hacer uso del API de Twitter seguiremos los pasos que siguen a continuación. El objetivo de esta primera fase es configurar una cuenta para establecer las contraseñas que permitirán la comunicación entre el API y nuestras librerías de R.

Twitter Search tiene ciertas limitaciones si lo que queremos es recoger tuits de forma dinámica. Para importar los relacionados con una palabra de interés, primero se ha de crear una aplicación. Seguid el siguiente enlace, [TwitterApps](https://apps.twitter.com), o bien poned directamente en el explorador «[apps.twitter.com](https://apps.twitter.com)».

Es necesario disponer de un usuario de Twitter para crear una aplicación (figura 13).

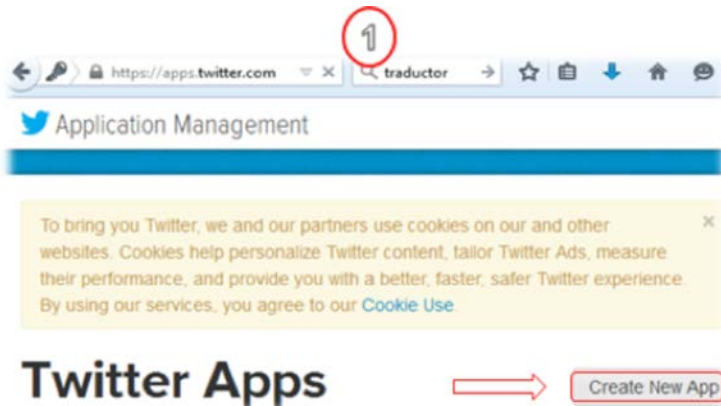
- 1) Desde Twitter Apps, haremos clic en «sign in».
- 2) A continuación, introduciremos nuestro usuario y nuestra contraseña.

Figura 13. Registro



Una vez abierta la sesión haced clic en «Crear nueva aplicación» (figura 14).

Figura 14. Creación de app

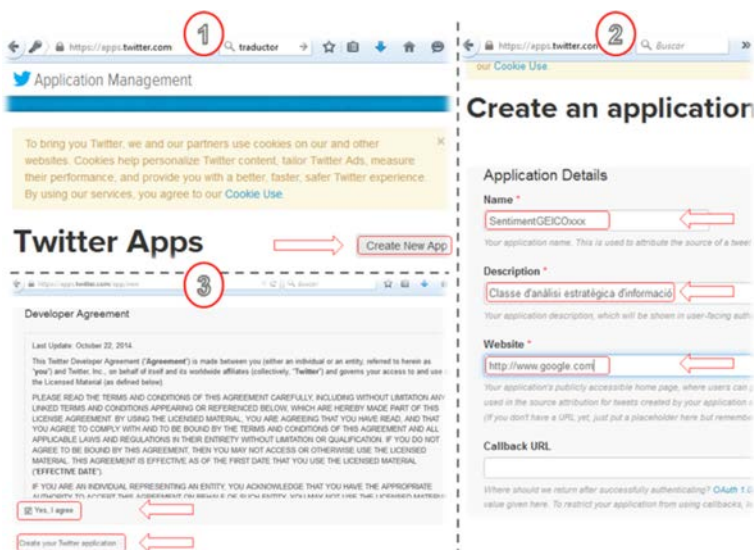


1) A continuación, introduciremos la siguiente información (figura 15):

- **Nombre de la aplicación.** No tiene que existir previamente; por lo tanto, en caso de introducir un nombre y que Twitter nos indique que ya existe, introduciremos otro, por ejemplo: app001, app002, app003...
- **Descripción:** cualquier descripción, por ejemplo: «Clase de AEI».
- **Website:** añadiremos un dominio válido, por ejemplo: <http://www.google.com>. Es importante no olvidar añadir «http://».
- **Callback URL:** no hay que rellenar.

- 2) Haremos clic en «Yes, I agree».
- 3) Para finalizar, clicad en «Create your Twitter application». Vuestro aplicación ya se ha creado.

