3.1 *Big data*

En los últimos años se ha hablado mucho de ***big data***. Si bien es un término bastante amplio que hace referencia a grandes volúmenes de datos, en analítica digital ha cobrado una gran relevancia.

Esto se debe, sobre todo, a que las empresas deben modificar sus equipos, formas de trabajo, recursos y herramientas para adecuarse a un mundo en el que ya no se trabaja con datos planos y lineales, sino con enormes cantidades de datos, estructurados y no estructurados (imágenes, videos, audios). Esto les supone un desafío al momento de recolectarlos, explorarlos, analizarlos y utilizarlos en beneficio de los objetivos de negocio.

Cuando se hace mención a *big data*, o a grandes volúmenes de datos, se tienen en cuenta 4 aspectos que la diferencian de la *small data* (pequeñas cantidades de datos):

- Volumen: tamaño de los datos.
- Velocidad: rapidez con la que esos datos se modifican.
- Veracidad: incertidumbre acerca de lo que esos datos contienen.
- Variedad: cantidad de fuentes de datos y formas distintas que puede tener la *data*.

Existen 3 desafíos que nacen junto al *big data*:

- **Almacenamiento**: es necesario contar con herramientas o plataformas que almacenen gran cantidad de datos.
- **Diversidad**: capacidad de poder almacenar en un mismo repositorio tanto los datos estructurados como los no estructurados.
- **Procesamiento**: utilizar las técnicas adecuadas para procesar estos grandes volúmenes de datos.

Hay, además, tres roles que juegan un papel fundamental:

- **Ingeniero**: será el responsable de utilizar las herramientas adecuadas para facilitar el acceso a los datos.

- **Consultor de negocio**: serán los responsables de comprender el problema de negocio y definir los pasos para resolverlo.
- **Científico**: definirá el modelo más idóneo para resolver el problema.

3.1.1 *Data mining*

Data mining (o minería de datos, en español) incluye el proceso y las técnicas que permiten explorar grandes volúmenes de datos (*big data*) con el objetivo de encontrar patrones comunes de comportamiento.

Se utilizan técnicas de *data mining* cuando no es posible explorar los datos de forma manual debido a su cantidad, volumen o estructura. Se requieren, entonces, técnicas automatizadas o semiautomatizadas que permitan descubrir, extraer y almacenar esos datos para que luego los analistas puedan sacar conclusiones y tomar decisiones.

La minería de datos permite buscar e identificar información valiosa dentro de grandes volúmenes de datos.

Para la exploración de los datos se utilizan técnicas de estadística o algoritmos capaces de identificar estos patrones. Por eso, quienes trabajan con grandes volúmenes de datos deben tener conocimientos avanzados en estas áreas: estadística, matemática o, incluso, inteligencia artificial.

Las fases dentro del *data mining* comienzan con los datos en bruto, y finalizan con la interpretación y toma de decisiones. A nivel general, se pueden identificar cuatro etapas principales:

1) **Lectura y limpieza de los datos (*cleansing*):** en esta etapa se procede a filtrar todos aquellos datos inválidos, incorrectos, duplicados, desconocidos, que pueden ensuciar una base de datos y afectar los resultados.

2) **Selección de variables:** según cuáles sean los objetivos del análisis, se seleccionarán algunas variables y se descartarán otras. Esto permitirá reducir la cantidad de datos y aumentar la velocidad de procesamiento, ya que no todas las variables tendrán la misma relevancia. Ejemplo: si tenemos la base de datos del registro automotor del país, pero nuestro análisis se centra en la posesión de autos antiguos, se seleccionará como variable el año de patentamiento, pero no la variable que define el color del auto.

