Big data en el mundo del *retail:* segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa

César Rogelio Cam Gensollen
https://orcid.org/0000-0003-1935-3853
Universidad de Lima, Facultad de Ingeniería y Arquitectura, Lima, Perú

Recibido: 31 de mayo del 2021 / Aprobado: 29 de junio del 2021
doi: https://doi.org/10.26439/ing.ind2022.n.5808

RESUMEN. En esta investigación se presentan los conceptos y técnicas utilizados en un proyecto de *big data* para una compañía europea de supermercados. Se propuso la segmentación de clientes, utilizando el algoritmo k-medias, y un sistema de recomendación a través de la librería LightFM de Python. Entre las principales conclusiones, se puede indicar la importancia de definir adecuadamente el problema por resolver, el uso correcto de la infraestructura de *big data*, y la relevancia del análisis exploratorio del conjunto de datos y su preprocesamiento, así como la aplicación de la metodología de proyectos TDSP (*Team Data Science Process*), orientada a los proyectos de *big data*.

PALABRAS CLAVE: *retail* / segmentación / sistema de recomendación / *big data* / aprendizaje automático

BIG DATA IN THE RETAIL WORLD: CUSTOMER SEGMENTATION AND RECOMMENDER SYSTEM IN A EUROPEAN SUPERMARKET CHAIN

ABSTRACT. In this research, we want to present the concepts and techniques used in a big data project for a European supermarket company, through a customer segmentation proposal, using the k-means algorithm, and a recommender system, via Light FM library. The main conclusions include the importance of adequately defining the problem to be solved, the correct use of the big data infrastructure, the relevance of the exploratory analysis of the dataset and its pre-processing, as well as the use of the TDSP methodology (Team Data Science Process), oriented to big data projects.

KEYWORDS: retail / segmentation / recommender system / big data / machine learning

Correo electrónico: crcam@ulima.edu.pe

INTRODUCCIÓN

De acuerdo con Schermann et al. (2014), el término *big data* resume los desarrollos tecnológicos en el área de almacenamiento y procesamiento de datos, que brindan la posibilidad de manejar aumentos exponenciales en el volumen de datos presentados en cualquier tipo de formato en periodos de tiempo que disminuyen constantemente (Chen, Chiang & Storey, 2012; Lycett, 2013). El *big data* brinda la oportunidad no solo de manejar, sino también de usar y agregar valor a grandes cantidades de datos provenientes de redes sociales, imágenes y otras tecnologías de información y comunicación (Schermann et al., 2014).

Por otro lado, en estos tiempos de gran incertidumbre, el negocio de *retail* necesita reinventarse por múltiples motivos: pandemia, exceso de competencia, consumidores más exigentes, presión regulatoria, globalización, entre otros. Y es ahí donde el *big data* y toda su potencia pueden entrar en juego para ayudar a las empresas a monetizar los datos, en particular para comprender hábitos de consumo y poder atender a los clientes de forma más eficiente (Cam et al., 2020). En ese sentido, se usará la información de una cadena de supermercados localizada en más de veinte países de Europa. Los datos para el presente trabajo corresponden a tiendas localizadas en España.

El propósito de este trabajo es presentar los pasos que se pueden desarrollar para poder atender estos retos a través de una propuesta de *big data*, desplegando la infraestructura en nube para procesar grandes volúmenes de información y así comprender los hábitos de consumo de los clientes. Esto permite desarrollar una propuesta de segmentación de clientes y un sistema de recomendación que dan soporte al desarrollo de acciones comerciales focalizadas por cada segmento de clientes.

Como antecedentes de la propuesta de segmentación de clientes utilizando el algoritmo de k-medias, la revisión de literatura menciona, entre otros, los trabajos de Chen, Sain y Guo (2012); Pascal et al. (2015); Aryuni et al. (2018), y Kansal et al. (2018). De igual forma, con respecto a los sistemas de recomendación, en la revisión de literatura correspondiente, se pueden hallar, entre otros, los trabajos de Christodoulou et al. (2017) y de Fang et al. (2018).

La contribución principal del presente trabajo puede considerarse en dos sentidos. El primero de ellos es la aplicación práctica del aprendizaje automático a dos situaciones específicas: la segmentación de clientes y el desarrollo de un sistema de recomendación en una cadena de supermercados. El segundo aporte corresponde a la utilización de forma práctica y concreta de la metodología TDSP a la gestión de este proyecto.

De acuerdo con la metodología de proyectos propuesta, se han seguido estas etapas: la comprensión del negocio, la captura de datos, el modelado y, finalmente, la aceptación y puesta en producción. En la parte de arquitectura de *big data*, se han utilizado los microservicios de AWS.

MATERIALES Y MÉTODOS

Definición del problema

El problema consiste en identificar hallazgos relevantes en los datos, que permitan proponer acciones comerciales. Por lo tanto, los retos que plantea este problema son estos:

- Entender qué le interesa al consumidor
- Proponer una segmentación de consumidores
- Crear una infraestructura de *big data*
- Proponer un sistema de recomendaciones que permita personalizar ofertas para incrementar la facturación y fidelizar la cartera de clientes

Propuesta de solución

La segmentación de clientes consiste en clasificar a los consumidores en diferentes grupos según ciertas características, necesidades o deseos comunes. En ese orden de ideas, es importante indicar que, en la gran mayoría de los casos, las empresas son conscientes de la importancia de la segmentación de mercado; sin embargo, no conocen cómo desarrollar un proceso de segmentación eficiente o cómo aplicarla, por lo que pierden mucha efectividad cuando se dirigen al consumidor. Esto provoca que los programas de fidelización y las promociones no tengan éxito, además del desperdicio de recursos de *marketing* (Doğan et al., 2018).

Asimismo, la segmentación es importante para que la empresa pueda crear segmentos rentables y reaccionar al segmento seleccionado en función de sus ventajas competitivas (Doğan et al., 2018).

Se plantea la propuesta de solución a través de las siguientes etapas:

- Primera etapa:
  - Análisis descriptivo de la muestra de datos
  - Propuesta de fuente externa para lograr una mejor segmentación
  - Primera segmentación
- Segunda etapa:
  - Creación de una infraestructura de *big data* en nube
  - Análisis descriptivo del conjunto de datos
  - Optimización del modelo de segmentación utilizando fuentes de datos externas propuestas en la primera segmentación

  - Definición e implementación de un modelo de recomendación de productos sobre la base de los segmentos obtenidos
- Tercera etapa:
  - Propuesta y ejecución de análisis avanzado
  - Presentación de resultados a la unidad de negocio:
    - Segmentos obtenidos
    - Hallazgos relevantes
    - Resultados del modelo de recomendación
    - Propuesta de acciones o comunicaciones segmentadas a los consumidores, en función de sus intereses