Figura 15. Detalles de la aplicación



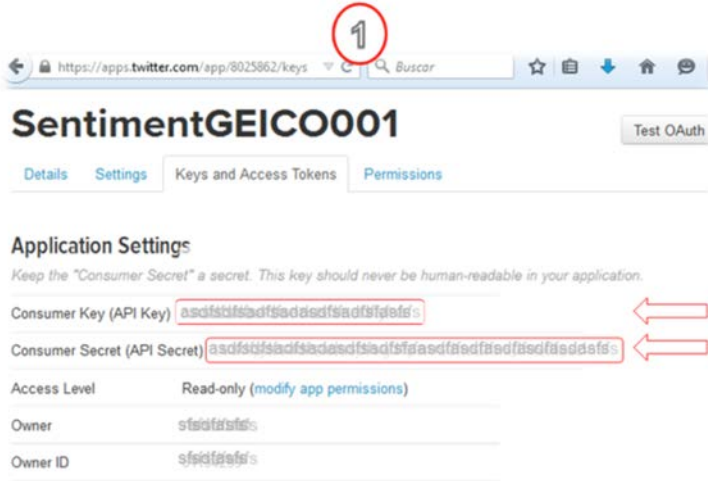
Es posible que Twitter nos pida registrar nuestro teléfono. Entonces, seguiremos las instrucciones del siguiente enlace y lo podremos añadir de forma muy rápida: Teléfono. Siempre lo podremos desvincular después de leer esta guía.

Es posible que Twitter no nos envíe el código si nuestro operador es extranjero. Será necesario tener un operador admitido por Twitter.

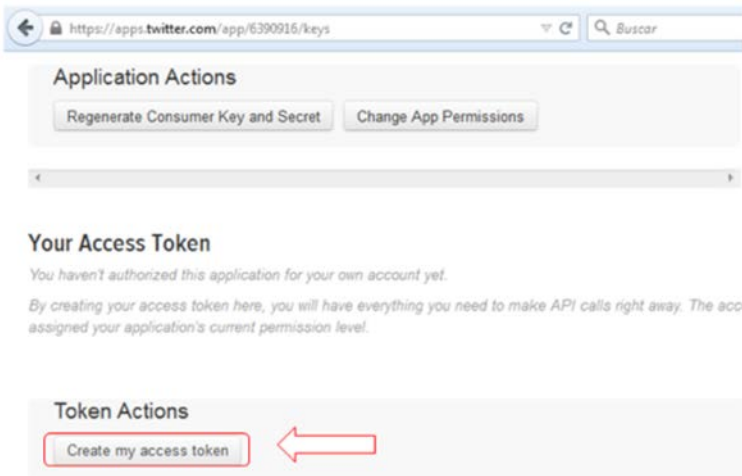
Seguidamente guardaremos la siguiente información del apartado «Key and Access Tokens», que nos será de utilidad más adelante (figura 16):

- *Consumer Key.*
- *Consumer Secret.*

### Figura 16. Configuración



Más abajo veremos la opción: «Token actions» y «Create my access token» (figura 17).

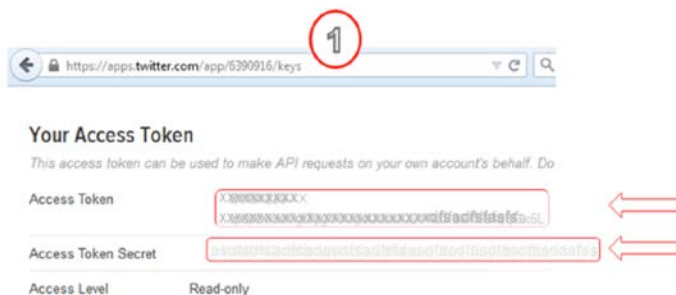
Figura 17. *Tokens*

Haremos clic en «Create my access token», y esto nos generará dos códigos adicionales que encontraremos en la parte inferior, con los nombres (figura 18):

- Access Token
- Access Token Secret

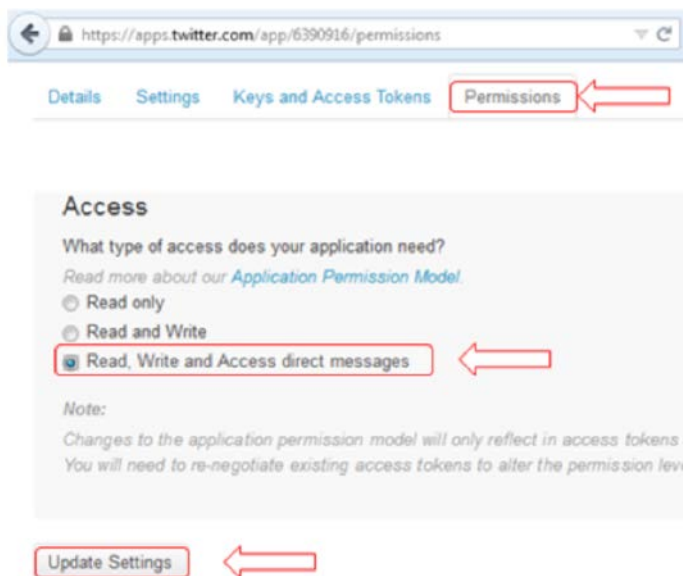
Tomaremos nota también de estos códigos, que utilizaremos más adelante.