3) **Extracción de datos**: con el empleo de distintas técnicas se pueden obtener patrones de comportamiento basados en las variables seleccionadas o relaciones de asociación entre distintas variables. Existen distintas técnicas y lenguajes comúnmente usados para la extracción de datos. Algunos de ellos son:
    a. **Redes neuronales**: están inspirados en la manera en que funciona el sistema nervioso.
    b. **Regresión lineal**: hace referencia a la relación entre dos variables.
    c. **Árboles de decisión**: son modelos predictivos que utilizan inteligencia artificial y emplean relaciones lógicas basados en reglas.
    d. **Modelos estadísticos**: sirven para indicar si existe o no una relación entre dos variables.
    e. ***Clustering***: agrupa distintos valores según su cercanía o distancia respecto a una característica específica.

4) **Interpretación y evaluación:** en esta etapa, el análisis se centra en la coherencia de los datos en base a los resultados obtenidos. El modelo debe ser validado antes de tomar decisiones.

3.1.2 *Datawarehouse*: arquitectura de la información

Cuando hablamos de *big data* y de grandes volúmenes de datos, no podemos dejar de mencionar el término ***datawarehouse***. Como

su nombre en inglés lo indica, hace referencia a un almacén de datos, un repositorio.

En informática, en estadística y en *data science*, estos repositorios de datos se caracterizan por tener tres cualidades que los diferencian de cualquier otro. El *datawarehouse* debe ser:

- **Integrado**: permite integrar en un mismo lugar los datos de muchas fuentes de información distintas.
- **No volátil**: la información debe ser permanente, es decir, no debe sufrir modificaciones.
- **Variable en el tiempo**: los datos siempre se encuentran actualizados y permiten realizar análisis históricos.

Los *datawarehouse* suelen contener datos de una gran cantidad de fuentes de información distintas. Esta información, a su vez, puede provenir de otras plataformas que, al mismo tiempo, ya han procesado los datos.

Nivel de agregación

La *data* que se incluye en el repositorio puede ser ***data* cruda** o ***data* agregada**. Cuando se encuentra de forma cruda, permite una mayor variedad de análisis. Por ejemplo, una base de datos tradicional puede constituir una fuente de *data* cruda, mientras que la información que proviene de Google Analytics ya se encuentra procesada por esa herramienta.

Origen de la *data*

Existen tres tipos de fuentes de información que se pueden incluir en un *datawarehouse*:

- ***First party data*:** es la *data* de fuentes propias. Por ejemplo, en una marca, corresponde a la información que poseen de sus propias bases de datos, redes sociales, sitio web o aplicaciones.
- ***Second party data*:** es la data obtenida de otras empresas con consentimiento explícito de los usuarios y mediante un acuerdo entre las partes. Por ejemplo, cuando una marca realiza un acuerdo con un banco, es probable que puedan compartir sus

bases de datos, siempre y cuando los usuarios lo hayan permitido.
- ***Third party data***: es la *data* proveniente de fuentes externas. En general, suele adquirirse a gran escala, lo que permite poseer bases de datos muy grandes, aunque poco detalladas.

Tipo de recolección

La forma de recolectar la *data* que se va a incluir en el *datawarehouse* es variada:

- **Declarativa**: es toda aquella *data* que fue proporcionada voluntariamente por un usuario. Un ejemplo muy común son los formularios presentes en sitios web en los cuales los usuarios completan datos y los envían aceptando términos y condiciones del uso de los mismos.
- **Comportamental**: surge de herramientas de análisis de comportamiento de usuarios. Por ejemplo, toda la información proporcionada por Google Analytics.
- **Transaccional**: proviene de diversas plataformas que brindan información transaccional de los usuarios. Esto incluye compras *online*, solicitudes *online*, etcétera.

Nivel de identificación

Los repositorios tienen información clasificada en distintos niveles. En lo que respeta al usuario, esta información puede ser:

- **Personal**: es decir, cada persona se encuentra identificada de manera única. Puede ser nombre, teléfono, geolocalización, foto, datos bancarios, etc.
  Se debe tener cuidado con el manejo de este tipo de información ya que existen varias reglamentaciones y leyes (especialmente en Europa) sobre el uso de la información personal, por considerarse datos sensibles y protegidos.
- **Seudonimizada**: corresponde a la identificación indirecta. Es decir, se pueden individualizar comportamientos, pero sin saber a quién pertenecen.
- **Anónima**: los datos son anónimos y no se pueden individualizar comportamientos.

