Qué es la ciencia de datos

What is data science?

Por Rosa María Valdovinos Rosas, Héctor Alejandro Montes Venegas, David Joaquín Delgado Hernández y Javier Salas García

Resumen: Denominada la profesión más sexi del siglo XXI, la ciencia de datos se ha popularizado y aplicado ampliamente a problemas que involucran análisis de datos en distintos ámbitos de la vida cotidiana. Para el logro de su cometido, se auxilia de otras disciplinas propias de la inteligencia artificial, la probabilidad y la estadística, lo que la convierte en una estrategia verdaderamente robusta para atender grandes retos, como los que presentan la nueva era de la información, la transformación digital y la 4ª revolución industrial. No obstante, debido a su versatilidad, es confundida con otras áreas. En este artículo, se realiza una breve descripción del contexto global de la generación masiva de datos, así como de la forma en que la ciencia de datos es clave para su análisis y cómo se relaciona con otras áreas.

Palabras clave: ciencia de datos, aprendizaje automático, minería de datos, estadística.

Abstract: Dubbed the sexiest profession of the 21st century, data science has become popular and widely applied to problems involving data analysis in different areas of everyday life. To achieve its mission, it is aided by other disciplines such as artificial intelligence, probability and statistics, which makes it a truly robust strategy to address major challenges, such as those presented by the new information age, digital transformation and the 4th industrial revolution. However, due to its versatility, it is often confused with other areas. In this article, a brief description is made of the global context of massive data generation, as well as how data science is key to its analysis and how it relates to other areas.

Keywords: data science, machine learning, data mining, statistics.

La humanidad se encuentra inmersa en la era tecnológica o digital, que se caracteriza por la elevada conectividad y la generación masiva de datos de distinta índole. La utilización de teléfonos celulares, el uso de redes sociales, el *e-learning*, el *e-commerce*, la computación vestible, el internet de las cosas y el *big data* generan millones y millones de datos por segundo alrededor del mundo. De hecho, la generación de datos realizada por máquinas, como aquellas provenientes de ciudades inteligentes y mediciones de dispositivos inteligentes, ocurre con solo ser utilizados, incluso sin la intervención humana.

GENERACIÓN MASIVA DE DATOS

Estudios recientes han determinado las principales tecnologías que han transformado la forma en que los seres humanos realizan actividades cotidianas, entre las que resaltan:

1. Relaciones interpersonales. Si bien este cambio ya se veía antes de la pandemia por covid-19, durante esta se disparó la utilización de redes sociales, las *date app* y las redes de opinión, convirtiéndose en un medio preferido por mu-



Ilustración: Vivian Amador

chos para comunicarse, toda vez que estas son enriquecidas con texto, llamadas, video llamadas o llamadas en grupo.

2. **Orientación al navegar.** Anteriormente, la *Guía Roji* era el principal medio para moverse cuando no se conocía la ruta para llegar a algún destino o cuando se visitaba una nueva ciudad. Hoy día, aplicaciones como Waze, TomTom y Google Maps son tan populares que, en gran medida, se depende totalmente de ellas para trasladarse caminando, en bicicleta, en carro o en transporte público.
3. **Planeación de viajes**. Hasta hace poco, la forma más tradicional y fácil de viajar consistía en visitar una agencia de viajes en la que se podía tener información de hoteles, renta de autos, lugares recomendados para visitar y para comer. Ahora, las aplicaciones como Airbnb, Despegar, Booking e, incluso, Uber, han facilitado ampliamente esta tarea, dando soluciones a la medida.
4. **Música y video.** Algunos años atrás, la principal distribución musical era en CD y otro tanto en formato MP3 y MP4 en algunas plataformas digitales. Asimismo, la generación de videoclips no profesionales era muy escasa, con difusión limitada. Hoy, debido a empresas como Spotify, Amazon, Netflix, YouTube, TikTok y otras similares, la distribución masiva de contenido musical y video profesional o no es muy fácil y, en muchos casos, de forma gratuita.