Figura 18. Tokens de acceso



Finalmente daremos acceso de lectura y escritura a la aplicación, yendo al apartado «Permissions» donde escogeremos «Read, Write Access direct messages», y para finalizar haremos clic en «Update settings» (figura 19).

Figura 19. Actualizar



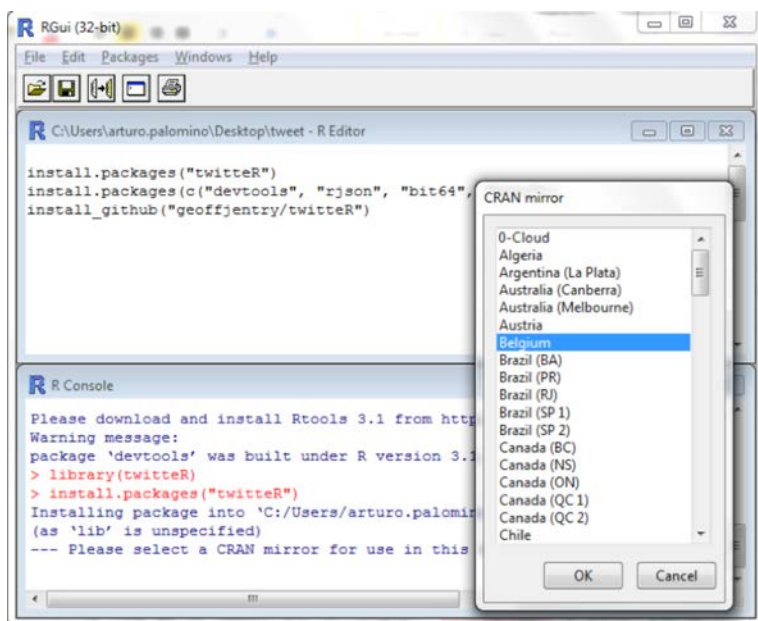
Para recoger tuits con R, introducimos los siguientes comandos en el editor.

```
install.packages("devtools")  
install.packages("twitter")
```

Ejecutaremos el código seleccionando las líneas y haciendo «control + R». Nos aseguraremos de que tenemos conexión a Internet antes de ejecutar.

El comando «install.packages» busca las librerías en un servidor externo. Escogeremos un servidor cercano y haremos clic en «OK». Puede tardar varios minutos. Habrá acabado cuando veamos en la consola el carácter «>» (figura 20).

Figura 20. Instalación



Una vez instalado el paquete «twitter», reiniciaremos R. Podemos guardar los *scripts* y, al reiniciar, abrirlos con la opción de menú: File > «open script». No es necesario guardar sesiones, solo los *scripts*.

Después de ejecutar los comandos «install.packages», los paquetes quedan guardados en R y no hay que ejecutarlos de nuevo en cada sesión. Para no volver a lanzarlas por error, dejaremos las líneas «comentadas» poniendo el carácter «#» delante, así:

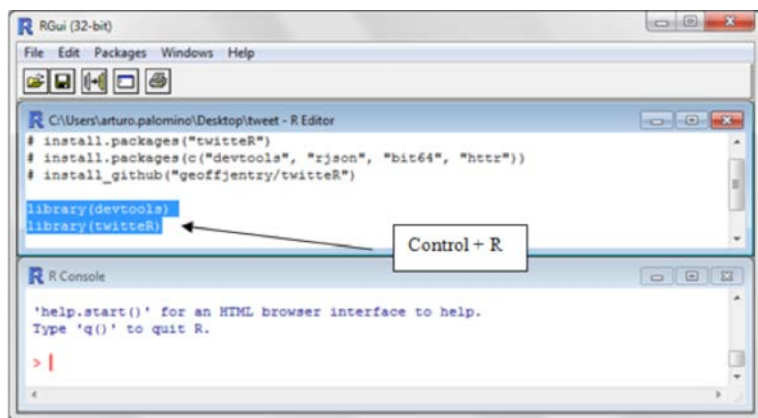
```
# install.packages("devtools")
# install.packages("twitterR")
```

A continuación, después de reiniciar R y volver al editor, se escriben los siguientes comandos:

```
library(devtools)
library(twitterR)
```

Ejecutaremos, seleccionando las dos líneas a la vez y haciendo «control + R» (figura 21).