Cloud vs on premise

Una de las primeras definiciones que debe tomar una empresa al comenzar a desarrollar un *datawarehouse* será dónde prefiere tener guardados sus datos. Existen dos opciones: guardarlos en la nube (*cloud*) y acceder a ellos a través de una conexión a internet, lo cual evita necesitar un *hardware* propio; o guardarlos en servidores locales de la empresa, con lo cual estarán disponibles solo mediante dispositivos locales.

Tabla 1. Ventajas y desventajas *on cloud* y *on premise*

ON CLOUD	Características	*ON PREMISE*
Menor	Tiempo de implementación	Mayor
Baja	Inversión por adelantado	Alta
No	Costos adicionales asociados	Si
Predecible	Costo total	Impredecible
Menor	Personalización	Mayor
Del proveedor	Estándar de seguridad	Propia

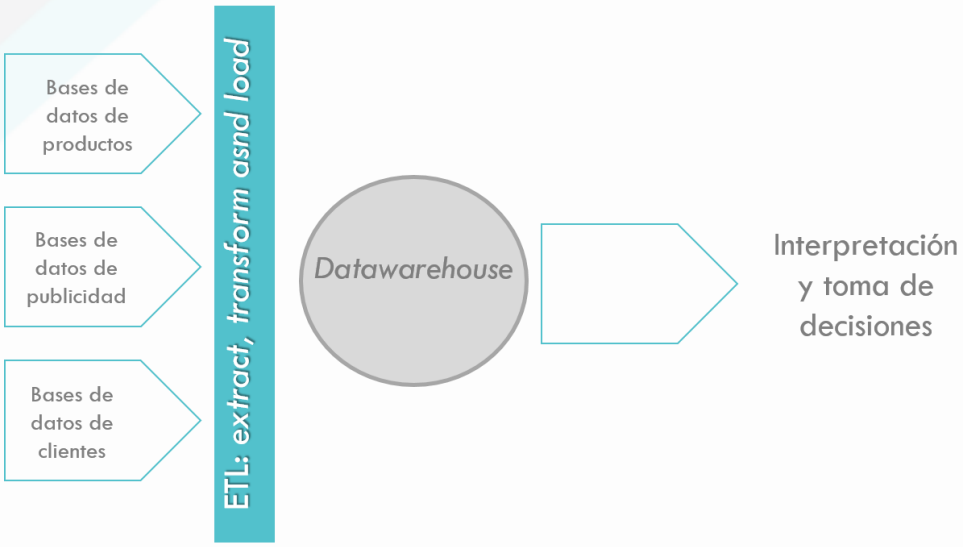
Fuente: elaboración propia.

Esquema de un *datawarehouse*

Al comenzar a estructurar un *datawarehouse* es necesario distinguir claramente dos áreas:

- **Operación**: aquí se define toda la *data* que recibirá el repositorio, las fuentes de datos y herramientas que se puedan conectar de forma automática.
- **Inteligencia de negocio**: son quienes harán uso de la información que se encuentra en el *datawarehouse* para la toma de decisiones.

Figura 1. Ejemplo de un esquema *datawarehouse*



Fuente: elaboración propia.

El ETL (*extract, transform and load*) es el proceso por el cual se extrae la información necesaria de las fuentes de información externas; se transforma para que se adecúe a una base de datos integrada y se carga en el repositorio para ser explotada por los analistas.

En resumen, un *datawarehouse* proporciona una herramienta clave para la toma de decisiones de cualquier área a partir de datos e información integrada y global de todo el negocio. Permite, además, encontrar relaciones entre distintos datos, variables y fuentes de información que aportan valor agregado a la toma de decisiones.

