No. 1
Vol. 2023
Katra

Ciencia, arte, tradición, opinión, reflexión y meditación.

Artículo: **¿Qué es la ciencia de datos?**

Autor(es): **Dr. Salvador Godoy Calderón**
sgodoyc@gmail.com

Publicación: **No. 1, vol. 2023**

No. 1
Vol. 2023

¿Qué es la ciencia de datos?

Salvador Godoy-Calderón
sgodoyc@gmail.com

Una de las decisiones más difíciles de tomar en la vida, es sin *duda*, la elección de la carrera o profesión que se desea estudiar. Para un estudiante de nivel medio-superior que esté próximo a concluir sus estudios, resulta muchas veces, una decisión inevitable y para la cual ha tenido muy poca o ninguna preparación previa. La decisión es complicada por varias razones de diferentes tipos:

- **Incertidumbre**. El contexto y las costumbres sociales suelen sugerir, a veces de forma explícita, y a veces a nivel intuitivo o subconsciente, que se trata de una decisión que, de alguna forma, va a *determinar* el tipo y calidad de vida, durante gran parte del resto de la vida. Sin embargo, no se tiene certeza, de la veracidad ni del alcance de esa afirmación.

- **Ruido informativo**. La orientación y consejos que recibe el estudiante que debe tomar la decisión, siempre resulta confusa e incluso contradictoria. No hay claridad sobre el principio fundamental que debe guiar la decisión: el análisis de las habilidades personales, el gusto o interés personal, la perspectiva de trabajo a corto, mediano o largo plazo, la coyuntura política y social del momento, las tendencias y modas en las redes sociales, o inclusive la tradición familiar.

- **Falta de información**. En última instancia, existe una ignorancia fundamental sobre cuáles son las opciones disponibles para estudiar, cuáles son las condiciones de exigencia y/o competencia que imperan durante el estudio de cada opción o qué tipo de actividad profesional implica cada opción.

Aunque tratar de resolver todas las categorías anteriores parece poco menos que imposible, en ***Katra*** tenemos el convencimiento de que sí podemos ayudar en la solución de la última categoría, la falta de información. Por eso, estableceremos una

sección especial, fija en cada número, en la que analizaremos alguna de las opciones para estudios de nivel licenciatura.

Una ciencia para estudiar los datos

Aunque la historia de las carreras universitarias nos enseña que, en muchos momentos se les ha dado un nombre un tanto exótico a algunas carreras, el caso de la *ciencia de datos* (*CD*), afortunadamente, no es uno de esos casos. El nombre de esta carrera claramente indica dos cosas muy evidentes:

- Que se trata de una ciencia o disciplina científica,
- Que su objeto de estudio son los datos.

En esencia, esta disciplina no se limita, de forma natural, a estudiar datos de algún tipo específico. Como disciplina científica, aspira a poder analizar cualquier tipo de datos, aunque en términos reales, está limitada por la capacidad computacional para almacenar, organizar y procesar conjuntos de datos.

Ahora bien, más allá de esos elementos evidentes, es necesario profundizar un poco en el conocimiento de esta disciplina para poder conocerla lo suficiente para tener elementos sólidos con los cuales tomar la decisión de estudiarla y dedicarse de forma profesional a ella. Para esa profundización se pueden seguir diferentes visiones. La visión más tradicionalista comienza por buscar algunas definiciones de la disciplina. Veamos algunas de esas definiciones:

- En el *Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas* (IIMAS) de la *Universidad Nacional Autónoma de México* (UNAM), se ofrece la carrera de *Licenciatura en Ciencia de Datos* (se puede consultar en *https://www.iimas.unam.mx/ciencia-de-datos/*) y se presenta con el siguiente texto:

  > "Es una ciencia transdisciplinaria que involucra métodos científicos, procesos y sistemas para obtener un mejor entendimiento de grandes cantidades de datos, con el fin de identificar patrones en fenómenos reales para que de esta manera se tomen decisiones asertivas".

- En el sitio de *Amazon Web Services* (*AWS*), que se encuentra en

https://aws.amazon.com/es/what-is/data-science/ la definen de la siguiente forma:

> "La ciencia de datos es el estudio de datos con el fin de extraer información significativa para empresas. Es un enfoque multidisciplinario que combina principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para analizar grandes cantidades de datos. Este análisis permite que los científicos de datos planteen y respondan a preguntas como "qué pasó", "por qué pasó", "qué pasará" y "qué se puede hacer con los resultados".