Gestión de proyectos

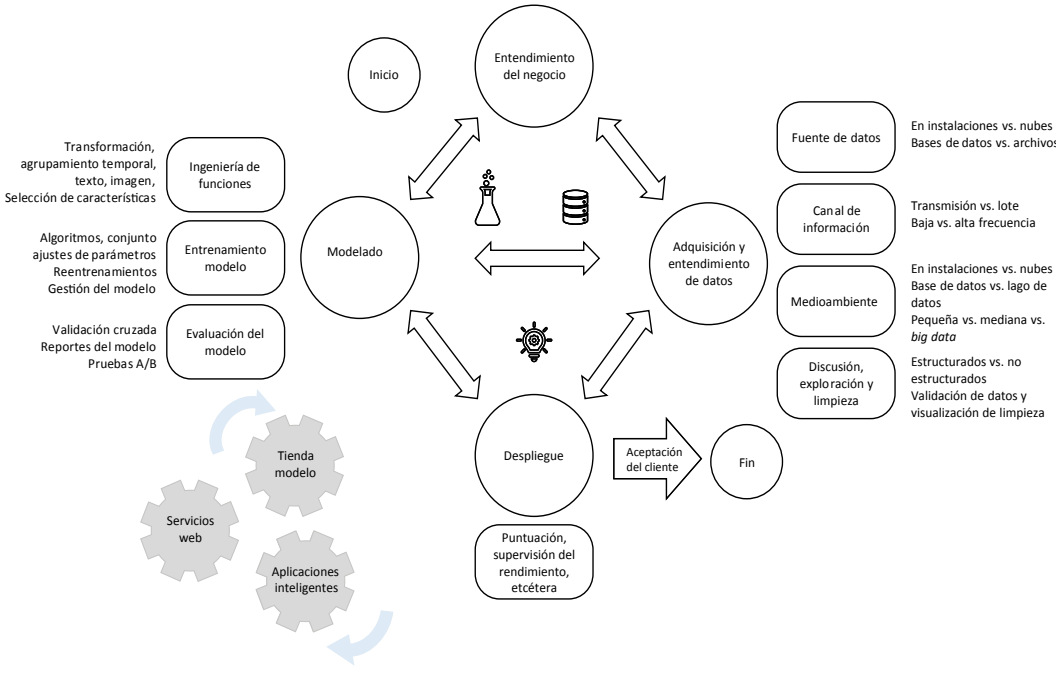
Un componente relevante dentro de un proyecto de *big data* es la propia gestión del proyecto, porque ayuda a cumplir con los objetivos que se está buscando alcanzar, y a definir los roles y tareas asociadas al proceso de entrega de valor.

Otro componente importante es la exploración y experimentación. Las áreas de negocio no son capaces de definir requisitos detallados al principio, y lo más probable es que, antes de encontrar un buen modelo, se tendrá que probar y descartar otros. Los sucesivos refinamientos de modelos ya incluyen el cambio como un elemento fundamental y beneficioso (Cam et al., 2020).

Desde ese punto de vista, se considera que la metodología TDSP (*Team Data Science Process*) de Microsoft proporciona el marco de referencia necesario, por tratarse de una metodología ágil e iterativa que permite entregar soluciones y aplicaciones en contextos de *big data*, así como definir los roles que intervienen en el proyecto (Microsoft, 2021). La figura 1 muestra la representación gráfica del ciclo de vida de TDSP.

Figura 1

Representación gráfica del ciclo de vida de TDSP



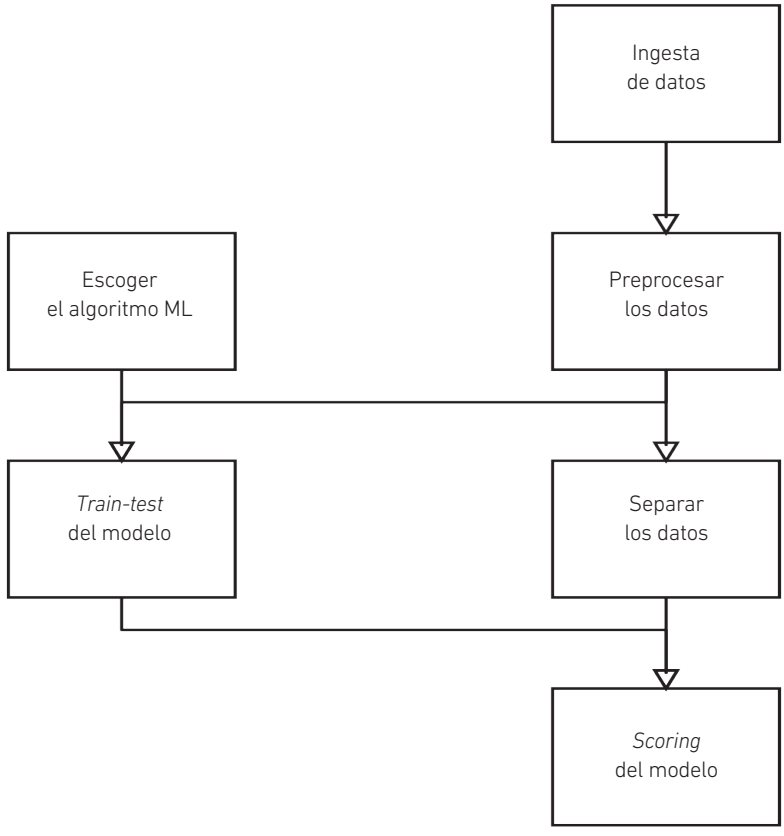
Nota. De *What is the Team Data Science Process*, por Microsoft, 2021 (https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview). Derechos de autor 2021 Microsoft.

El ciclo de vida de un proyecto desde la metodología TDSP comprende las siguientes etapas iterativas:

- *Comprensión del negocio*. Conocer el modelo de negocio y, por ende, el modelo de datos de la empresa en donde se está ejecutando el proyecto.
- *Captura de datos*. Producir un *dataset* de alta calidad y desarrollar una tubería (*pipeline*) que permita trabajar adecuadamente con los datos. Para lograr esto, se deben desarrollar tres actividades: ingesta de datos, análisis exploratorio de los datos (preprocesar los datos) y configuración de una arquitectura de *big data*.
- *Modelado*. En esta fase se está considerando el proceso de modelado (véase la figura 2).

Figura 2

Proceso de modelado



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 6), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

- *Aceptación y puesta en producción*. Entre los detalles del producto de datos que se entregará al área usuaria, se tienen las siguientes consideraciones:
  - Iterar y entregar lo más pronto posible al área usuaria a fin de conseguir retroalimentación temprana.
  - El área usuaria valida los modelos y se procede a su operacionalización, considerando que se deben desplegar en la forma en que serán utilizados por el área usuaria.
  - Con retroalimentación del área usuaria, definir un producto mínimo viable (MVP) que puede darse a través de los servicios web, tableros o aplicaciones corporativas.

Entonces, al utilizar esta metodología de proyectos TDSP en este caso de aplicación, se tiene lo siguiente:

Comprensión del negocio

De acuerdo con lo presentado anteriormente, el contexto de negocios para este caso de aplicación es el de una cadena de supermercados. Por lo tanto, el modelo de negocios se sustenta en la atención presencial a través del formato de tiendas (canal presencial) y la atención virtual a través de la página web (canal digital). En ambos canales, la empresa tiene una oferta amplia y variada de productos perecibles y envasados, orientados principalmente al consumo de alimentos. Una gran cantidad de tiendas a lo largo de la nación aumenta la presencia y cobertura de la cadena. El canal virtual está dirigido a atender a los clientes que buscan experiencias más digitales al momento de la compra; por lo tanto, es de especial importancia el contenido relevante que se vaya a colocar en la página web, así como la facilidad de navegación y de pago con medios digitales.

Los retos para esta empresa están en entender qué segmentos de clientes pueden identificar desde la perspectiva de los datos y desarrollar un sistema de recomendación que ayude a aumentar las ventas, de acuerdo con los gustos y preferencias de los clientes.