3.1.3 Herramientas y lenguajes esenciales

En el desarrollo de un *datawarehouse* se encuentran involucradas una gran variedad de herramientas y lenguajes de programación, necesarios para cada una de las etapas o necesidades. A continuación, detallaremos algunos ejemplos:

Amazon web services, Google Cloud Platform, Mixrosoft Azure

Estas plataformas son servicios integrales que se encuentran en la nube (es decir, se accede a ellas a través de internet), ofrecen la posibilidad de integrar inmensas infraestructuras de datos y las tienen disponibles para la explotación de información y decisiones de negocio.

Los precios varían de una plataforma a otra, pero, en general, los valores están asociados al volumen de almacenamiento que se brinda y a la cantidad de consultas que habilitan realizar a la base.

Hadoop y Spark

Son *frameworks* de computación de tipo *open source* (es decir, de código abierto) que permiten una lectura más simple, ágil e interpretable de los datos almacenados en un *datawarehouse*.

En analítica digital estas aplicaciones son muy usadas ya que permiten buscar en millones de páginas web resultados relevantes mediante algoritmos propios o desarrollados especialmente para un objetivo específico.

SQL

El **lenguaje de consulta estructurado** (*structural query language*) hace referencia es un tipo de lenguaje utilizado para realizar consultas en grandes bases de datos de manera ágil y sencilla.

Utiliza modelos el álgebra y cálculos relacionales para realizar las consultas y extraer información.

3.1.4 Visualización de la información

Uno de los desafíos más grandes frente a la manipulación de *big data*, es la manera en que se pueden sintetizar y presentar los datos de forma clara y entendible para cualquier analista.

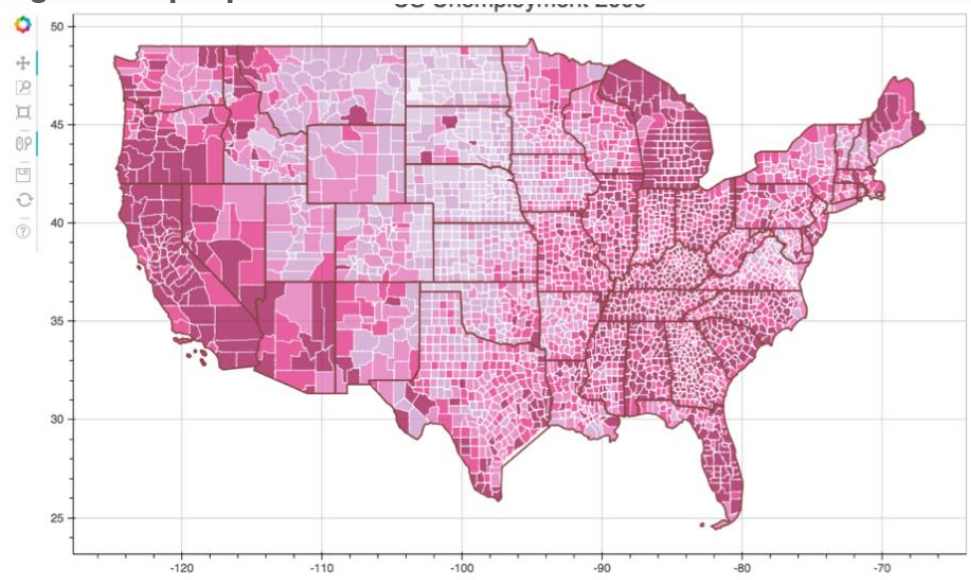
Python y R

Estos son los lenguajes más utilizados en la ciencia de datos. Los científicos de datos, estadísticos u otros analistas, hacen uso de estos lenguajes para poder traducir los datos en visualizaciones.