- En el sitio de la empresa *International Busines Machines* (*IBM*), que se puede consultar en *https://www.ibm.com/mx-es/topics/data-science*, se define de forma muy semejante:

> "La ciencia de datos combina las matemáticas y la estadística, la programación especializada, los *analytics* avanzados, la inteligencia artificial (IA) y el aprendizaje automático con conocimientos específicos en la materia para descubrir *insights* procesables ocultos en los datos de una organización. Estos *insights* pueden utilizarse para orientar la toma de decisiones y la planificación estratégica".

A pesar de todos esos intentos de definición, la verdad es que resulta muy difícil crear una definición universal para el término *ciencia de datos,* sobre todo, si la intención de esa definición es que el término se use correctamente en conversaciones cotidianas y que su significado esté siempre claro.

Una forma de comenzar a comprender la disciplina de *ciencia de datos* es examinando su *ciclo de operación.* Ese ciclo representa las distintas actividades que un profesional de la *ciencia de datos,* debe realizar durante la solución de algún problema concreto planteado por la empresa u organización para la que trabaja. El ciclo se divide en siete etapas y, durante cada una de ellas se realiza una actividad diferente:

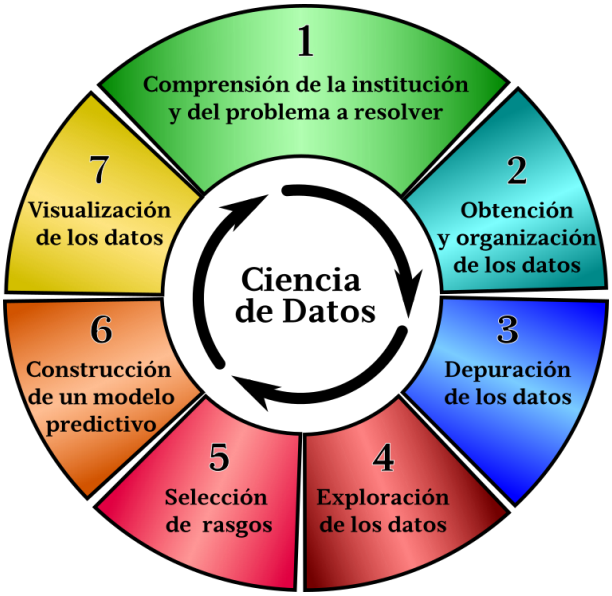


Figura 1. Ciclo de operación de la ciencia de datos

La figura 1 muestra el ciclo de operación de la *ciencia de datos.* Examinemos brevemente en qué consiste cada etapa:

1) La etapa más relevante del ciclo, por ser la primera y la última. Antes que cualquier otra actividad es necesario comprender cabalmente la forma de operar de la empresa u organización (el *modelo de negocio*), así como entender sus necesidades de toma de decisiones. Con esa información se puede definir con precisión el problema a resolver con ayuda de datos.

2) Cuando se ha definido el problema a resolver se requiere recopilar todos los datos disponibles acerca del fenómeno en cuestión, organizarlos, integrarlos y, en ocasiones, homogeneizarlos para hacerlos compatibles.

3) La limpieza de los datos consiste fundamentalmente en solventar la ausencia de información, ya sea eliminando los datos incompletos o estimando su valor con técnicas probabilísticas, resolver las incongruencias o contradicciones presentes en los datos, y muchas veces, reestructuralos creando catálogos de referencia; un proceso que en el ámbito de las bases de datos se denomina *normalización* de la base de datos.

4) La exploración inicial de los datos ya depurados es un análisis estadístico clásico que indica las características más relevantes de ese conjunto de datos. Para cada atributo de los objetos descritos, conocer su dominio, valor mínimo y máximo, su media, moda y desviación estándar, así como caracterizar la distribución de probabilidades de esos valores.

5) No todos los atributos de los objetos descritos resultan relevantes al analizar datos. La *selección de rasgos* es la identificación de los atributos que son relevantes y de utilidad para el problema que se desea resolver. Considerar sólo esos atributos y no todos los disponibles permite simplificar los modelos que se obtendrán en la siguiente etapa.

6) El análisis estadístico y probabilístico de los datos no es suficiente para tomar decisiones importantes a partir de los datos. Es entonces cuando se acude a una disciplina muy relacionada: el *aprendizaje automático* (*machine learning*) para crear, a partir de los datos analizados, un *modelo predictivo* del fenómeno estudiado.

7) Finalmente, se realizan suficientes experimentos con el modelo predictivo y se hace uso de métodos y herramientas para visualización de los resultados. Esa visualización permite formar una expectativa del comportamiento del fenómeno y estimar el alcance de las decisiones que se toman al respecto.

A partir de todos los intentos anteriores de definición, y de examinar con detenimiento el ciclo anterior, podemos extraer otros dos elementos que parecen formar parte de esta disciplina:

- Resulta de especial interés para empresas y organizaciones de todo tipo,
- Cotidianamente debe trabajar con «grandes volúmenes de datos»

Precisar con exactitud a qué se refiere el término «grandes volúmenes de datos» tampoco es sencillo. En términos cotidianos, significa un volumen tal de datos que no tiene sentido pensar en analizarlos sin la ayuda de computadoras. Aunque, desde el punto de vista de la computación, se usa el término «grandes volúmenes

de datos» para referirse a conjuntos de datos de tamaño superior al tamaño de la memoria de la computadora. Es decir, que en esa computadora, el conjunto de datos no puede ser cargado en memoria de forma completa. Por supuesto, esa es una apreciación subjetiva que depende de la computadora con la que se pretende analizar los datos, pero nuevamente, nos da una idea de la importancia de la tecnología computacional para hacer *CD.*

En cuanto al contexto empresarial, resulta natural que a las empresas e instituciones de todo tipo, les interese poder estudiar sus datos, para así tomar decisiones informadas en todo tipo de procesos; mercadotecnia, logística, política, etc. Estudiar datos significa algo muy concreto: encontrar *patrones interesantes e implícitos* en los datos.

¿Qué significa buscar patrones en los datos? En el contexto del análisis de datos, los *patrones* son datos, secuencias de datos o estructuras que se repiten de forma regular. y que, por lo tanto, nos dicen algo sobre la naturaleza de los fenómenos cuyos datos estamos analizando. Estos patrones no es fácil identificarlos de primera vista, a menos que se trate de un conjunto muy pequeño de datos. Por eso, para analizar grandes volúmenes de datos, es indispensable hacerlo por medios computacionales; y eso explica la relación de la *CD* con la *computación.* Ahora bien, existe una gran variedad de patrones que se pueden encontrar en los datos. Tradicionalmente, cada tipo de patrón ha sido estudiado por una disciplina diferente y, en *CD* se pretende desarrollar las habilidades necesarias para encontrar todos estos tipos de patrones. Veamos algunos ejemplos de tipos de patrones en los datos:

- ***Medidas de tendencia central y dispersión.***- Son medidas estadísticas, patrones numéricos que tradicionalmente pretenden resumir un conjunto de valores en un solo valor con significado bien definido para generar una «visión global» de un conjunto de datos. Las medidas de tendencia central indican los valores en torno a los cuales se pueden agrupar los datos: medias, medianas y modas; mientras que las medidas de dispersión indican los límites de variabilidad en los valores de cada variable y con respecto a las medidas de tendencia central. Estos patrones son

estudiados por la disciplina de *Estadística Descriptiva.*

- ***Asociaciones relacionales.-*** También llamados patrones asociativos, son asociaciones entre dos o más atributos de los objetos o fenómenos descritos por los datos. Por ejemplo, en una base de datos de habitantes de una ciudad, podemos encontrar patrones asociativos como: «el 90% de los adultos mayores de 65 años y que están jubilados, viven en la zona sur de la ciudad» o «el 45% de los estudiantes graduados en universidades de la zona norte de la ciudad, durante los últimos 3 años, trabaja actualmente en empresas extranjeras». Este tipo de patrones ha sido estudiado por la disciplina llamada *Minería de reglas de asociación* y nació como una herramienta para analizar las tendencias de compra de artículos en tiendas departamentales.

- ***Series de tiempo.-*** Se trata de secuencias cronológicas de datos acerca de un mismo fenómeno y en las que el valor de cada dato depende, tanto de los datos anteriores, como de las circunstancias de cada momento. Al analizar series de tiempo, generalmente se pretende *predecir* uno o más datos que aparecerán en los siguientes elementos de una serie. Este tipo de patrones constituye el elemento fundamental para el estudio de diversos fenómenos, desde señales biológicas (electrocardiografía, magnetoencefalografía, e imagenología médica en general), hasta mercados financieros (bolsa de valores). Por ejemplo, se pueden encontrar patrones que indiquen que: «Durante los últimos 10 años, el consumo de los usuarios ha sido 85% superior durante los meses de noviembre y diciembre, pero 25% inferior durante los meses de enero y febrero» .

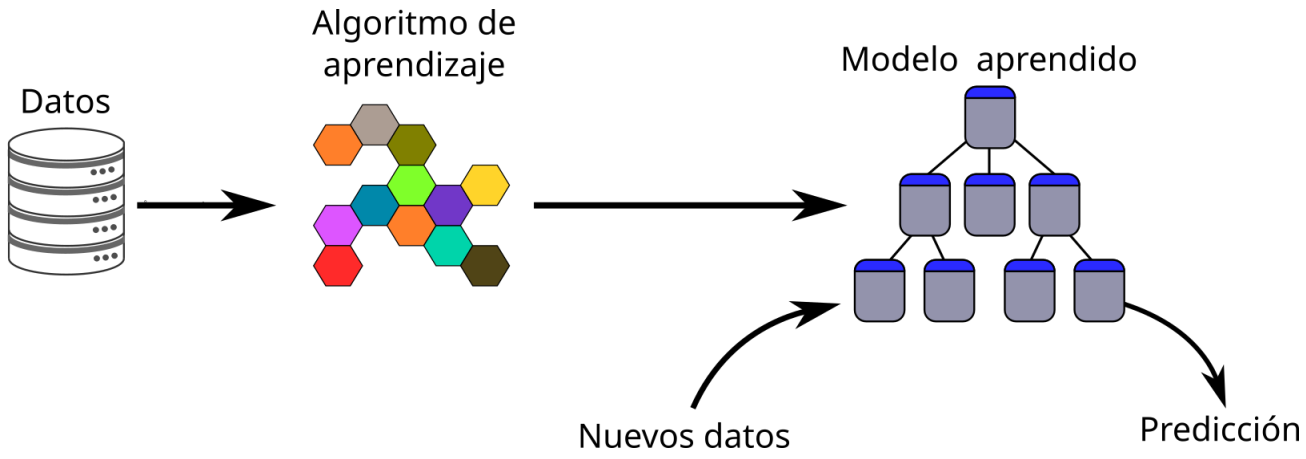
Un *científico de datos,* el profesional de la *Ciencia de Datos,* necesita comprender la naturaleza de estos y otros tipos de patrones en los datos, así como conocer técnicas y algoritmos para encontrar dichos patrones en los conjuntos de datos. En casos extremos, en los que no existan algoritmos previos para encontrar algún tipo particular de patrones, el profesional de la *Ciencia de Datos* debe tener la capacidad

de proponer nuevas estrategias y algoritmos.

La relación con Inteligencia Artificial

Un aspecto sumamente interesante de la *Ciencia de Datos*, y que vale la pena analizar y aclarar, es su relación con la disciplina de *Inteligencia Artificial*. Evidentemente no se puede pretender que un profesional de la *Ciencia de Datos*, sea también un experto en *Inteligencia Artificial (IA).* Entonces, ¿qué parte de la *IA*, requiere conocer un *científico de datos*? Aunque diversas ramas de la *IA* pueden ser extremadamente útiles para las etapas de *toma de decisiones,* una vez que se han analizado los datos, para el profesional de la *CD,* que es el encargado del proceso previo de analizar los datos, la rama de *IA* que tiene directa relación es el *aprendizaje automático,* conocido por algunos como *aprendizaje de máquina* (*machine learning*). Esta es una de las ramas más importantes, pero menos entendidas de la *IA. Aprender* significa encontrar o generar un *modelo* de algún fenómeno, a partir de analizar algunos datos previos sobre ese mismo fenómeno. ¿Parece familiar?, por supuesto.

El *aprendizaje automático* es una rama de la *IA* que también analiza conjuntos de datos, pero con el propósito particular de "aprender" de ellos, eso significa generar un *modelo* para el fenómeno que describen esos datos. El *modelo* que se obtiene debe ser capaz, tanto de *explicar* el fenómeno, como de *predecir* ese mismo



fenómeno. Sin embargo, como ocurre en la mayoría de las disciplinas científicas que han llegado a un nivel de madurez, exiten en *IA* corrientes de pensamiento, o

paradigmas que proponen diferentes formas de realizar los mismos procesos.

Figura 2. **La dinámica de operación del aprendizaje automático**

En el caso del *aprendizaje automático,* el modelo que se aprende de un conjunto de datos, puede tener forma *explícita* o sólo *implícita.* Cuando un modelo es *explícito* (también llamado *modelo simbólico*), significa que se puede observar directamente el modelo, se puede saber qué estructura tiene e incluso se pueden realizar ciertos ajustes a ese modelo para incorporar conocimiento previo sobre el fenómeno que se está aprendiendo. Este tipo de modelo se puede expresar con ecuaciones matemáticas, pero lo más común es expresarlo con *fórmulas lógicas.* Ambos tipos de expresiones son de tipo *simbólico,* es decir que tienen dos características principales:

1) Incluye símbolos especiales llamados *variables* que representan, ya sea elementos conocidos del fenómeno que se está aprendiendo (a esto se le llama *variable de primer orden*), o bien otros elementos del mismo modelo (*variables de orden superior*).

2) SIEMPRE, sin excepción alguna, se tiene clara la *semántica* de cada variable, así como la *semántica* de las expresiones que incluyen esas variables. Esto garantiza que cualquier analista, que comprenda el lenguaje en el que se expresó el modelo (lenguaje lógico o lenguaje matemático), siempre va a poder *comprender* lo que el modelo está representando, y por tanto podrá también, modificarlo para mejorar su poder predictivo y estimar el alcance de esas modificiaciones.

Los modelos de este tipo se denominan *modelos simbólicos* y el proceso de aprendizaje se conoce como *aprendizaje simbólico automático* (*symbolic machine learning*).

Por el contrario, cuando un modelo es *implícito* (también llamado *modelo conexionista*), no es posible observarlo directamente, porque está *codificado* de alguna forma que no resulta trivial de entender. Este tipo de modelos NO son

simbólicos y, por lo tanto, *NO* incluyen variables o símbolos de ningún tipo. Inclusive si se entiende la forma en que está codificado el modelo, debido a que no incluye variables simbólicas, no resulta recomendable modificar el modelo aprendido, ya que los efectos de esa modificación son impredecibles. Este es el tipo de modelo que generan todas las *redes neuronales artificiale* (*RNA*). Cuando una *RNA* aprende un conjunto de datos, el *modelo* aprendido se codifica en forma de pesos para cada una de las conexiones entre neuronas artificiales. Entonces, para usar el modelo aprendido por una *RNA* para realizar *predicción* de algún fenómento, la única opción es usar esa misma *RNA* y alimentarla con nuevos datos. Por supuesto, eso implica que el modelo aprendido es muy inflexible; cualquier cambio en la estructura de los datos de entrada, la conexión interna de la *RNA,* o las condiciones del experimento para el cual se desea hacer la predicción, puede implicar que el modelo aprendido dejó de ser útil y es necesario volver a realizar el proceso de entrenamiento de la *RNA.*

Sobre la actividad profesional

Como ocurre con cualquier otra carrera que se estudie de manera formal, a veces la práctica profesional puede no coincidir con la actividad durante el tiempo de estudio.

Lo común durante el proceso de estudio, es que se haga mucho énfasis en *comprender* la forma de operar de los diferentes algoritmos para analizar datos y que se invierta mucho tiempo en *programar* desde cero esos algoritmos, así como en desarrollar la habilidad para *diseñar, proponer* y *poner en operación* nuevos algoritmos que se adapten a condiciones muy específicas de operación. Sin ese nivel de comprensión, ni el desarrollo de esas habilidades, no tendría ningún valor diferenciante un profesional graduado en *ciencia de datos,* se reduciría a sólo ser un técnico que cotidianamiente usa algunas *herramientas* para realizar análisis de datos y, consecuentemente, su nivel de demanda en el mercado laboral se reduce considerablemente, al igual que su espectativa de salario.

Desafortunadamente, también resulta común que, en muchas empresas e instituciones, se adoptan políticas para usar programas y sistemas comerciales para realizar análisis de datos.

Para saber más, consulta...

- El sitio *web* de Ciencia de Datos en Wikipedia (español), en *https://es.wikipedia.org/wiki/Ciencia_de_datos*
- El sitio *web* de *Microsoft Azure,* ¿Qué es la ciencia de datos? *https://azure.microsoft.com/es-mx/resources/cloud-computing-dictionary/what-is-data-science*
- El sitio *web* de *Oracle México,* ¿Qué es la ciencia de datos? *https://www.oracle.com/mx/what-is-data-science/*
- El sitio *web* de la *Escuela británica de artes creativas y tecnología,* ¿Qué es la ciencia de datos? *https://ebac.mx/blog/que-es-la-ciencia-de-datos*
- El sitio *web* de *sudeep.co, Understanding the Data Science Lifecycle* *https://www.sudeep.co/data-science/2018/02/09/Understanding-the-Data-Science-Lifecycle.html*
- El sitio *web* de la *SAAS,* ¿Qué es un científico de datos? *https://www.sas.com/es_mx/insights/analytics/what-is-a-data-scientist.html*