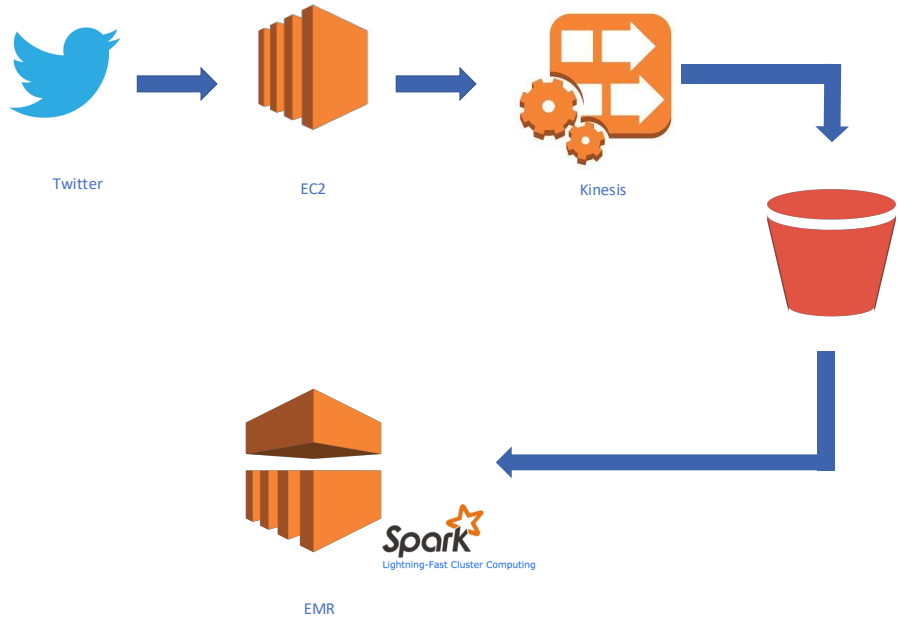
Captura de datos

Para el presente proyecto, se usarán dos tipos de fuentes de datos:

- *Fuentes de datos internas*. Conformadas por la información interna de la empresa que fue entregada para este propósito, se trata de fuentes supervisadas y son cuatro bases de datos que contienen la información de productos (*products*), tiendas (*stores*), clientes (*customers*) y boletas (*tickets*). En la primera segmentación, se utilizará una muestra del conjunto de datos y, en la segunda segmentación, se usa la totalidad de datos disponibles. Esto es así con el objetivo de hacer una primera aproximación a la solución.
- *Fuentes de datos externas*. Con la finalidad de enriquecer el proceso de segmentación de clientes, se propone la captura de tuits de los clientes, dado que la compañía realiza una fuerte promoción al uso de la aplicación móvil a fin de aumentar la venta en el canal digital. Posteriormente, se propone la utilización de la información oficial del Instituto Nacional de Estadística (INE) y de Nutriscore (Cam et al., 2020).

Figura 3

Arquitectura de big data



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 9), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Una vez concluido el primer análisis exploratorio con la muestra de datos, se procede a desplegar la siguiente arquitectura de *big data* en Amazon Web Services (AWS):

- *EMR (siglas en inglés de Amazon Elastic MapReduce)*. Clúster configurado con un nodo maestro y dos nodos esclavos para el entrenamiento y calificación del modelo, la generación de la segmentación y el sistema de recomendación. El criterio principal para elegir las instancias fue la memoria, dado que el procesamiento del *script* de Python utiliza *dataframes* de Pandas que carga a memoria los datos en tiempo de ejecución. Se seleccionan instancias de 32 GB de memoria (un maestro y dos esclavos), y se utiliza la configuración "m5.2xlarge", pues se realiza una prueba con la configuración inmediatamente menor a esta, "m5.xlarge" de 16 GB de memoria; sin embargo, al ejecutar la unión entre las tablas de *tickets* y productos, se obtiene un error de *out of memory* y no fue posible ejecutar el *script* de Python (Gulabani, 2017).
- *EC2 (abreviatura en inglés de Amazon Elastic Compute Cloud)*. Se configura una instancia de EC2 independiente para la captura de la información de Twitter y se apalanca en el servicio de Kinesis Firehose (servicio para cargar datos en

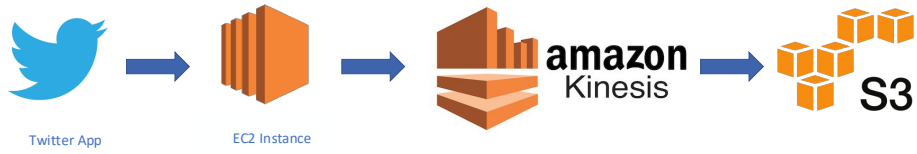
tiempo real de manera fiable) para la ingesta de la información al *bucket* de S3 (abreviatura en inglés de Amazon Simple Storage Service). Un *bucket* es un contenedor de objetos. Un objeto es un archivo y cualquier metadato que describa ese archivo.

- *S3*. Se utiliza un *bucket* de S3 como sistema de almacenamiento.

Tal como se propone en la etapa de captura de datos, se utiliza Twitter como fuente de datos externa. Para tal efecto se ha considerado la tubería representada en la figura 4.

Figura 4

Tubería para captura de tuits



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 10), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Modelado

En esta etapa del proyecto, la idea es poder identificar las variables más representativas que serán utilizadas en los modelos de aprendizaje automático. Este paso es de vital importancia, pues requiere que el equipo del proyecto sea capaz de identificar adicionalmente las variables explícitas, aquellas que contienen potenciales hallazgos relevantes y que pueden generar un alto valor para el negocio. En otras palabras, si bien los datos "crudos", una vez preprocesados, se pueden usar directamente en los modelos de aprendizaje automático, a menudo es necesario identificar ciertas relaciones o variables que no son tan explícitas y que pueden enriquecer el modelamiento; de ahí que sea de especial interés en esta actividad la participación de los analistas del negocio a fin de contextualizar los requerimientos y preguntas que los diferentes modelos deben tratar de solucionar (Cam et al., 2020).

Al identificar las mejores relaciones y variables, se deben realizar dos tareas en paralelo: seleccionar los modelos de aprendizaje automático que se van a utilizar y separar los datos en grupos de datos para entrenamiento y prueba (*training* y *testing*), con la finalidad de ejecutar los modelos seleccionados. Una vez obtenidos los resultados, se procede a medir la *performance* de estos, por ejemplo, con la métrica RMSE (el error

cuadrático medio mide la cantidad de error entre dos conjuntos de datos y es una de las estadísticas más usadas) o la matriz de confusión, según corresponda. Esta revisión del rendimiento de los algoritmos debe hacerse de forma conjunta con los analistas del negocio, de forma tal que se puedan recoger sus impresiones respecto de cuáles son los algoritmos que se ajustan más a las necesidades del negocio y si es necesario replantear las relaciones y variables o refinar los modelos de aprendizaje automático con la finalidad de obtener mejores resultados (Cam et al., 2020). Todo esto, tal como se muestra en figura 1.

Primera segmentación

En este primer ejercicio de generación de segmentos (clústeres), se utilizará la muestra de datos del primer análisis descriptivo (muestra de datos). Para este fin, se hará la generación de clústeres a través de k-medias, que es un algoritmo iterativo donde el número *k* de clústeres o segmentos está predeterminado y el algoritmo iterativamente asigna cada dato a uno de los *k* clústeres en función de la similitud de características (Cam et al., 2020). Este algoritmo pertenece al ámbito de los algoritmos no supervisados, dado que las observaciones que se desea segmentar no cuentan con una etiqueta que permita determinar de qué grupo es cada dato.

De acuerdo con Pérez (2013), k-medias es el algoritmo más importante de clasificación no jerárquica desde los puntos de vista conceptual y práctico. Parte de unas medias arbitrarias y, mediante pruebas sucesivas, contrasta el efecto que sobre la varianza residual tiene la asignación de cada uno de los casos a cada uno de los segmentos. En otras palabras, se busca que cada dato se encuentre muy cerca de los de su mismo segmento y los segmentos lo más lejos posible entre ellos.

En el presente trabajo, se utiliza el criterio gráfico del "codo" para especificar el número *k* de segmentos (clústeres) por ser encontrados, ya que este es un método que utiliza la distancia media de los datos a su centroide. Eso significa que se fija en las distancias dentro del clúster; por lo tanto, cuanto más grande es el número de segmentos *k*, la varianza intraclúster tiende a disminuir.

En esta primera segmentación, se va a construir un juego de datos con la base de datos *customers* como la base principal, buscando asociaciones con las otras tres bases de datos a través de variables sintéticas. A efectos de identificación, se mantienen el código del cliente y la edad.

Las variables sintéticas que se van a crear son las siguientes: antigüedad de cliente (fecha del último *ticket* menos la fecha de registro como cliente) en la tabla de *customers*; y, en la tabla de *tickets*, promedio de visitas al mes (promedio de *tickets* generados por un cliente en un mes), artículos por mes (promedio de artículos adquiridos por mes por un cliente), cancelado por mes (promedio de pagos mensuales de un cliente) y descuento

por mes (promedio de descuentos recibidos al mes por un cliente). Estas variables se complementan con siete variables sintéticas que se desarrollan para los siete grupos de mercancías de la tabla de *products* (promedio mensual de consumo por grupo de mercancías por cliente en cada visita).