Son gratuitos, ambos permiten interactuar con herramientas y plataformas en la nube. Además, al ser de código abierto, se encuentran disponibles para manipular, modificar, corregir, agregar funciones, de acuerdo con las necesidades de cada empresa.

Gracias a las numerosas librerías que contienen, se pueden realizar gráficos muy variados como histogramas, diagramas de dispersión, mapas, etcétera.

Figura 2. Ejemplo de visualización realizada con R



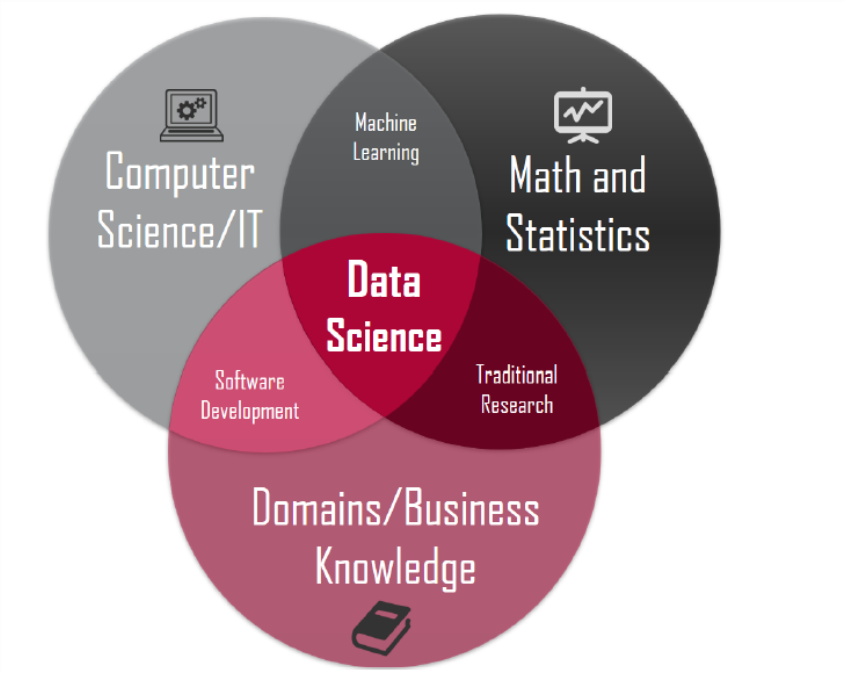
Fuente: BBVA API Market, 2015, https://bbvaopen4u.com/es/actualidad/cinco-librerias-en-python-para-cientificos-de-datos-como-visualizar-informacion.

3.2 *Data science*

Los *data scientist* han sido definidos por Thomas H. Davenport (2012) como "*the sexiest job of the 21st century*". La búsqueda de *data scientists* en grandes y medianas empresas ha crecido de forma exponencial en los últimos años.

La ciencia de datos hace referencia a un campo interdisciplinario que incluye los procesos y sistemas para extraer conocimiento de grandes volúmenes de datos. Decimos que es un área interdisciplinaria porque incluye ciencias de la computación, matemática, estadística, negocio.

Figura 3. ***Data science***



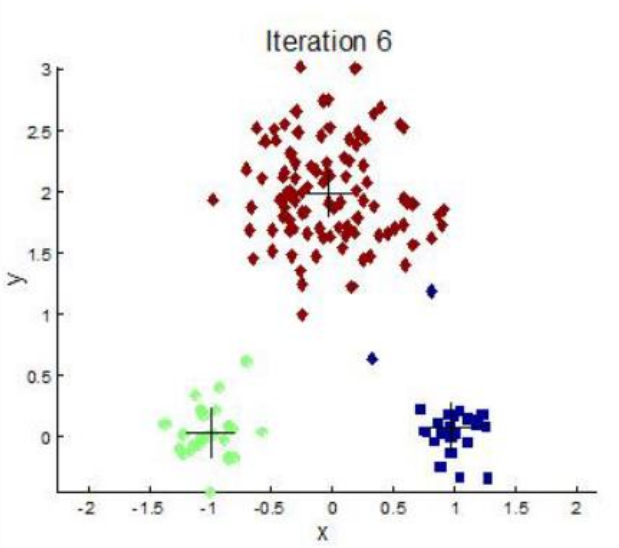
Fuente: Barber, 2018, https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

La minería de datos, la estadística y el *machine learning* son parte fundamental de la ciencia de datos.