Figura 21. Librerías



A continuación, introduciremos entre las comillas («xxx») los códigos que hemos guardado de TwitterApps (Consumer Key, Consumer Secret, Access Token y Access Token Secret). Seleccionaremos las líneas y haremos Control+R.

```
consumerKey <- "XXX"
```



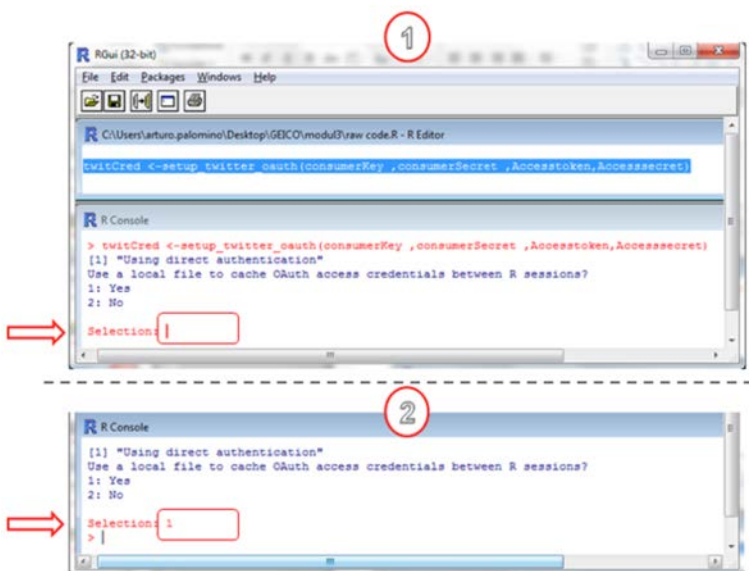
```
consumerSecret <- "XXX"  
Accesstoken <- "XXX"  
Accesssecret <- "XXX"
```

Añadiremos y ejecutaremos la siguiente función que hará la conexión con nuestra App y haremos «Control+R».

```
twitCred <-setup_twitter_oauth(consumerKey ,consumerSecret  
,Accesstoken,Accesssecret)
```

La consola nos pedirá introducir una opción; escogeremos 1. Pondremos 1 y pulsaremos «Enter» (figura 22).

Figura 22. Autenticación



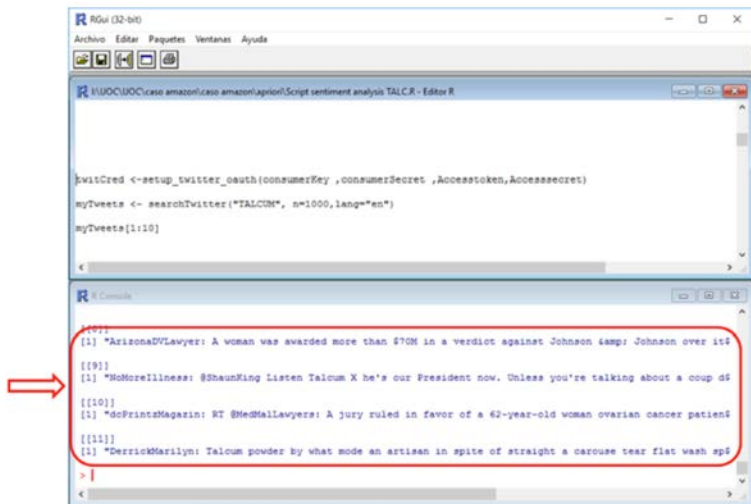
Para recoger todos los tuits que incluyan la palabra «TALC», por ejemplo, utilizaremos el siguiente comando. Con `n=1000` limitamos el número de tuits a mil. Por otro lado, escogemos el idioma inglés con la variable `lang=«en»`. El comando exacto a ejecutar es el siguiente:

```
myTweets <- searchTwitter("TALCUM", n=1000,lang="en")
```

Para ver los primeros diez tuits usaremos la siguiente línea de código, que tendremos que ejecutar (figura 23).

```
myTweets[1:10]
```