Asimismo, se procederá en ambas segmentaciones al proceso de normalización de datos a fin de evitar que las variables con mayores unidades tengan mayor influencia en la distancia. Asimismo, los valores nulos, duplicados o inconsistentes serán eliminados.

Segunda segmentación

En este segundo ejercicio de generación de segmentos, se explora la posibilidad de enriquecer los datos de la compañía de supermercados con datos externos a la organización, como los tuits que los clientes emiten en la cuenta oficial de la empresa y que permiten valorar si esta red social puede ayudar a enriquecer el proceso de segmentación. Asimismo, se contribuirá con el área de Mercadotecnia y Digital, que tiene interés en investigar sobre cómo el coronavirus ha afectado al sector *retail* y el apoyo que las redes sociales suponen para la compañía.

Para este efecto, se usarán las infraestructuras de *big data* presentadas en las figuras 5 y 6; de esta manera, se pueden gestionar mayores volúmenes de datos y recoger tuits en tiempo real. La propuesta consiste en analizar la información que se recoge en tiempo real y así poder evaluar si es posible obtener información que pueda complementar el presente trabajo, tanto para la segmentación como para el sistema de recomendación.

Dadas las lecciones aprendidas de la segmentación inicial, los atributos seleccionados para este proceso de segmentación son antigüedad, promedio de visitas, cancelado por mes, descuento por mes y la nueva columna que se crea al calcular el gasto en el supermercado en relación con la renta media de la comunidad. Con este juego de datos, se procede a utilizar el algoritmo de k-medias, con un valor sugerido de cinco segmentos (criterio gráfico del "codo"), lo cual arroja una clasificación adecuada. De las diversas iteraciones, se puede concluir que la variable antigüedad es la predominante en el proceso de segmentación; por ende, la integridad entre las tablas de clientes y *tickets* resulta fundamental para la propuesta.

Sistema de recomendación

El sistema de recomendación se desarrollará en función de dos estrategias. La primera busca identificar el cliente y sus hábitos de consumo para sugerir, según este historial, un conjunto de diez productos que se asemejen a lo que un cliente consume, pero que nunca ha consumido. La segunda está orientada a sugerir un conjunto de diez productos, pero en función del producto seleccionado en el instante por el consumidor. Dicho esto, se

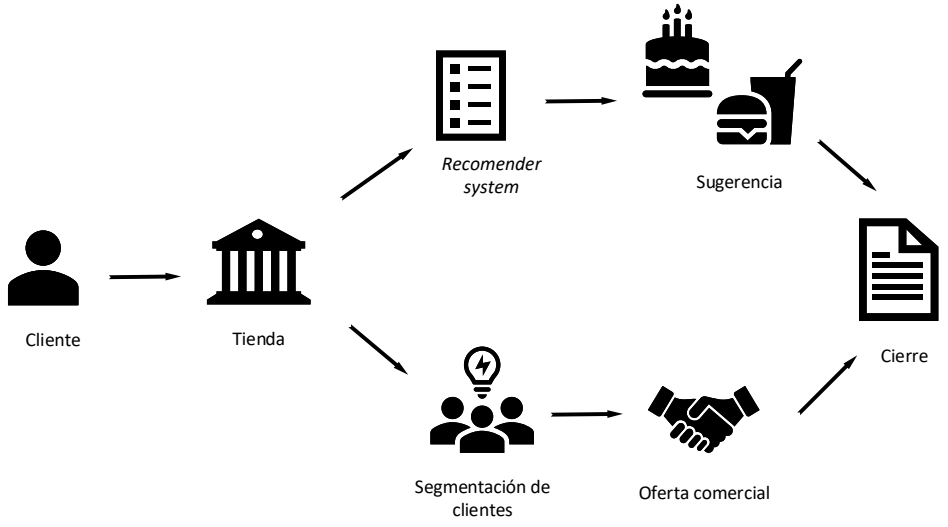
concluye que para la primera estrategia se requiere del conocimiento del cliente, de sus hábitos de compra y de un entendimiento de la frecuencia de compra de los productos. Mientras que la segunda estrategia solamente necesita que un cliente, así sea nuevo, seleccione un producto para poder sugerir la nueva compra.

De forma complementaria, se ata a la estrategia comercial la segmentación de los clientes a través de un clúster. Así el sistema podrá identificar la categoría a la que pertenece el cliente cuando esté haciendo el ofrecimiento de productos, a través de cualquiera de las dos estrategias anteriores, y se podrá hacer un ofrecimiento atractivo con el fin de que se materialice la sugerencia hecha por el sistema de recomendación. A continuación, se presentan los respectivos flujos para cada una de las estrategias propuestas.

En el flujo 1, el cliente ingresa a la tienda, el sistema de recomendación identifica sus hábitos de consumo y, con base en ese historial, ofrece nuevos productos. Asimismo, según la segmentación del cliente, se realiza una oferta comercial esperando incrementar la tasa de conversión (productos sugeridos/productos comprados).

Figura 5

Flujo 1: cliente habitual llega a la tienda



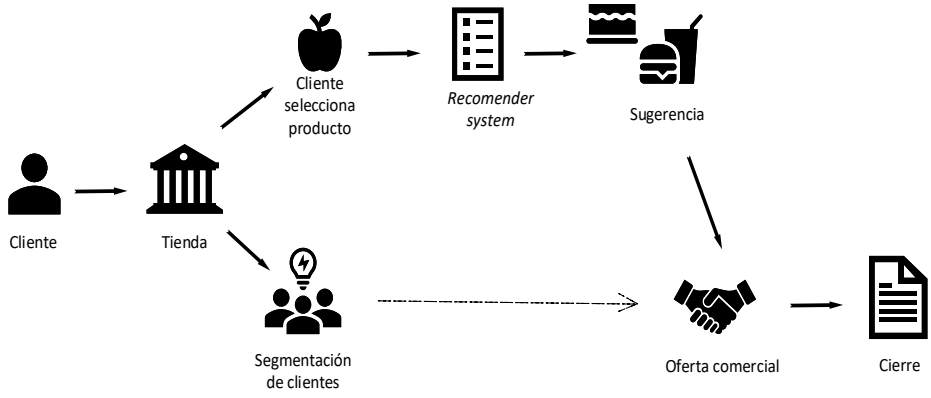
Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 36), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

En el flujo 2, el cliente no habitual ingresa a la tienda, selecciona uno de los productos disponibles, el sistema de recomendación sugiere, según afinidad al producto escogido, los nuevos productos. Adicionalmente, si el cliente cuenta con atributos suficientes para

ser segmentado, se realiza una oferta comercial con el objetivo de incrementar la tasa de conversión (productos sugeridos/productos comprados).

Figura 6

Flujo 2: cliente no habitual llega a la tienda



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 37), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Para la construcción del modelo de recomendación, se utilizará la librería LightFM de Python optimizada para la creación de este tipo de sistemas, así como las tablas de productos y *tickets*.

Aceptación y puesta en producción

Entre los detalles del producto de datos que se entregará al área usuaria, se presenta las siguientes consideraciones:

- Se debe iterar y entregar lo más pronto posible al área usuaria a fin de conseguir retroalimentación temprana.
- El área usuaria decide qué es importante y qué no lo es.
- Con la aceptación de los modelos, se procede a su operacionalización considerando que se debe desplegar en la forma como serán utilizados por el área usuaria: a través de una interfaz API (*application programming interface* o interfaz de programación de aplicaciones). Esta es un conjunto de definiciones y protocolos que permiten la comunicación entre dos aplicaciones de *software* mediante un conjunto de reglas.
- Con retroalimentación del área usuaria, se define un producto mínimo viable que puede darse a través de aplicaciones web, tableros de control o

aplicaciones corporativas, y que contenga las funcionalidades requeridas por el área usuaria.