La ciencia de datos sirve para:

- Predecir: cuando se posee información de entrada (X), pero no de salida (Y).
- Entender relaciones entre X e Y, y cómo una afecta a la otra.
- Comprender la existencia de patrones en los datos.

Figura 4. Patrones de comportamiento - *Clustering*



Fuente: elaboración propia.

Los científicos de datos suelen ser matemáticos, economistas, estadísticos u otras profesiones afines.

Sus objetivos principales son:

- Preparar la información para que pueda ser analizada.
- Explorar los datos para encontrar patrones.
- Ajustar el modelo para hallar la mejor manera de resolver un problema de negocio.

Big data se diferencia de *data science* porque, mientras el primero se centra en los problemas de almacenaje de grandes volúmenes de información; el segundo hace referencia a la manera de transformar los datos en información de valor.

3.2.1 Estadística y modelos

Como mencionamos, existen varias disciplinas que participan de la **ciencia de datos**. Una de ellas es la estadística.

La estadística hace referencia a la formalización de relaciones entre variables que se manifiesta mediante ecuaciones matemáticas.

Los modelos probabilísticos o estadísticos son ecuaciones matemáticas que se utilizan para indicar los distintos factores que modifican la variable de respuesta. Sirven para explicar la variabilidad de un fenómeno particular.

Por ejemplo, si se pretende realizar un análisis de la tasa de variación media de la población de una ciudad determinada, las variables que explican esas variaciones pueden estar relacionadas a su tamaño, perfil económico o clima.

Existen varios modelos estadísticos, aunque los más utilizados son:

- **Modelo de correlación**: es un índice que permite medir la fuerza de correlación entre dos variables.
- **Modelo de regresión lineal**: es la relación que existe entre una variable dependiente y una variable independiente.

Los **modelos interpretables** son aquellos que utilizan modelos estadísticos, como las regresiones lineales o árboles de decisión, para generar *insights* sobre el impacto de ciertas variables.

En cambio, los **modelos no interpretables** poseen reglas demasiado complejas, difícil de llevar a cabo por una persona. Uno de los ejemplos más conocidos son los modelos de redes neuronales que permiten predecir comportamientos. Estos modelos se inspiran en el comportamiento del cerebro humano y permiten predecir y resolver problemas usando técnicas más complejas que los algoritmos tradicionales o las técnicas estadísticas probabilísticas.

En cualquiera de los casos, los modelos predictivos se utilizan en marketing y publicidad para predecir comportamientos. Emplean, para esto, inferencias estadísticas.

3.2.2 *Machine learning*

El aprendizaje automático (o *machine learning*) "(…) tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender. Se trata de crear programas capaces de generalizar comportamientos a partir información no estructurada suministrada en forma de ejemplos". (Ecured, 2009, https://www.ecured.cu/Aprendizaje_Autom%C3%A1tico). Es decir, estos modelos se alimentan de información del pasado para predecir y clasificar nuevos datos relevantes.

Estos modelos se pueden clasificar en:

- Modelos supervisados: hallan respuestas basados en otras respuestas ya conocidas. Estos modelos se suelen emplear para que los datos históricos ayuden a predecir probables eventos futuros.
- Modelos no supervisados: aprenden paulatinamente sin tener respuestas previas. Es decir, el sistema no posee una respuesta correcta y no existen etiquetas históricas. Por ejemplo, se pueden utilizar modelos de *clustering* en donde se identifican comportamientos similares de un grupo de clientes, que se puede utilizar, luego, en acciones de *marketing* afines a sus intereses.

3.2.3 Herramientas esenciales

Al igual que para el caso de *big data* o *datawarehouse*, también hay una gran variedad de herramientas que permiten trabajar con aprendizaje automáticos y lenguajes para manipular grandes cantidades de datos no estructurados.

Detallamos a continuación algunos ejemplos:

H2O

Es una plataforma de código abierto desarrollada en Python y R, y se puede utilizar para analizar conjuntos de datos en la nube. Permite predecir situaciones de fraude o realizar análisis avanzados de impacto en publicidad.

Link: https://www.h2o.ai/products/h2o/

OpenNN

Es una biblioteca de programación que permite implementar redes neuronales.

Link: http://www.opennn.net/

PredictionIO

Es un servidor que permite desarrollar motores predictivos de cualquier tipo, personalizarlos e implementarlos en un sitio web.

Link: https://predictionio.apache.org//

3.2.4 Casos de aplicación de modelos

Existen numerosos casos de aplicación de modelos de aprendizaje automático que forman parte de nuestra vida cotidiana y que, muchas veces, apenas percibimos.

Por ejemplo, en su celular usted puede habilitar el teclado predictivo, el cual aprende, con el tiempo, las palabras que más utiliza.

Figura 6. Teclados predictivos

Fuente: Melenciano (2014) [Captura de pantalla] Recuperado de https://tecnobitt.com/como-funciona-el-texto-predictivos-en-nuestro-movil/

Los modelos de aprendizaje automático pueden tener muchas utilidades: detectar fraude, predecir qué usuarios se darán de baja de un servicio, predecir fallos tecnológicos, prever qué empleados serán más rentables en el futuro, predecir el clima, el tráfico, realizar diagnósticos médicos o predecir el impacto que puede tener una comunicación en redes sociales.

En *marketing* y publicidad, los usos del aprendizaje automático son muy valiosos para mejorar la experiencia del usuario, ofrecerles productos y servicios alineados a sus intereses y comportamientos, o para fidelizar usuarios.

Algunos ejemplos de esto son:

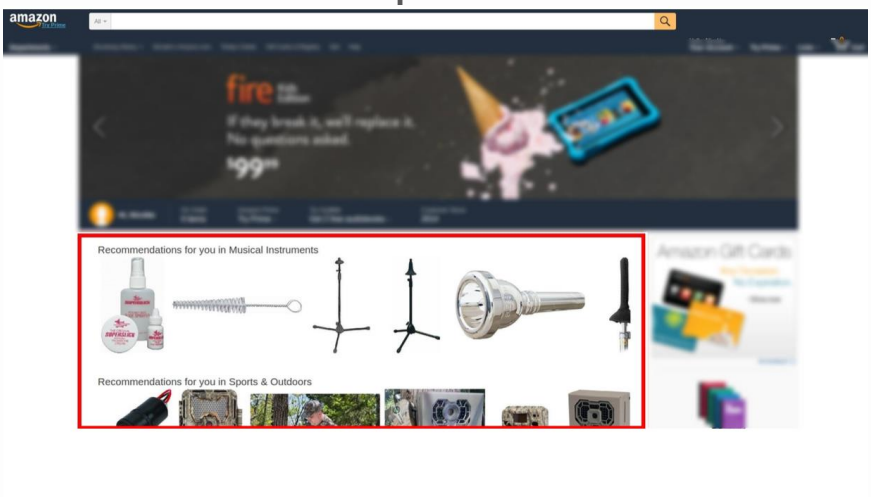
Recomendación de productos

Los sitios de comercio electrónico utilizan muchas veces técnicas de *machine learning* basadas en el comportamiento de los usuarios: productos que vieron, productos que agregaron al carrito,

categorías más buscadas (sobre las cuales los usuarios hicieron clic), etc.

Esta información es utilizada para ofrecer productos relacionados o recomendados a otros usuarios con similares características.

Figura 7. Recomendación de productos

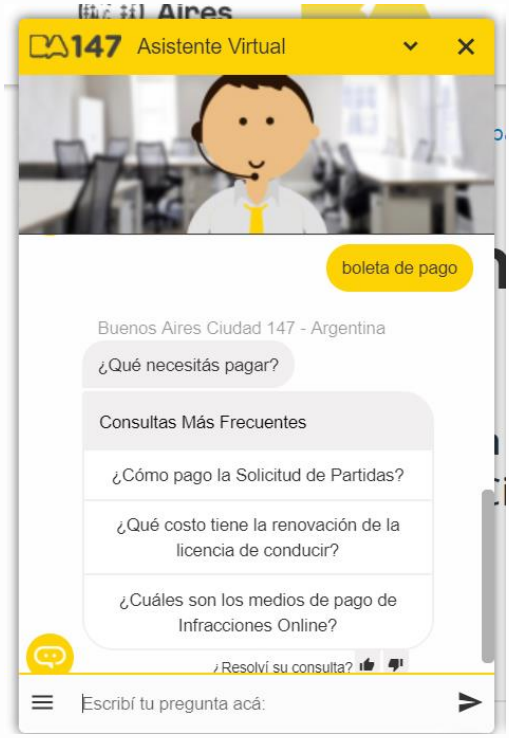


Fuente: Amazon.com. Inc (1996-2019) [captura de pantalla].
Recuperado de www.amazon.com,

Bots

Los *bots* o asistentes virtuales son un claro ejemplo de uso de técnicas de *machine learning*. Estos "robots" aprenden a medida que se interactúa con ellos y se les provee la información necesaria para que acumulen conocimiento y puedan brindar las respuestas correctas.

Figura 8. Asistente virtual



Fuente: Gobierno Ciudad de Buenos Aires (2019) [captura de pantalla]. Recuperado https://www.buenosaires.gob.ar/aplicacionesmoviles/ba-147

Lead scoring y *clustering*

Los modelos de aprendizaje automático son óptimos también en el uso de técnicas de retención y fidelización de usuarios.

Por ejemplo, el *lead scoring* o calificación de *leads* es una técnica automatizada que permite calificar a los potenciales clientes según su grado de afinidad con el cliente ideal, su interacción con la empresa y el punto del *journey* de compra en el cual se encuentra. Para esto, se utilizan herramientas de que permiten cruzar la información histórica con datos demográficos, sociales o de comportamiento.

El modelo aprende a medida que pasa el tiempo de cada uno de esos factores y genera, como resultado, un *ranking* que indican la propensión de compra (conversión) de un segmento de clientes.

A su vez, la utilización de modelos de *clustering* permite identificar segmentos de clientes que comparten comportamientos y características similares, y buscar así otros potenciales clientes con características afines a esas, pero que aún no se han convertido en clientes (*lookalike*).

Figura 9. *Lead scoring*

Fuente: elaboración propia.

Referencias

Amazon.com, Inc. [captura de panyalla]. Recuperado de www.amazon.com

Barber, M. (14 de enero de 21018). *Data science concepts you need to know! Part 1*. Recuperado de https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

BBVA API Market (2015). *Cinco librerías en Python para científicos de datos: cómo visualizar información*. Recuperado de https://bbvaopen4u.com/es/actualidad/cinco-librerias-en-python-para-cientificos-de-datos-como-visualizar-informacion

Davenport, T. H. (2012). *The seciest job of the 21st century* [documento en línea]. Recuperado de https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Ecured. (2009) *Aprendizaje automático*. [documento en línea]. Recuperado de https://www.ecured.cu/Aprendizaje_Autom%C3%A1tico

Gobierno de la Ciudad de Buenos Aires. *BA 147*. [captura de pantalla]. Recuperado de https://www.buenosaires.gob.ar/aplicacionesmoviles/ba-147