Figura 23. Revisión

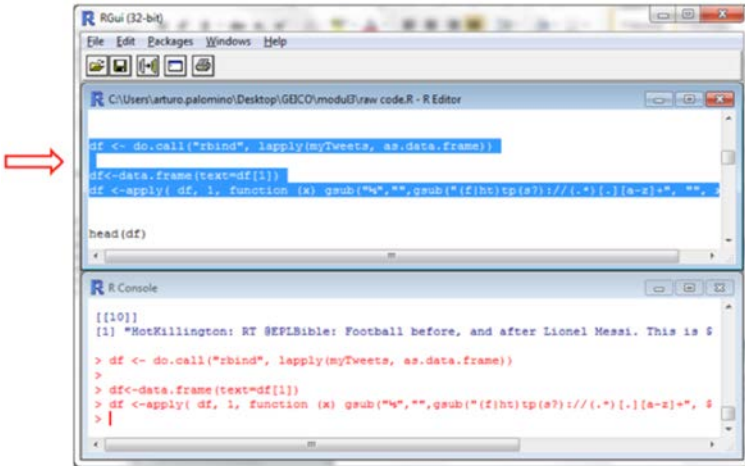


A continuación, nos quedaremos con los tuits del *crawling*, guardándolos en un *dataframe* denominado «df» (figura 24)<sup>3</sup>:

```
df <- do.call("rbind", lapply(myTweets, as.data.frame))
df<-data.frame(text=df[1])
df <-apply( df, 1, function (x) iconv(gsub("(f|ht)tp(s?)://
(.*)[.][a-z]+", "", x), "latin1", "ASCII", sub="")) )
```

3 Para información sobre *dataframes* revisaremos el capítulo 2.2.4 de Castillo, A. J. S. «Métodos Estadísticos con R y R Commander». [Recurso electrónico gratuito] <<http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>>

Figura 24. Dataframe



Ahora cargaremos una librería de *text mining* para la parte del *preprocessing* de los tuits.

```
install.packages("tm")
library(tm)
```

A continuación, para hacer el preprocesamiento, crearemos el corpus<sup>4</sup>.

Ejecutaremos la siguiente línea:

```
myCorpus <- Corpus(VectorSource(df))
```

---

4 Para más información leed:

Feinerer, I. (2015). «Introduction to the tm Package Text Mining in R». 2013-12-01.

Liau, Bee Yee; Tan, Pei Pei (2014). «Gaining customer knowledge in low cost airlines through textmining». *Industrial Management & Data Systems* (vol. 114, núm. 9, pág. 1344-1359).

Nguyen, Tung Thanh; Quan, Tho Thanh; Phan, Tuoi Thi (2014). «Sentiment search: an emerging trend on social media monitoring Systems». *Aslib Journal of Information Management* (vol. 66, núm. 5, pág. 553-580).

Pasaremos a minúscula el texto, eliminaremos puntuación, numeración, *stopwords*, espacios en blanco, etc.

```
myCorpus<- tm_map(myCorpus, content_transformer(tolower))
myCorpus <- tm_map(myCorpus , PlainTextDocument)
myCorpus <- tm_map(myCorpus, removePunctuation)
myCorpus <- tm_map(myCorpus, removeNumbers)
myStopwords <- c(stopwords('english'), "available", "via",
"http")
myCorpus <- tm_map(myCorpus, removeWords, c(myStopwords))
myCorpus <- tm_map(myCorpus , stripWhitespace)
```

Guardaremos una copia del corpus.

```
dictCorpus <- myCorpus
```

Cargaremos librerías adicionales.

```
install.packages("SnowballC")
install.packages("RWeka")
install.packages("rJava")
install.packages("RWeKajars")

library("SnowballC")
library("RWeka")
library("rJava")
library("RWeKajars")
```

Conservaremos la raíz léxica de las palabras haciendo *stemming*.

```
myCorpus <- tm_map(myCorpus, stemDocument)
```

Completaremos los lexemas (puede tardar varios minutos).

```
myCorpus <- tm_map(myCorpus, stemCompletion,
dictionary=dictCorpus)
```

Limpiaremos el corpus.

```
dataframe<-data.frame(text=unlist(sapply(dictCorpus[1:1000]
[1:1000], `[`, "content")), stringsAsFactors=F)
myCorpus<-Corpus(VectorSource(dataframe$text))
```

Filtraremos aquellas palabras con una frecuencia mínima; en este caso frecuencia 4 mínimo.

```
myDtm <- TermDocumentMatrix(myCorpus, control =
list(minWordLength = 4))
```

Visualizaremos las palabras más frecuentes, por encima de cierto umbral; y las más asociadas a TALC.

```
findFreqTerms(myDtm, lowfreq=5)
findAssocs(myDtm, 'TALCUM', 0.10)
```

Cambiaremos el formato de los datos a matriz.

```
m <- as.matrix(myDtm)
```

Para visualizar la información instalaremos un paquete adicional.<sup>5</sup>

```
install.packages("wordcloud")
library(wordcloud)
```

Haremos un *WordCloud* para ver la nube de palabras que rodean «TALC» más habitualmente a los tuits.