Igualmente, es recomendable considerar las siguientes perspectivas, como componentes importantes de la mejora continua para futuros proyectos:

- Perspectiva de la alta dirección
  - Presentar los resultados del proyecto, que sean tangibles y alineados con los objetivos estratégicos de la organización.
  - Asegurar el presupuesto necesario para el mantenimiento y operación por el ciclo de vida del proyecto a través de un plan de negocio debidamente sustentado.
- Perspectiva del área usuaria
  - Es nuestra razón de ser, es importante recibir su retroalimentación a fin de mejorar.
  - Detectar de forma conjunta nuevas oportunidades de negocio y repetir el ciclo.
  - Dar soporte y capacitación a los usuarios.
  - Revisar posibles mejoras luego de tres meses de uso.
- Perspectiva de datos
  - Actualizar los datos con regularidad (mensual, trimestral).
  - Identificar posibles cambios en las tendencias.
  - Monitorear redes sociales a fin de seguir enriqueciendo los datos institucionales.
  - Recomendar al área de finanzas que revise regularmente las variables económicas que puedan tener impacto en el modelo: cambio de tarifas, alteraciones en el tipo de cambio, cambios regulatorios.
- Perspectiva de la arquitectura de *big data*
  - Establecer políticas de ciberseguridad.
  - Monitorear la escalabilidad de la arquitectura.
  - Optimizar costos de iteración.
  - Prever picos de demanda, sobre todo en campañas.
  - Revisar qué se puede mejorar de los procesos ya realizados.

  - Definir qué nuevas prácticas de la industria pueden incorporarse.
  - Revisar qué componentes de la arquitectura merecen actualizarse, mejorarse o descartarse según la política de AWS.
- Perspectiva del modelo
  - Evaluar constantemente, sobre todo al inicio, la *performance* de los modelos a fin de buscar refinamiento.
  - Medir si los resultados obtenidos por los modelos se acercan a lo esperado en los resultados del negocio.

RESULTADOS

Primer análisis exploratorio de los datos

Para poder efectuar esta tarea con la muestra inicial de datos, se desarrolló un código en lenguaje de programación Python con la finalidad de realizar el análisis descriptivo de este primer juego de datos. A continuación, se hace una breve descripción de la muestra de datos utilizada.

Tabla 1

Fuentes de datos internos para el primer análisis exploratorio

Nombre de la tabla	Peso	Formato	Número de registros	Número de campos
Products	3,360 KB	JSON	7 917	17
Stores	84 KB	JSON	1 127	4
Customers_deliver1	24 KB	JSON	168	7
Tickets_deliver1	9,962 KB	JSON	40 693	11

Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 17), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

La idea de este primer análisis descriptivo de los datos es identificar registros duplicados, registros vacíos y valores atípicos, así como empezar a entender la estructura de datos en cada archivo, en donde podemos resaltar, entre otras, la siguiente información:

- Se encontró dos valores atípicos en la base de datos Products, totalizando 222 ocurrencias.
- El 94,2 % de los productos no requieren pesaje.
- Solo el archivo *stores* presentó cuatro registros duplicados.

- La mayor cantidad de tiendas se ubica en ciudades principales. La ciudad con mayor representación de tiendas es Madrid con 71 (12,86 %), le sigue Barcelona con 39 (7,06 %) y Valencia con 19 (3,44 %). Al parecer, la cantidad de locales está muy ligada al índice poblacional de España.
- La primera tienda se inauguró hace cinco años y la última hace un año.
- Considerando los registros, se observa una fuerte concentración de aperturas en el segundo semestre del 2018 y el primer semestre del 2019.
- Las personas que son clientes se encuentran en un rango de 20 a 76 años, con una media de 50 años; la mayor concentración se presenta en las personas de 50 y 60 años.
- Si bien el campo género tiene vacíos, la mayoría de los clientes son mujeres, que representan el 75,2 %, mientras que los hombres solo llegan al 24,8 %.
- El primer cliente registrado fue hace cinco años y el último fue hace un año.
- Los clientes están registrados solo en diez tiendas de todas las que se encuetran en el país. Por otro lado, el registro de los clientes se ha mantenido constante en el tiempo, pero se evidencian tres momentos claves en que el registro fue superior al promedio: diciembre del 2015 a febrero del 2016, julio del 2017 a agosto del 2017 y febrero del 2018 a marzo del 2018. El último periodo fue el más representativo de los registros.

Asimismo, de acuerdo con lo analizado, la relación entre las bases de datos se aprecia en la tabla 2.

Tabla 2

Relaciones entre bases de datos

Nombre de la tabla	Campo de cruce	Tabla de relación
Customer	Customerid	*Tickets*, *stores*
Stores	Storeid	*Tickets*
Products	Productid	*Tickets*
Tickets	Ticketid	*Products*, *stores*, *customer*

Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 20), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

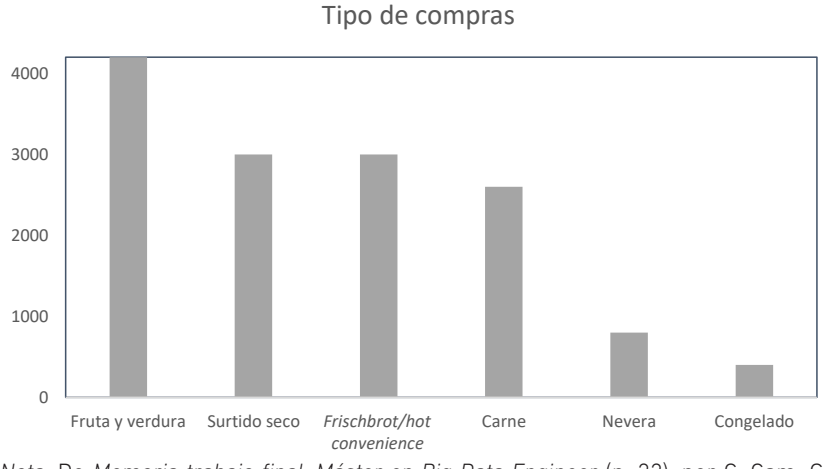
- Cuando se relacionan los *tickets* con los productos, se evidencia que todos los productos que se venden no se encuentran relacionados en la tabla de productos. Solo se tienen 14 263 registros de *tickets*, lo cual equivale al 35,05 % de la

información. La cantidad de productos que podemos identificar es solo el 44 % de los existentes en la tabla de productos.

- La relación de las tablas de *tickets* y *customers* es perfecta y no tenemos pérdida de información.
- La relación de las tablas de *tickets* y *stores* es perfecta y no tenemos pérdida de información.
- Las compras realizadas están ligadas a las tiendas de las ciudades de Zaragoza, Santander y Pamplona, las cuales no son las ciudades donde existe la mayor cantidad de tiendas que tiene la compañía. Adicionalmente, las compras no corresponden a las tiendas recientemente abiertas.
- Los productos que más se consumen según el tipo de mercado se muestran en la figura 7.

Figura 7

Productos más consumidos por tipo de mercado

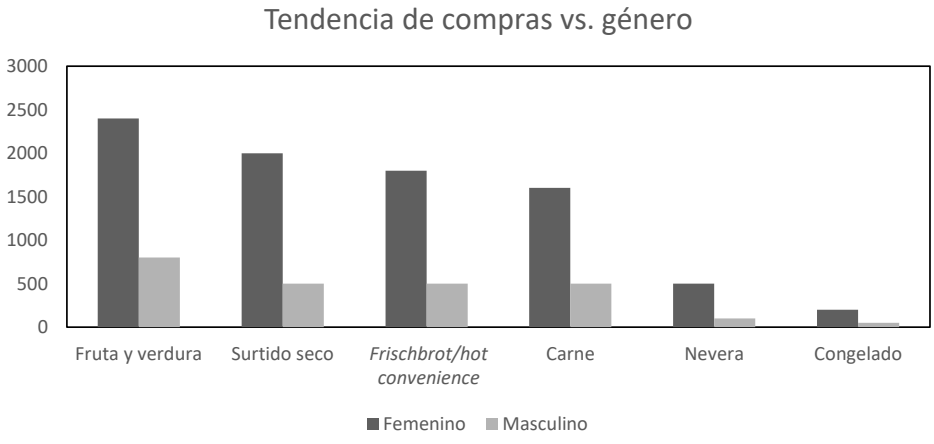


Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 22), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

- Analizando quién compra más, encontramos que el 78,2 % de los compradores son mujeres y el 21,8 % son varones.
- Analizando la relación de lo que más se compra con respecto al género, la tendencia entre hombres y mujeres se observa en la figura 8.