5. **Streaming.** Con el confinamiento obligado al que la población se enfrentó, la mejor estrategia implementada para la difusión masiva de conocimiento, noticias, eventos y ocio se basó en canales de *streaming*. Con esto, la comunidad científica, política, educativa y de todo ámbito encontró un medio de comunicación masivo.
6. **Dispositivos inteligentes y computación vestible.** La utilización de relojes, ropa, lentes, pañales y otros dispositivos inteligentes para censar y monitorear el estado de salud físico o medioambiental se han convertido en accesorios muy populares.
7. **Asistentes digitales.** El apoyo de un ejecutivo en un banco o un gran corporativo solía ser bastante tedioso por la realización de una fila casi interminable. Actualmente, con el desarrollo de chatbots, la atención al cliente se ha facilitado enormemente y las largas filas se han acortado significativamente, salvo aquellos casos que no pueden ser resueltos en línea.
8. **Moneda digital.** Las conocidas criptomonedas surgieron hace algunos años y son ampliamente utilizadas en el manejo de divisas. Recientemente, esta infraestructura cifrada es prometedora para el manejo no solo de dinero, sino de información sensible o el tránsito de documentos valiosos por la red.

LA CIENCIA DE DATOS APOYA LA TOMA DE DECISIONES DE LOS USUARIOS, A PARTIR DE LOS MILLONES Y MILLONES DE DATOS QUE SE GENERAN EN CELULARES, REDES SOCIALES, COMERCIO EN LÍNEA Y MUCHAS OTRAS APLICACIONES

Todas estas aplicaciones recopilan información personal, de identidad y de preferencias o intereses, rutas, frecuencia de visitas y opiniones, entre otros datos que son utilizados para mostrar ofertas, lo más visto por otros y brindar recomendaciones de compra o acción.

CIENCIA DE DATOS

El reporte *Data Never Sleeps* de la empresa DOMO (Marquina, 2021) indica que, del 2018 al 2021, se incrementó de 4.3 a 5.2 billones el número de usuarios de internet, generando un orden de 79 Zettabytes (Zb) de forma anual, proyectados a 180 Zb para el 2025. Pero, ¿qué significa esto en números? Si consideramos que un byte es la mínima unidad de medida, un Zb sería 10247 bytes, es decir 1'180,591,620,717,411,303, 424 bytes, esto, multiplicado por 180, nos daría los bytes generados de forma anual para el 2025. Las estrategias para enfrentar esta situación van en dos vertientes: el almacenamiento de los datos y el procesamiento de estos. En el primero se hace uso de *hardware*, servidores y súper computadoras, en tanto que, para el procesamiento, la ciencia de datos es la principal protagonista.

La ciencia de datos es un área del conocimiento interdisciplinar que, apoyada de otras disciplinas, permite encontrar un mejor entendimiento de los datos, con independencia de su fuente y de la estructura que tengan, para la toma de decisiones. En pro de ese cometido, esta ciencia acude a disciplinas orientadas al tratamiento masivo de datos, desde las más clásicas, como probabilidad, estadística, matemáticas, bases de datos y programación, hasta las menos convencionales, que involucran algún tipo de algoritmos de inteligencia artificial, como la minería de datos, el aprendizaje automático, el reconocimiento de patrones y el *big data*.

Se debe entender que solo hay una ciencia de datos, pero se relaciona con otras varias disciplinas. Es decir, una solución propuesta de ciencia de datos podría incluir tecnologías de *big data* (Spark, Hadoop,

Flink, etc.) para la recogida de los datos de entrada provenientes de distintas fuentes (*streaming*, imágenes de tránsito, señales sísmicas, texto, etc.), estrategias de minería de datos para realizar la limpieza de los datos (imputación, filtrado, condensado de datos, etc.), métodos de aprendizaje automático para construir el modelo (supervisado, no supervisado, de refuerzo, profundo, etc.), métodos estadísticos para el análisis de resultados (media geométrica, curvas ROC, F-estadístico, etc.) y presentar la visualización de resultados obtenidos.