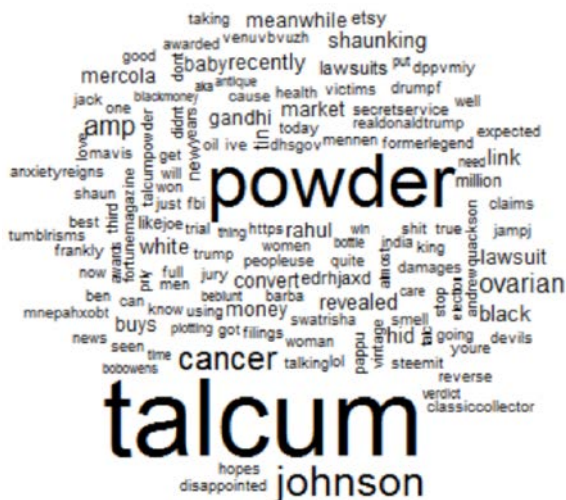
```
v <- sort(rowSums(m), decreasing=TRUE)
myNames <- names(v)
d <- data.frame(word=myNames, freq=v)
wordcloud(d$word, d$freq, min.freq=10)
```

---

5 Para más información leed Fellows, I.; Fellows, M. I.; Rcpp, L. (2012). «Package “wordcloud”». <<http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>>

La nube de palabras nos permite observar de una forma muy visual cuáles son las palabras que están más asociadas al término «TALC» en los mil tuits (figura 25).

**Figura 25. Nube de palabras**



Fuente: elaboración propia.

Podemos ver palabras como: Johnson, veredicto, legalidad, daños, cáncer, decepción, mujer, cáncer de óvulo. Hacen alusión a la alarma social que generó el producto y su relación con esta enfermedad mortal.

Para tener un indicador de sentimiento u opinión general de los tuits recogidos, cargaremos unas tablas de palabras subjetivas de diccionario.<sup>6</sup>

6 Para más información leed:

Liau, Bee Yee; Tan, Pei Pei (2014). «Gaining customer knowledge in low cost airlines through textmining». *Industrial Management & Data Systems* (vol. 114, núm. 9, pág. 1344-1359).

Nguyen, Tung Thanh; Quan, Tho Thanh; Phan, Tuoi Thi (2014). «Sentiment search: an emerging trend on social media monitoring Systems». *Aslib Journal of Information Management* (vol. 66, núm. 5, pág. 553-580).

Dinsoreanu, Mihaela; Potolea, Rodica (2014). «Opinion driven communities' detection». *International Journal of Web Information Systems* (vol. 10, núm. 4, pág. 324-342).

Copiaremos los dos ficheros facilitados con este documento, «**positive-words.txt**» y «**negative-words.txt**» en una ruta que recordamos fácilmente.

A continuación, copiaremos en los *scripts* las siguientes líneas, que tendremos que modificar, sustituyendo las rutas por aquella donde hemos guardado los dos ficheros adjuntos.

```
pw<-read.table("C:/RUTA DEL FITXER/positive-words.
txt",sep=";", stringsAsFactors = F)
nw<- read.table("C:/ RUTA DEL FITXER/negative-words.txt",
stringsAsFactors = F)
```

Deberemos sustituir la ruta «C:/RUTA DEL FICHERO/» por la ruta donde se encuentren los dos ficheros que hemos copiado. Seguidamente, ejecutaremos las dos líneas. Vigilaremos no confundir «/» con «\».

A continuación, copiaremos la siguiente función en el *script* y también lo ejecutaremos.

```
sentimen<- function(x){
  sentiment = 0
  palabras = 0
  NEGATIVAS=0
  POSITIVAS=0
  ratio1=0
  ratio0=0
  for (i in 1:length(myCorpus)) {#i<-1
    doc <- myCorpus[[i]]
    vw <- strsplit(doc[[1]],' ')
    s = 0
    P=0
    N=0
    for (w in vw[[1]]) { #w<-vw[[1]][6] w<-"unbeatable"
      palabras<-palabras+1
      if (length(which(w%in%pw[[1]])) > 0) {
        s = s + 1
        P=P+1
      }
    }
  }
}
```

```

    }
    if (length(which(w%in%nw[[1]])) > 0) {
        s = s -1
        N=N+1
    }
}

sentiment = sentiment + min(max(-1,s),1)
if (s>0) {ratio1=ratio1+1}
if (s<0) {ratio0=ratio0+1}
NEGATIVAS=NEGATIVAS+N
POSITIVAS=POSITIVAS+P
}

cat(paste("OPINIONS SOBRE TOTAL TWEETS: ",
round((ratio0+ratio1)/length(myCorpus),2)))
cat("\n")cat(paste ("-----
-----"))
cat("\n")
cat(paste("TOTAL PARAULES POSITIVES ",
POSITIVAS))
cat("\n")
cat(paste("TOTAL PARAULES NEGATIVES ",
NEGATIVAS))
cat("\n")
cat(paste("TOTAL PIULADES POSITIVES ",
ratio1))
cat("\n")
cat(paste("TOTAL PIULADES NEGATIVES ",
ratio0))
cat("\n")
cat(paste ("-----
-----"))
cat("\n")
cat(paste("SENTIMENT (0=NEUTRE, >0 POSITIU, <0 NEGATIU)
", sentiment))
cat("\n")
cat(paste("SENTIMENT POSITIU SOBRE TOTAL OPINION (0.5
neutre) ", round(ratio1/(ratio0+ratio1),2) ))
cat("\n")

```



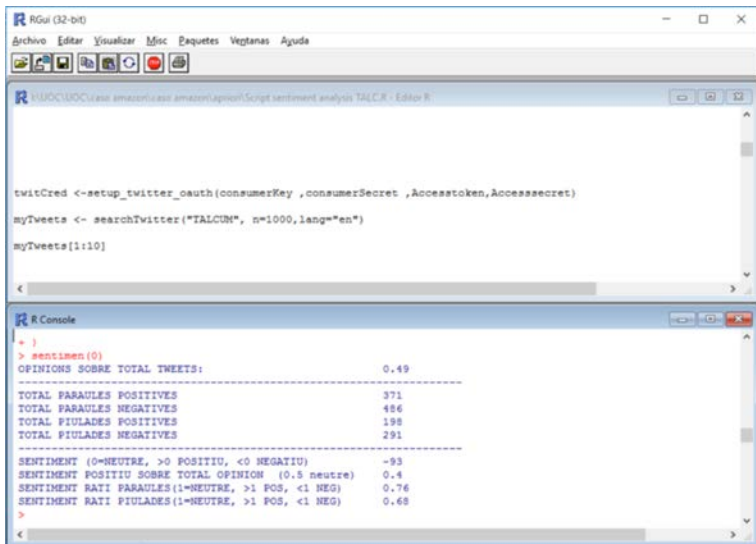
```

cat(paste("SENTIMENT RATI PARAULES(1=NEUTRE, >1 POS, <1
NEG)      ", round(POSITIVAS/ ifelse(NEGATIVAS==0,0.001,NEGA
TIVAS),2 ) ))
cat("\n")
cat(paste("SENTIMENT RATI PIULADES(1=NEUTRE, >1 POS, <1
NEG)      ", round(ratio1/ ifelse(ratio0==0,0.001,ratio0),2 )
))
cat("\n")
}
sentimen(0)

```

El resultado será similar al siguiente (figura 26). No coincidirá exactamente, dado que depende del momento de ejecución:

**Figura 26. Resultado**



Lo que indica este resultado (figura 26) es lo siguiente:

- **OPINIONES SOBRE TOTAL DE TUIITS:** Indica el número de tuits que expresan una opinión no neutra (positiva o negativa).

En este caso, el 49% de los tuits expresaban una opinión en uno u otro sentido (polaridad).

- **TOTAL PALABRAS POSITIVAS:** Indica el número de palabras que expresan una opinión positiva. En este caso, se han encontrado 371 palabras positivas del total.
- **TOTAL PALABRAS NEGATIVAS:** Indica el número de palabras que expresan una opinión negativa. En este caso, se han encontrado 486 palabras negativas del total.
- **TOTAL TUIITS POSITIVOS:** Indica la cifra de tuits donde el número de palabras positivas por tuit superaba el número de palabras negativas. En este caso se han encontrado ciento noventa y ocho tuits con sentido positivo.
- **TOTAL TUIITS NEGATIVOS:** Indica la cifra de tuits donde el número de palabras negativas por tuit superaba el número de palabras positivas. En este caso, se han encontrado doscientos noventa y un tuits con sentido negativo.
- **SENTIMIENTO:** Indica el número de tuits positivos, menos el número de tuits negativos. De tal forma que, si el resultado es igual a 0, el sentimiento general es neutro; si el resultado es superior a 0, entonces el sentimiento general es positivo; negativo en el otro caso.
- **SENTIMIENTO SOBRE TOTAL DE OPINIONES:** Expresa una ratio de número de tuits positivos entre el total tuits que expresan opinión positiva o negativa. Valores inferiores a 0,5 indican opinión negativa general; valores superiores a 0,5 indican opinión positiva general.
- **SENTIMIENTO RATIO PALABRAS:** Expresa la ratio de número de palabras positivas entre el total palabras negativas. Una ratio superior a 1 indica opinión positiva general; mientras que una ratio inferior a 1 indica opinión general negativa.
- **SENTIMIENTO RATIO TUIITS:** Expresa la ratio de número de tuits positivos entre el total tuits negativos. Una ratio superior a 1 indica opinión positiva general; mientras que una ratio inferior a 1 indica opinión general negativa.