Figura 8

Productos más consumidos por género



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 22), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Segundo análisis exploratorio de los datos

En este caso, se utiliza un conjunto de datos mayor que la muestra: los incrementos se han producido en los archivos de clientes y *tickets*, mientras que las demás tablas permanecen iguales; de ahí la necesidad de construir las arquitecturas de *big data* en la nube. En la tabla 3, se describe cada una de las tablas utilizadas.

Tabla 3

Fuentes de datos internos para el segundo análisis exploratorio

Nombre de tabla	Peso	Formato	Número de archivos	Número de registros	Número de campos
Products	3,360 KB	JSON	1	7 917	17
Stores	84 KB	JSON	1	1 127	4
Customers	7,299 KB	JSON	1	49 998	8
Tickets	3,13 GB	JSON	20	13 238 241	11

Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 17), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Para llevar a cabo este segundo análisis descriptivo, se vuelve a utilizar el mismo código de lenguaje de programación en Python (Singh, 2019) del primer análisis. Los principales hallazgos en los archivos *customers* y *tickets* son los siguientes:

- La base de datos de clientes contiene 547 registros duplicados.
- Los clientes están en un rango de 17 a 119 años; podemos considerar edades atípicas a partir de los 80 años en adelante. La media de edad de las personas está sobre los 45 años.
- El campo género de clientes tiene las dos categorías y presenta una mayor concentración en las mujeres, las cuales constituyen el 66,3 % y los hombres el 33,6 %. Existen valores que están fuera de ambas categorías, los cuales se tendrán que eliminar.
- La variable *store* contiene 551 comercios.
- El primer cliente registrado fue hace cinco años y el último fue hace un año. Podemos notar que existe una mayor concentración de vinculaciones entre el 2018 y el 2019. Estas vinculaciones representan el 92 % del total.
- La base de datos *tickets* cuenta con 117 122 registros duplicados y no contiene vacíos.
- El 88 % de las compras se realizaron con la tarjeta DigitalCard y el 11,5 % con la tarjeta Mobile.
- Las compras se realizaron solamente en 557 tiendas de las 1127 que en total tiene la marca.
- Las compras registradas corresponden a cuatro meses de actividad de las tiendas.

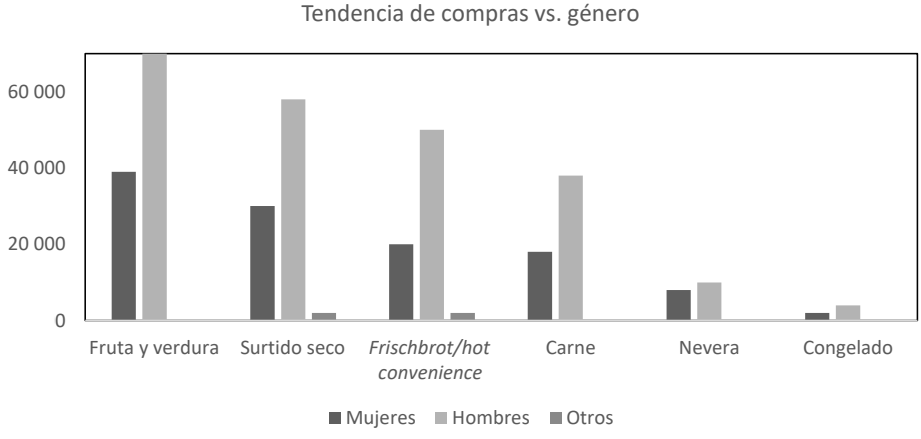
Asimismo, se debe mencionar que en este segundo análisis descriptivo se mantienen las relaciones entre las bases de datos, tal como se indicó en el primer análisis (véase la tabla 2). Entonces, al realizar el análisis correspondiente de acuerdo con estas relaciones, se observa lo siguiente:

- Cuando se relacionan los *tickets* con los productos, se halla que todos los productos que se venden no se encuentran relacionados en la tabla de productos. Solo hay 4 597 204 registros de *tickets*, lo cual equivale al 34,72 % de la información. La cantidad de productos que se pueden identificar es de solo el 27,28 % de los existentes en la tabla de productos.
- El cruce de las tablas de *tickets* y de clientes solo permite determinar 1 083 325 registros, lo cual equivale al 8,18 % de la información. Ahora bien, este cruce representa solo 3800 clientes únicos, que equivalen al 7,6 % de los clientes de la tabla de *customers*.
- La relación de las tablas de *tickets* y *stores* es perfecta y no existe pérdida de información.

- Las compras realizadas están ligadas a las tiendas en las ciudades de Madrid, Zaragoza, Sevilla y Barcelona, donde la marca tiene mayor presencia de tiendas.
- De la base de datos de productos, podemos observar que las categorías que más se consumen son fruta y verdura, surtido seco, *frischbrot*/*hot convenience*, carne, nevera, congelado.
- Analizando quién compra más, puede verse que la distribución por género es de 58,5 % mujeres, 30,1 % hombres y 11,3 % con género no clasificado.
- Analizando la relación de lo que más se compra con respecto al género, la tendencia entre hombres y mujeres se observa en la figura 9.

Figura 9

Productos más consumidos por género

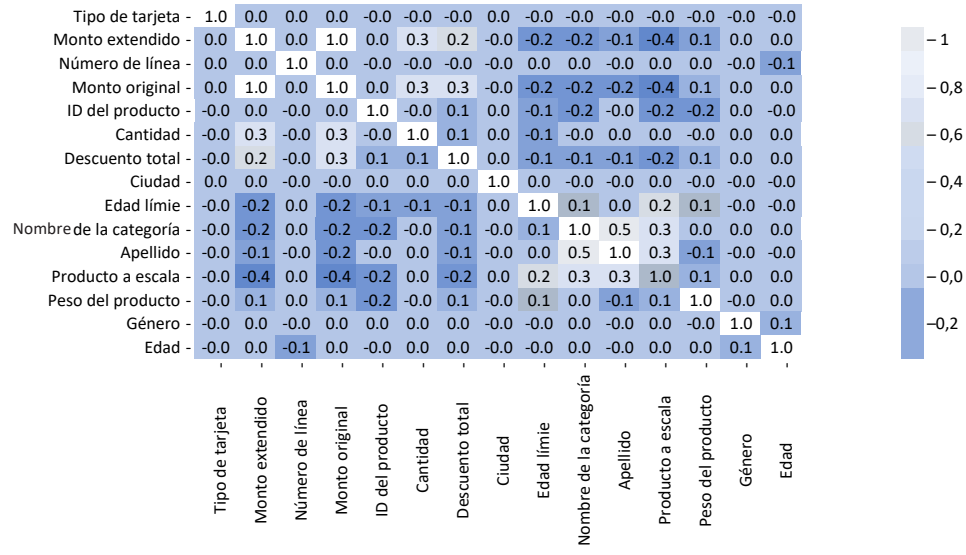


Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 22), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Se observa, al revisar la correlación entre las variables continuas, que las variables monto original y monto extendido (*original amount* y *extended amount*) de la tabla de *tickets* tienen una alta correlación.

