Rocío C. Romero Zaliz

CON-CIENCIA DE DATOS: *TRAS LAS PISTAS DEL CONOCIMIENTO*

Granada
2024

Colección
Tecnologías de la Información y la Comunicación

ISBN: 978-84-338-7450-4. Depósito legal: GR./1359-2024
Edita: Editorial Universidad de Granada
Campus Universitario de Cartuja. 18071 Granada
Telfs.: 958 24 39 30 – 958 24 62 20
web: editorial.ugr.es

Maquetación: CMD. Granada
Diseño de cubierta: Tarma. Estudio Gráfico
Imprime: Printhaus. Bilbao

Printed in Spain / *Impreso en España*

Contenido

Prólogo

Datos para compartir conocimiento, datos para comprender la vida

Los datos son al siglo XXI lo que el oro supuso para la humanidad 4.000 años a.C. Su conocimiento, obtención, manipulación y comercialización trae a la par grandes oportunidades y grandes riesgos. El mundo es así.

Familiarizarnos con la Ciencia de Datos ya no es una cuestión de cultura científica, si no de adaptación al medio, de supervivencia. Desde que nos levantamos hasta que nos acostamos nuestra vida se traduce en datos, no todo, pero casi todo. Decidir qué, cómo y hasta dónde, no es exclusivamente una cuestión de científicos y tecnólogos, es una cuestión que nos atañe a cada uno de nosotros como sociedad. Para asumir esa parte de responsabilidad que tenemos a la hora de decidir el mundo que queremos, no hay otra vía que lograr buena información, y no es una cuestión baladí, en tiempos en los que la desinformación tiene tantas vías de entrada.

Que las mismas personas que se dedican al estudio y aplicación de la Ciencia de Datos centren su empeño en que todos podamos comprender este ámbito de conocimiento, nos familiaricemos con sus conceptos, identifiquemos sus oportunidades y nos muestre las aristas en las que hay que poner atención y cuidado, es, sin duda, una fortuna. Nadie es capaz de comunicar mejor la ciencia, por compleja que sea, que quien siente pasión por ella y tiene el don de la comunicación. Si además dedica esfuerzo a formarse en divulgación, lo tenemos todo. Y este es el caso de Rocío Romero, la autora de la publicación que tienen entre manos, o frente a su pantalla.

Es un libro necesario. El último estudio de *Percepción Social de la ciencia y la Tecnología 2022,* publicado por FECYT, nos aporta evidencias de cómo la sociedad valora beneficios y riesgos de algunas aplicaciones tecnológicas. De todas las analizadas en el estudio, la única aplicación con un balance claramente positivo entre riesgos y beneficios son los aerogeneradores. Otras como la inteligencia artificial (con un balance ligeramente positivo), la robotización del trabajo, la experimentación animal o el cultivo de plantas modificadas genéticamente son percibidas con tantos beneficios como riesgos; y por último, la energía nuclear y el *fracking* son claramente valoradas con más perjuicios que beneficios.

La información a la que tenemos acceso y que manejamos impacta en la conformación de nuestra opinión, y en la postura pública que adoptamos y trasladamos.

Cuanta más y mejor información divulgativa seamos capaces de ofrecer a la sociedad, mejor sustentadas estarán estas opiniones y menos vulnerables seremos ante las informaciones falsas. Por eso, *Con-ciencia de datos: Tras las pistas del conocimiento* es una propuesta de cultura científica tan necesaria.

Las opiniones se conforman en base a los inputs que se reciben y de los valores previos de cada persona. Si queremos promover la cultura científica de la Ciencia de Datos, quien lo cuenta, es tan importante como la manera de contarlo. El hecho de que la Universidad de Granada haya decidido lanzar, a través de su Editorial una nueva colección de libros divulgativos sobre Tecnologías de la Información y la Comunicación, dirigida por el profesor Verdegay, de la que esta publicación forma parte, incrementa exponencialmente su capacidad de influencia. El informe de *Percepción Social de la Ciencia* de FECYT lo deja claro, las Universidades y los centros públicos de investigación son las organizaciones que la ciudadanía considera *"más adecuadas para explicar el impacto de los avances científico tecnológicos", seguidos de lejos por los centros de investigación privados, los divulgadores científicos en redes y blogs o los museos de Ciencia*, dejando claro que quienes suscitan más reticencia son las asociaciones de protección del medio ambiente, periodistas e industria y empresa privada.