El análisis claramente nos revela que este producto no está bien visto por el consumidor; ante este resultado, la decisión del distribuidor tiene que ser evitar su venta al cliente, teniendo en cuenta que no

solo va a perjudicar su imagen y le va a traer menos ventas; en este caso tan extremo, también estamos poniendo en riesgo la salud de nuestros clientes.

# Bibliografía

- Castillo, A. J. S. «Métodos Estadísticos con R y R Commander». [Recurso electrónico gratuito] <<http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>>
- Dinsoreanu, M.; Potolea, R. (2014). «Opinion driven communities' detection». *International Journal of Web Information Systems* (vol. 10, núm. 4, pág. 324-342).
- Feinerer, I. (2015). «Introduction to the tm Package Text Mining in R». 2013-12-01.
- Fellows, I.; Fellows, M. I.; Rcpp, L. (2012). «Package "wordcloud"». <<http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>>
- Liau, B. Y.; Tan, P. P. (2014). «Gaining customer knowledge in low cost airlines through textmining». *Industrial Management & Data Systems* (vol. 114, núm. 9, pág. 1344-1359).
- Nguyen, T. T.; Quan, T. T.; Phan, T. T. (2014). «Sentiment search: an emerging trend on social media monitoring Systems». *Aslib Journal of Information Management* (vol. 66, núm. 5, pág. 553-580).
- Santana, J. S.; Farfán, E. M. «El arte de programar en R: un lenguaje para la estadística». <[http://cran.r-project.org/doc/contrib/Santana\\_El\\_arte\\_de\\_programar\\_en\\_R.pdf](http://cran.r-project.org/doc/contrib/Santana_El_arte_de_programar_en_R.pdf)>

## Apartado Estadística

- Baró Llinàs, J.; Alemany Leira, R. (2002). «Anàlisi cluster». *Estadística II*. Barcelona: Universitat Oberta de Catalunya.
- Baró Llinàs, J.; Alemany Leira, R. (2002). «Sèries temporal s». *Estadística II*. Barcelona: Universitat Oberta de Catalunya.
- Baró Llinàs, J.; Alemany Leira, R. (2008). «El model de regressió múltiple». *Estadística II* (4a. ed.). Barcelona: Universitat Oberta de Catalunya.

- Gibergans Bàguena, J.; Gil Estallo, A. J.; Rovira Escofet, C. (2009).**  
«Anàlisi de la variància». *Estadística* (4a. ed.). Barcelona: Universitat Oberta de Catalunya.
- Gibergans Bàguena, J.; Gil Estallo, A. J.; Rovira Escofet, C. (2009).**  
«Combinatòria i tècniques de recompte». *Estadística* (4a. ed.). Barcelona: Universitat Oberta de Catalunya.
- Greenacre, M. M. (2008).** «Inferència estadística: part 4 i 5». *Estadística I* (3a. ed.). Barcelona: Universitat Oberta de Catalunya.
- Salvador Figueras, M. (2003).** «Análisis de correspondencias».

# ¿Qué es H2PAC?

El modelo H2PAC resuelve propuestas clave a partir de ACTIVIDADES.

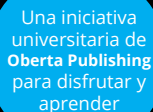
Esta forma de aprendizaje parte de un **RETO**: la actividad que deberás resolver. Para ello te facilitamos un contenido teórico, **EL CONOCIMIENTO IMPRESCINDIBLE**, que te ayudará a entender los conceptos esenciales para poder afrontar el desafío planteado inicialmente.

Además del contenido teórico, el modelo también te facilita **LAS SOLUCIONES**, una propuesta de resolución del reto expuesto.

El reto de esta obra es la aplicación del análisis de información como herramienta estratégica en un contexto empresarial. La obra hace una recopilación de técnicas de análisis, y plantea ejemplos contextualizados y aplicados.



**H2PAC**



Una iniciativa  
universitaria de  
**Oberta Publishing**  
para disfrutar y  
aprender

**UOC** Universitat  
Oberta  
de Catalunya



PID\_00242822