Figura 10

Correlación entre variables



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 24), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Primera segmentación

A fin de facilitar esta primera aproximación a la segmentación, se excluyeron los registros con valores nulos en la columna edad. Se llevaron a cabo tres iteraciones. En la primera, con todas las variables, se obtuvo superposición de los puntos de varios clústeres, pero no se logró una agrupación limpia; por lo tanto, se desechó esta opción. Para la segunda iteración, se decidió eliminar la edad, pues excluye registros y no contribuye a la clusterización. En este caso, se observa que la superposición ha mejorado; sin embargo, se obtiene una agrupación que no es muy limpia y se puede ver que las variables sintéticas de los grupos de mercancías tampoco aportan al agrupamiento; por lo tanto, también se desecha esta opción. En la tercera iteración, se elimina la variable edad y las variables sintéticas de los grupos de mercancías. En esta oportunidad sí se obtienen seis clústeres claramente definidos con baja superposición (solo en dos grupos) y con los centroides bien ubicados. Por lo tanto, se considera a la tercera opción como el mejor resultado de la segmentación (Cam et al., 2020).

Tabla 4

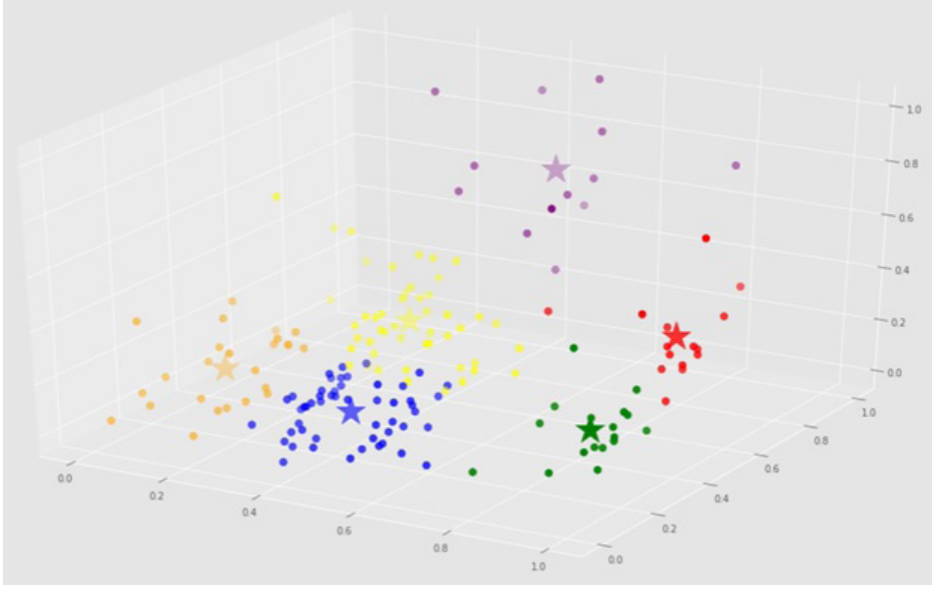
Resultados de la primera segmentación

Clúster	Color	Cantidad	Antigüedad en días	Promedio de visitas al mes	Gasto promedio por mes	Descuento promedio por mes
0	Amarillo	48	601,6	74,4	119,2	3,7
1	Púrpura	13	725,0	135,0	246,0	5,0
2	Azul	51	643,6	32,0	58,1	1,3
3	Naranja	24	245,4	40,2	71,7	2,0
4	Rojo	15	1 295,1	84,2	138,9	4,7
5	Verde	17	1 271,5	38,8	76,0	1,7

Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 18), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Figura 11

Representación gráfica de la primera segmentación



Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 28), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Segunda segmentación

La revisión posterior de los tuits extraídos en tiempo real muestra que no existe información relevante para mejorar la segmentación de clientes y que el flujo de estos es muy bajo. Dada esta situación, se decide capturar los tuits antiguos (es decir, los que figuran "escritos" en la cuenta oficial, ya no en tiempo real). El resultado es igualmente deficiente: no hay información que ayude a mejorar significativamente el modelo de segmentación. Asimismo, se debe mencionar que no se pudo relacionar el código de cliente con el usuario de Twitter.

Ante los resultados obtenidos, se consideran dos alternativas adicionales de fuentes externas. La primera fue la exploración sobre la información nutricional, pero esto no fue posible porque la información nutricional (Nutriscore) no era de directa aplicación sobre la tabla de productos, por lo que esta alternativa se descartó. La segunda alternativa fue utilizar la renta media por hogar por comunidad autónoma; para tal efecto, se consiguió esta información oficial a través del portal web del Instituto Nacional de Estadística (Cam et al., 2020).

A continuación, se mezcla esta información de renta media con los datos del archivo stores.csv, a fin de que cada *store id* tenga asociada su respectiva renta media por comunidad. Para lograr eso, se tuvo que completar la comunidad autónoma por ciudad/distrito del *store id*.

Con este nuevo archivo de *stores* y usando el *store id* como elemento de conexión, agregamos a cada código de cliente el ratio de compra/renta, que se obtiene al dividir la compra mensual entre la renta media mensual de la comunidad. Esta nueva columna compra/renta se usará como nueva fuente externa a fin de mejorar la segmentación. Los resultados de la segunda segmentación se muestran en la tabla 5.

Tabla 5

Resultados de la segunda segmentación

Clúster	Color	Cantidad	Antigüedad (días)	Promedio de visitas por mes	Gasto promedio por mes	Descuento promedio por mes	Ratio gasto/ renta
0	Rojo	367	627,77	77,34	135,68	3,19	4 %
1	Verde	336	151,16	144,6	281,82	10,37	1 %
2	Azul	1796	131,92	51,3	99,48	2,51	3 %
3	Amarillo	36	1 254,16	71,48	120,33	4,25	4 %
4	Morado	639	297,7	71,69	124,74	2,51	5 %

Nota. De *Memoria trabajo final. Máster en Big Data Engineer*, por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

Con base en estos hallazgos, proponemos los siguientes segmentos de clientes:

- *Clientes tradicionales (segmento 0 - rojo)*. El segmento 0 representa el 12 % del total de clientes, con una antigüedad media de 627 días, 9 visitas al mes en promedio (más que la media general) y el segundo promedio más alto de gastos de los segmentos presentados. El descuento proporcionado a este segmento se encuentra muy cercano al promedio de descuento general.
- *Clientes potenciales (segmento 1 - verde)*. Estos clientes (el 11 % de la base) se caracterizan por ser los segundos más jóvenes en términos de vinculación con la empresa (promedio de 151,16 días), pero con el más alto consumo en comparación con el resto de segmentos (281,8 euros por mes). Para favorecer su crecimiento, se encuentra que es el segmento con el mayor descuento ofrecido por la compañía.
- *Clientes masivos (segmento 2 - azul)*. Este segmento representa el 57 % de la base de clientes. Son clientes con antigüedad promedio de 132 días, su frecuencia de visita es la más baja (51,3 en promedio al mes), al igual que el consumo que realizan en las tiendas, donde pagan 99 euros por mes en promedio. Dadas sus características, son también los clientes con el menor porcentaje de descuentos aplicados.
- *Clientes fieles/antiguos (segmento 3 - amarillo)*. Corresponden al 1 % de la base y se caracterizan por haberse vinculado en promedio hace 3,4 años; además, las visitas por mes son un poco más elevadas que la media general. Tienen la segunda mayor tasa de descuento, atendiendo de esta forma su fidelidad.
- *Clientes recientes (segmento 4 - morado)*. Este segmento representa el 20 % de la base de clientes y ronda el año desde su vinculación. Son clientes con consumo, visitas y descuentos cercanos al promedio.

Esta segunda segmentación, caracterizada por los hábitos de consumo, tiene como objetivo principal clasificar a los clientes por la gama de productos que consumen, con la finalidad de ejecutar el sistema de recomendación por cada segmento. Así pues, la recomendación estaría atada a la afinidad de consumo del segmento propio sin interactuar con los segmentos adyacentes (Cam et al., 2020).