El estudio también nos dice que los grupos profesionales de ciencia, ingeniería, así como de medicina o educación, son los más valorados por la sociedad. Un libro de la Universidad de Granada escrito por una

ingeniera investigadora, parte con las mejores condiciones para convertirse en un manual de referencia en la divulgación de la Ciencia de Datos.

En la era digital, este es un libro oportuno, y es así, porque sabemos que Internet, los libros y las revistas de divulgación científica están ganando posiciones como fuentes de información en Ciencia en España, mientras que medios como la televisión, la prensa en papel y la radio están perdiendo posiciones.

Es un libro que muchas personas esperan. El perfil de uso de los diferentes medios de obtención de información científica y tecnológica también nos orienta hacia quién va a ser el público mayoritario de este libro: serán más probablemente adultos, y personas familiarizadas con la lectura. También, y esto es una interpretación propia, supone un recurso perfecto para tantos profesores de escuelas y de institutos que están formando a las generaciones que nos darán relevo, así como para los profesionales de la comunicación que busquen una fuente fiable y amena para documentarse, y por supuesto, para las miles de personas que disfrutan adquiriendo cultura.

Y además es un libro ameno. La autora, generosa en sus apreciaciones, nos convierte en ingeniosos detectives que avanzan por el conocimiento de este campo, nos quita el temor, y nos ayuda a situarnos cogidos de su mano. Este esfuerzo por hacer divertido y riguroso lo complejo no nos sorprende a quienes hemos visto a Rocío en acción en su activo rol de divulgadora.

Les invito a disfrutar de su lectura y sobre todo, a compartirlo, porque compartir conocimiento científico será probablemente uno de los mejores regalos que puedan hacer.

Teresa Cruz Sánchez
Directora General de la Fundación Descubre

1. Introducción

¡Bienvenido al fantástico mundo de la Ciencia de Datos! A lo largo de estas páginas estudiaremos juntos los conceptos fundamentales necesarios para convertirte en un verdadero detective de datos. No necesitas muchos conocimientos sobre matemáticas, y básicamente ninguno de programación, sólo te pediré un poco de tu tiempo y muchas ganas de aprender.

En los próximas secciones descubrirás:

- *Qué es y que no es Ciencia de datos.* Hablaremos de su definición y su relación con otros términos muy utilizados actualmente.
- *Historias detrás de los datos.* Haremos también un repaso por la historia reciente y veremos casos de éxito en el uso de la ciencia de datos que tal vez desconozcas.
- *Más allá de la estadística.* No podemos olvidar la necesidad de repasar algunos datos básicos relacionadas con la estadística que nos serán de mucha utilidad en nuestros proyectos.

- *Visualización de datos.* Dicen que "una imagen vale más que mi palabras" y esto es justamente lo que analizaremos, incluyendo recomendaciones para que tus gráficos destaque y sean comprensibles por cualquier persona.
- *Sesgos en los datos.* Lamentablemente no todos los datos con que contamos tienen la misma calidad. Debemos ser conscientes de los posibles sesgos tanto a la hora de realizar un experimento y recoger datos, como de utilizar datos ya existentes recopilados por terceros.
- *Proyectos de ciencia de datos.* Trabajar con datos no es tan fácil como utilizar una planilla de cálculo y hacer unos cuantos gráficos. Para ser un buen detective de datos debemos crear un proyecto y seguir al pie de la letra los distintos pasos que lo componen.
- *Modelos de datos.* La ciencia de datos, en muchos casos, requiere el uso de herramientas computacionales más allá del uso de un paquete de ofimática. Aquí comentaremos los distintos tipos de aprendizaje que se pueden utilizar en un proyecto de ciencia de datos, ejemplos incluidos.
- *Aplicaciones.* Una vez que conocemos los pasos básicos de un proyecto de ciencia de datos te propongo ver algunas aplicaciones que te pueden servir de inspiración para que tu mismo lleves a cabo tu propio estudio.

- *¡Quiero ser un científico de datos!* En esta sección nos ponemos manos a la obra para llevar a cabo un proyecto completo desde el inicio hasta el final.
- *Datos y ética.* Responsabilidad, fiabilidad, privacidad, confianza... Estas palabras cobran sentido cuando hablamos del uso ético de los datos y repasamos algunas de los elementos que debemos tener siempre en mente a la hora de trabar con ellos.
- *Desafíos y limitaciones.* La ciencia de datos es una disciplina que no lleva entre nosotros muchos años. Esta juventud hace que, actualmente, deba hacer frente a numerosos desafíos y presente algunas limitaciones que, con el tiempo, serán seguramente superadas.
- *El futuro de la ciencia de datos.* En este mundo que está en cambio constante lo único que podemos saber a ciencia cierta es que todo va a cambiar. Podemos ya atisbar parte de ese futuro y entrever hacia donde nos llevará la ciencia en los próximos años.
- *Recomendaciones finales.* No podemos acabar estas páginas sin resumir lo aprendido y dar unas últimas recomendaciones finales que te servirán para que te animes a ponerte manos a los... ¡datos!
- *Recursos adicionales.* Si te ha gustado el mundo de la ciencia de datos en esta última sección te dejo varios recursos adicionales en distintos formatos (e.g., libros, blogs, podcasts) para que te sigas formando y aprendiendo día a día.

Una vez concluida la lectura de estas páginas espero que seas consciente del poder que tienen los datos y de cómo estos pueden ayudarte a tomar mejores decisiones profesionales y personales.

Prepárate para iniciar esta aventura y aprender paso a paso cómo llevar a cabo tu propio proyecto de ciencia de datos. Sorprende a tus amigos y familiares descubriendo el porqué del mundo que te rodea y sácale el máximo partido a tus propios datos.

Por último, un dato aclaratorio antes de empezar. En todo lo que sigue, el uso del género masculino en el texto se emplea de manera neutra y no excluye a personas de otros géneros.

2. ¿Qué es y qué no es *Ciencia de Datos?*

Querido aprendiz de detective de datos, antes que nada necesitamos dejar algo en claro, no todo se considera *Ciencia de Datos*. Porque el término esté de moda no implica que tenemos que usarlo indiscriminadamente. Si..., sabemos que decir que hacemos ciencia de datos es mucho más glamuroso que decir que has hecho un gráfico en Excel. Pero la ciencia de datos va mucho más allá de un gráfico bonito, y estas páginas están para demostrarlo.

Por un lado, cuando hablamos de *Ciencia* nos referimos comúnmente al "conjunto de conocimientos obtenidos mediante la observación y el razonamiento". Por otro lado, cuando hablamos de *Datos* nos referimos a la "información sobre algo concreto que permite su conocimiento exacto"[1]. Pero entonces... ¿a qué nos referimos cuando hablamos de *Ciencia de Datos* en forma

1. Real Academia Española, http://rae.es

conjunta? La respuesta es sencilla, hablamos de obtener nuevo conocimiento sobre algo concreto, los datos que tengamos disponibles, a través de la observación y el razonamiento. La ciencia de datos en sí misma es un campo multidisciplinar que combina disciplinas tales como la estadística, la matemática y la informática, para analizar, interpretar y visualizar datos. El objetivo de la ciencia de datos es poder extraer información y conocimiento útil y relevante a partir de un conjunto o conjuntos de datos. Esto incluye también el hecho de poder comunicar los hallazgos de forma clara y sencilla.

Como buen aspirante a detective de datos habrás notado que en esta última década se han empezado a utilizar cada vez más varios términos que ahora mismo están muy, muy de moda: inteligencia artificial, minería de datos, *big data*, aprendizaje automático, y, por supuesto, ciencia de datos. Estos términos están tan estrechamente interrelacionados que a veces resulta complicado distinguir claramente dónde acaba uno y empieza el otro. Analicemos entonces sus similitudes y diferencias antes de ponernos manos a los... datos.

Comenzaremos con los dos términos más utilizados actualmente: inteligencia artificial y ciencia de datos. Por un lado, la ciencia de datos utiliza herramientas informáticas para poder procesar los datos, incluyendo herramientas dentro de la categoría de inteligencia artificial, como podrían ser las redes neuronales artificiales (si quieres aprender más de ellas no te pierdas el cuadernillo "Inteligencia Artificial" de esta misma colección). Sin embargo, también utiliza técnicas estadís-

ticas, tanto simples como complejas, como ser el cálculo de la media, la moda y demás medidas de la estadística descriptiva; o el análisis de componentes principales que permite reducir la complejidad de nuestros datos y representarlos de manera más sencilla. Por otro lado, la inteligencia artificial se usa en ámbitos que van más allá de la ciencia de datos, como puede ser la robótica o la visión artificial. Por tanto podemos decir que ambos términos tienen un cierto solapamiento.

¿Qué pasa entonces con los términos minería de datos y ciencia de datos? Muchas personas suelen utilizar ambos términos como sinónimos, tal vez por el hecho que su objetivo general es el mismo: extraer conocimiento útil a partir de conjuntos de datos. Sin embargo, hay diferencias claras entre ellos. La minería de datos se centra en descubrir patrones y tendencias ocultos en grandes conjuntos de datos, mientras que la ciencia de datos es un campo más amplio que incluye además de la minería de datos, la recopilación, limpieza, análisis y visualización de datos. Se puede decir entonces que la minería de datos es una parte integral de la ciencia de datos.

¿Y qué pasa con el Aprendizaje Automático? El aprendizaje automático, también conocido como *machine learning* en inglés, es una rama de la inteligencia artificial dedicada a desarrollar herramientas que permitan aprender a partir de datos. En otras palabras, podemos decir que son capaces de aprender a partir de ejemplos, en contraposición a ser explícitamente programadas para realizar una tarea específica siguiendo

una serie de pasos predefinidos. Es decir, estas técnicas de aprendizaje automático son capaces de mejorar su rendimiento a medida que se exponen a más y más datos. Veremos algunas de esta técnicas en la Sección 7 (Modelos de Datos). Por tanto, el aprendizaje automático es una herramienta esencial para la minería de datos. Sin embargo, su uso va más allá de la ciencia de datos y puede usarse en otras áreas de la inteligencia artificial, como el procesamiento del lenguaje natural o la detección de objetos en tiempo real.

Por último analicemos otro de los términos más utilizado últimamente: *Big Data*. Éste término se traduce al español como *macrodatos* o el "conjunto de datos que, por su gran volumen, requieren técnicas especiales de procesamiento"[1]. La ciencia de datos se aplica a cualquier conjunto de datos, tanto pequeños como grandes, pero si éstos son lo suficientemente grandes como para caer en la categoría de los macrodatos, entonces requerirán de herramientas específicas para poder tratarlos, tal como indica su definición.

Para resumir este análisis de la terminología y poder comprender mejor la relación entre todos los términos vinculados a la ciencia de datos, te invito a explorar la Figura 1 que presenta un resumen visual de estos términos y sus interrelaciones.

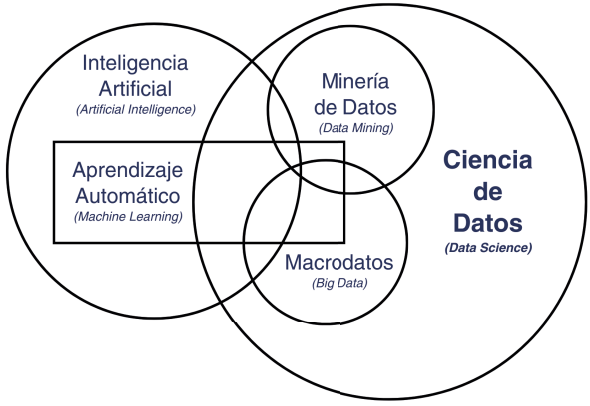

Figura 1. Relación entre los términos más comunes relacionados con la Ciencia de Datos.