Es importante precisar que lo anterior es solamente un ejemplo, ya que pudieran desarrollarse soluciones de ciencia de datos que no involucre métodos de todas las disciplinas, todo dependerá del objetivo que se persiga.

En consecuencia, la era de la transformación digital, la globalización y la alta conectividad entre dispositivos ha marcado un antes y un después, tanto en las formas en que el ser humano se relaciona como en el manejo y explotación de la información, abriendo nichos de oportunidad para los retos que hoy día los grandes corporativos y sistemas altamente conectados demandan.

Referencias

Kroese, P. Dick *et al.* (2019). *Data Science and Machine Learning: Mathematical and Statistical Methods*. *Chapman and Hall*/CRC, <https://doi.org/10.1201/9780367816971>.

Lemus-Delgado, Daniel y Ricardo Pérez-Navarro (2020). "Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos", en *Colombia Internacional*, núm. 102, pp. 41-62, <https://doi.org/10.7440/colombiaint102.2020.03>.

Marquina, Julián (2021) "¿Qué sucede en Internet en tan solo un minuto? (Cifras 2021)", en *Julián Marquina*. <https://www.julianmarquina.es/que-sucede-en-internet-en-tan-solo-un-minuto-cifras-2021/>.

Moreno Salinas, José Gerardo (2017). "Científico de datos: codificando el valor oculto e intangible de los datos", en *Revista digital universitaria*, vol. 18, núm. 7. <http://dx.doi.org/10.22201/codeic.16076079e.2017.v18n7.a2>.



Rosa María Valdovinos Rosas es doctora en Ciencias Computacionales con doble grado, miembro del SNI nivel II. Colabora en la organización de eventos de alcance nacional e intenacional para la divulgación de la ciencia, así como acciones orientadas a niñas en la computación y vocaciones tempranas con la organización de eventos, participación en ferias de ciencia y redes sociales a nivel básico, secundaria y superior. La producción científica reporta más de 1000 citas, con un índice H12 y de acuerdo con el International AD Scientific index 2023 es una las 10000 científicas más influyentes a nivel Nacional.



Héctor Alejandro Montes Venegas es doctor en Ciencias Computacionales por la Universidad de Lincoln en el Reino Unido. Ha sido asesor externo en desarrollo de software, capacitación de personal y sistemas de cómputo. Actualmente, es profesor de tiempo completo en la División de Cómputo de la Facultad de Ingeniería en la UAEMéx. Sus intereses de investigación incluyen varias ramas de la inteligencia artificial, en particular algoritmos de cómputo evolutivo y percepción robótica.



David Joaquín Delgado Hernández es doctor en Ingeniería por la Universidad de Birmingham, Inglaterra. Autor de cuatro libros, más de 100 artículos técnicos y líder de 10 proyectos de investigación científica sobre ingeniería. Ha brindado más de 25 conferencias internacionales. Ha trabajado para la Comisión Federal de Electricidad. Actualmente es profesor en la Facultad de Ingeniería de la UAEMéx, donde ha sido coordinador de posgrado, jefe de la División de Ingeniería Civil y subdirector académico.



Javier Salas García es ingeniero en Electrónica, con maestría y doctorado en Ciencias del agua. Su especialidad es el desarrollo de sistemas de medición y monitoreo aplicado a la hidrología subterránea y ciencias ambientales. Es profesor investigador en la Facultad de Ingeniería de la UAEMéx. Es integrante SNI y miembro del Registro Conacyt de Evaluadores Acreditados. Es el investigador con más patentes otorgadas a la UAEMéx, autor de más de 40 artículos publicados en revistas indexadas y en congresos nacionales e internacionales.