Sistema de recomendación

Tal como se indicó anteriormente, se definieron dos estrategias. En la primera, el sistema identifica los hábitos de consumo de un cliente y, de acuerdo con ellos, realiza la recomendación. En este caso, se tomó un cliente al azar y se encontró que los diez productos más consumidos y los productos sugeridos fueron los que se observan en la tabla 6.

Tabla 6

Resultados de la primera estrategia del sistema de recomendación

Consumo habitual	Productos sugeridos
Aspil Jumpers Mantequilla	Kinder Joy
Lacasitos Toy	Barra de pan
Burger de pavo/pollo con espinacas	Plátano canario FP
Barra de picos	Escalopines de lomo de cerdo adobado
Plátano canario 800 g	Lacasa Paraguas Chocolate surtido
Albóndigas de ave	Barra gallega
Floopy azúcar	*Floopy* bombón
Pulguita	Puerro
Pan de la abuela	Barra *premium* con masa madre
Kinder Sorpresa	Ninguna
Croissant margarina	Ninguna

Nota. De *Memoria trabajo final. Máster en Big Data Engineer* (p. 39), por C. Cam, G. Hidalgo, C. Huérfano y J. Medina, 2020, Universidad de Barcelona.

En la segunda estrategia, el consumidor escoge un producto y se le recomiendan diez productos afines. Se tomó como ejemplo el código de producto 82620, que corresponde al pimiento rojo, y se obtuvieron las siguientes sugerencias:

- Cuétara Mini Campurrianas
- Berenjena
- Cebolla 2 kg
- Pimiento rojo granel
- Casa Macán Queso, barra gallega
- Sanase Color n.° 7.77, rubio almendra
- Col picuda
- Brillante Tripack Integquinoa 2+1
- Cebolla 750 g

Al medir el rendimiento del modelo, se obtuvo una precisión del 90,88 %. La precisión mide la proporción de elementos positivos entre los *k* elementos mejor clasificados; en este caso, *k* corresponde a los diez productos recomendados. Como tal, la precisión está centrada en la calidad de la clasificación en la parte superior de la lista, sin considerar qué tan buena o mala sea el resto de su clasificación, siempre que los primeros

k elementos sean en su mayoría positivos. Esta métrica es adecuada si solo se va a mostrar a los usuarios la parte superior de la lista (Witten, 2017; Falk, 2019).

Aquí es importante recordar, que, según el segundo análisis descriptivo, solo se ha podido identificar el 27,28 % de los productos que figuran en la tabla de *tickets*, lo cual refuerza la idea acerca de la relevancia de la integridad y la consistencia de las tablas que intervienen en un proyecto de *big data*.

CONCLUSIONES

Inicialmente se tenía el objetivo de segmentar clientes según su consumo de productos; sin embargo, esto no fue posible porque era necesario aplicar varias variables sintéticas para completar el análisis y, finalmente, estas no fueron determinantes al momento de clasificar. El sentido común indicaba que debería existir esa segmentación cruzada, pero la realidad de los datos mostró lo contrario; la antigüedad fue la variable dominante en la formación de clústeres.

Esta dificultad permite resaltar la importancia de los análisis descriptivos de datos llevados a cabo antes de la primera y de la segunda segmentación, porque estos dan lugar al proceso de entender las características de los datos no solo desde la mirada estadística, sino, sobre todo, desde el punto de vista del negocio. Asimismo, cabe mencionar que la integridad de las bases de datos es de gran importancia, pues aporta sustancialmente al análisis exploratorio de los datos, así como al desarrollo en la etapa de modelado.

Es de suma utilidad el uso de una metodología de proyectos basada en agilidad, pues aporta flexibilidad y productividad en un entorno de incertidumbre y requisitos cambiantes, en donde hay que conjugar la experimentación con la entrega de resultados que tengan un impacto en el negocio. El trabajo en equipo en los proyectos de *big data* es fundamental para alcanzar los resultados.

Indudablemente, emplear la infraestructura adecuada para resolver problemas de *big data* no solo permite tener tiempos de respuesta ideales para la exploración, construcción y análisis de los modelos estadísticos, sino también enriquecer, por medio de fuentes alternativas de datos, todo el trabajo realizado con las fuentes de información tradicionales. No siempre las fuentes externas pueden resultar de utilidad; es necesario explorar fuentes alternativas de valor a fin de mejorar la fase de modelado. Debido a la estructura de las bases de datos, fue muy complicado encontrar una relación, por la baja frecuencia de tuits y su nula conexión con las variables con las que trabajamos.

En la misma línea de pensamiento, cabe resaltar la relevancia de los sistemas de recomendación dentro del mundo del consumo masivo. Su explotación y uso se produce en sectores tan diversos como películas, videos, música, libros, hoteles, restaurantes, etcétera, y se han convertido en una estrategia comercial muy potente y presente en el mundo actual.

REFERENCIAS

Aryuni, M., Didik Madyatmadja, E., & Miranda, E. (2018). Customer segmentation in XYZ Bank using k-means and k-medoids clustering. En *Proceedings of 2018 International Conference on Information Management and Technology, ICIMTech 2018* (pp. 412-416). https://doi.org/10.1109/ICIMTech.2018.8528086

Cam, C., Hidalgo, G., Huérfano, C., & Medina, J. (2020). *Memoria trabajo final. Máster en Big Data Engineer*. Universidad de Barcelona.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: a case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197-208.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, *36*(4), 1165-1188. https://doi.org/10.2307/41703503

Christodoulou, P., Christodoulou, K., & Andreou, A. S. (2017). A real-time targeted recommender system for supermarkets. En *Proceedings of the 19th International Conference on Enterprise Information Systems. Volumen 2: ICEIS 2017* (pp. 703-712). https://doi.org/10.5220/0006309907030712

Doğan, O., Ayçin, E., & Bulut, Z. A. (2018). Customer segmentation by using RFM model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*, *8*(1), 1-19.

Falk, K. (2019). *Practical recommender systems*. Manning.

Fang, Y., Xiao, X., Wang, X., & Lan, H. (2018). Customized bundle recommendation by association rules of product categories for online supermarkets. En *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)* (pp. 472-475). https://doi.org/10.1109/DSC.2018.00076

Gulabani, S. (2017). *Practical Amazon EC2, SQS, Kinesis, and S3: A hands-on to AWS*. Apress. https://doi.org/10.1007/978-1-4842-2841-8

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer segmentation using k-means clustering. En *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018* (pp. 135-139). https://doi.org/10.1109/CTEMS.2018.8769171

Kumar, V., & Reinartz, W. (2018). *Customer relationship management: concept, strategy, and tools* (3.ª ed.). Springer. https://doi.org/10.1108/IJBM-11-2014-0160

Lycett, M. (2013). "Datafication": making sense of (big) data in a complex world. *European Journal of Information Systems*, *22*(4), 381-386. https://doi.org/10.1057/ejis.2013.10

Microsoft. (2021, 11 de diciembre). *What is team data science process?* https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

Pascal, C., Ozuomba, S., & Kalu, C. (2015). Application of k-means algorithm for efficient customer segmentation: a strategy for targeted customer services. *International Journal of Advanced Research in Artificial Intelligence*, *4*(10), 40-44. https://doi.org/10.14569/ijarai.2015.041007

Pérez, C. (2013). *Análisis multivariante de datos. Aplicaciones con IBM SPSS, SAS y STATGRAPHICS* (1.ª ed.). Garceta.

Schermann, M., Hemsen, H., Buchmüller, C., Bitter, T., Krcmar, H., Markl, V., & Hoeren, T. (2014). An interdisciplinary opportunity for information systems research. *Business and Information Systems Engineering*, *6*(5), 261-266. https://doi.org/10.1007/s12599-014-0345-1

Singh, P. (2019). *Machine learning with PySpark*. Apress. https://doi.org/10.1007/978-1-4842-4131-8

Witten, I. H., Eibe, F., & Hall, M. A. (2017). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann.