Explorando la Ciencia de Datos: Desde la Estadística hasta el Big Data

José Luis Quiroz Fabián
Adriana Pérez Espinosa
Graciela Román Alonso
Miguel Alfonso Castro García
Manuel Aguilar Cornejo
Departamento de Ingeniería Eléctrica,
Universidad Autónoma Metropolitana-Iztapalapa

Casa abierta al tiempo

Foto de Markus Spiske en pexels.com

Resumen
Actualmente, la Ciencia de Datos es un campo interdisciplinario que se centra en la extracción de conocimiento útil de grandes conjuntos de datos. Se enfoca en la utilización de herramientas y técnicas para la recopilación, procesamiento y análisis de datos, con el objetivo de identificar patrones, tendencias y relaciones que puedan ser utilizados para mejorar la toma de decisiones. En este artículo, se busca dar un panorama general de la Ciencia de Datos y el conocimiento que implica, proporcionando una visión integral de sus conceptos, metodologías y requerimientos del tratamiento de datos masivos.

Palabras clave: Ciencia de Datos, Estadística, CRISP-DM, Datos Masivos, IA.

Abstract
Today, Data Science is an interdisciplinary field that involves extracting useful knowledge from large datasets. It focuses on using tools and techniques to collect, process, and analyze data in order to identify patterns, trends, and relationships that can be used to improve decision-making. This article aims to provide a general overview of Data Science and the knowledge it encompasses, offering a comprehensive view of its concepts, methodologies, and Big Data challenges.

Keywords: Data Science, Statistics, CRISP-DM, Big Data, AI.

1. Introducción

La **Ciencia de Datos** es una colección de técnicas utilizadas para extraer valor de los datos [1], donde se busca responder preguntas como **“¿Qué pasó?”, “¿Por qué pasó?”, “¿Qué pasaría?” y “¿Qué se puede hacer con los resultados?”.** La Ciencia de Datos abarca varias disciplinas, combinando Estadística y Matemáticas, Ciencias de la Computación y diversos campos de aplicación, incluyendo (pero no limitándose) a los negocios, la salud, la investigación científica y el sector público. A los profesionistas en el campo de la Ciencia de Datos se les conoce como **Científicos de Datos**.

1.1 Clasificación de problemas en la Ciencia de Datos

Los problemas en Ciencia de Datos se pueden clasificar según el tipo de problema a resolver, el tipo de datos utilizados, el enfoque del aprendizaje automatizado y el objetivo del problema:

Tipo de problema: por ejemplo problemas de regresión, problemas de agrupamiento (clustering), problemas de detección de anomalías, entre otros.

Tipo de datos: datos estructurados, datos no estructurados, datos semi-estructurados (como XML o JSON) y datos de secuencia (como los datos de series temporales).

Enfoque del aprendizaje automatizado: Modelos que usan datos con respuestas conocidas (supervisados) o modelos que buscan patrones sin respuestas predefinidas (no supervisados).

Objetivo del problema: Por ejemplo, se pueden realizar análisis predictivos o descriptivos para obtener estadísticas.

1.2 Ciencia de Datos y otras líneas de investigación

La Ciencia de Datos es un área de conocimiento que se conforma de tres grandes

áreas: las Ciencias de la Computación, las Matemáticas en su rama de Estadística y las Habilidades de Negocio (véase la Figura 1 con el Diagrama de Venn de Drew Conway [2]).

Por Ciencias de la Computación nos referimos a los principios y técnicas para diseño y desarrollo de los sistemas computacionales. Se abarca desde algoritmos simples hasta algoritmos de inteligencia artificial e interacción humano-computadora.

De igual forma las Ciencias de la Computación junto con las Matemáticas y la Estadística dan pie al Aprendizaje Automático o Machine Learning, que se enfoca en la creación de modelos predictivos automatizados [3]. Con ello la Ciencia de Datos extrae conocimiento y genera descubrimientos útiles (insights), generalmente a partir de grandes volúmenes de datos.

Finalmente, la Ciencia de Datos incorpora habilidades de diferentes disciplinas (Habilidades de Negocio), lo que implica una comprensión profunda del área de estudio para poder aplicar los resultados de manera efectiva en diversos contextos, incluyendo empresariales, académicos, científicos y de toma de decisiones.

2. Estadística en la Ciencia de Datos

La estadística es fundamental en la Ciencia de Datos, ya que proporciona las herramientas necesarias para interpretar y analizar datos. La estadística descriptiva se utiliza para resumir y visualizar datos, lo que permite identificar patrones y tendencias mediante medidas como la media, la mediana y la desviación estándar, así como a través de gráficos. Por otro lado, la estadística inferencial permite hacer predicciones y generalizaciones sobre una población a partir de una muestra, utilizando técnicas como intervalos de confianza y pruebas de hipótesis. Ambas ramas son esenciales para tomar decisiones informadas basadas en datos, evaluar modelos y validar resultados en la Ciencia de Datos.

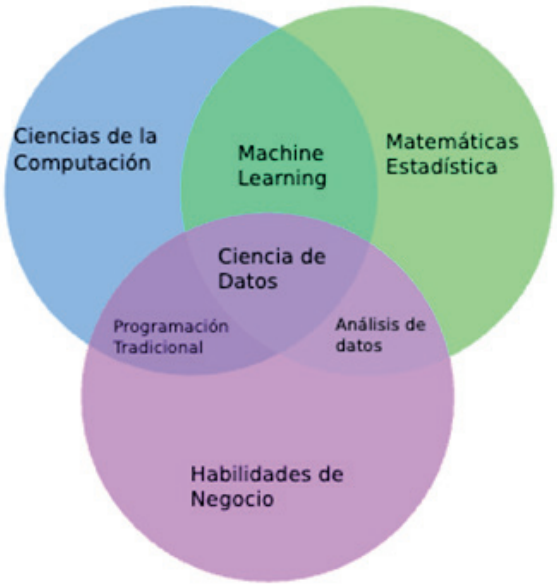

Figura 1. Diagrama de Venn de Drew Conway ilustrando las disciplinas que conforman la Ciencia de Datos.

3. Metodología o ciclo de vida de la Ciencia de Datos

Inicialmente, los equipos de análisis de datos seguían metodologías basadas en la experiencia personal, así como en estrategias de ensayo y error. Con el tiempo, organizaciones como Microsoft e IBM desarrollaron estándares metodológicos. El más importante y ampliamente utilizado en la industria es el Cross-industry standard process for data mining (CRISP-DM), o proceso estándar interindustrial para minería de datos, publicado en 1999 [4]. Este fue el primer gran esfuerzo por crear una metodología estándar para el

análisis de datos. Aunque existen otras metodologías, la mayoría se asemeja a CRISP-DM, que incluye las siguientes etapas: comprensión del negocio o problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implantación.

3.1 Comprensión del problema o negocio

Esta etapa es análoga a la toma de requerimientos en la ingeniería de software, se habla con el usuario o cliente y se busca establecer los objetivos claros, para luego traducirlos en requerimientos técnicos.

3.2 Comprensión de los datos

Se recaban los datos que se van a emplear durante el proyecto y se realiza un análisis preliminar con el fin de adquirir un mejor entendimiento de los datos. Se pueden empezar a formular hipótesis y a identificar datos faltantes. También se emplean herramientas como gráficas para adquirir un entendimiento de alto nivel de los datos.

3.3 Preparación de los datos

Se seleccionan los datos que se van a emplear durante el análisis y se realizan diferentes transformaciones sobre ellos, tales como la selección de subconjuntos de datos utilizables, el reemplazo de campos faltantes, combinación de datos para formar datos nuevos, adaptar el formato para ser utilizable en alguna herramienta, normalizaciones, etc.

3.4 Modelo de los datos

Se seleccionan una o varias técnicas de modelado que se van a llevar a cabo con los datos, por ejemplo redes neuronales, agrupamiento, regresiones, etc. También se diseñan pruebas mediante las cuales se va a evaluar la calidad de los distintos modelos producidos durante esta fase. En la Sección 4 se detalla más esta etapa.

3.5 Evaluación

Se realiza un análisis a fondo, tanto del modelo seleccionado en la etapa anterior como de todo el proceso seguido hasta este punto, para determinar si el modelo cumple con los objetivos establecidos. De acuerdo a este análisis se decide si se va a seguir con la fase de implantación, si se van a realizar más iteraciones del proyecto o si se va a empezar un proyecto nuevo.

3.6 Implantación

La información proporcionada por el modelo se organiza para poder ser presentada al cliente. También se elabora un plan de implantación con el fin de que el cliente tenga claros los pasos a realizar para poder hacer uso efectivo del modelo generado

4. Modelo de Datos (Modelo de Aprendizaje)

En Ciencia de Datos, se utilizan modelos para analizar, procesar y predecir información. Aunque todas las etapas de la metodología son importantes, los modelos de aprendizaje tienen un peso especial. Estos modelos incluyen técnicas de *Aprendizaje Automático* que se aplican a diversos tipos de problemas, como la predicción, el agrupamiento, la clasificación, la identificación de patrones, la optimización, entre otros. Los modelos seleccionados dependen del tipo de problema que se desea resolver.

Algunos modelos comunes incluyen:

- *Regresión lineal:* Modelo para aproximar o predecir el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido.

- *Redes neuronales:* Modelos que imitan un cerebro orgánico para reconocer patrones y tomar decisiones. Consisten en neuronas artificiales[1] conectadas, que procesan datos de entrada para producir valores de salida.

- *Máquinas de Soporte vectorial.* Modelos de aprendizaje que clasifican datos al crear un hiperplano que maximiza la distancia entre las clases.

- *Random Forest:* Modelo de aprendizaje que utiliza múltiples árboles de decisión para alcanzar un solo resultado.

- *Aprendizaje por refuerzo:* Modelo de aprendizaje automático en el que un agente aprende a tomar decisiones mediante la interacción con un entorno, recibiendo recompensas o penalizaciones.

- *K-means:* Modelo de agrupamiento que divide datos en k clústeres, asignando cada observación al clúster con el centroide más cercano. Los centroides se actualizan iterativamente para minimizar la variación interna de los clústeres.

- *Árboles de decisión:* Modelo de aprendizaje que clasifica datos mediante una estructura de árbol, donde cada nodo es una condición y cada hoja es un resultado. Las decisiones se toman siguiendo un camino desde la raíz hasta una hoja.

- *Algoritmos evolutivos:* Modelos que utilizan procesos inspirados en la evolución, como selección y mutación, para optimizar soluciones iterativamente.

Estos modelos son evaluados mediante métodos como la validación cruzada para asegurar su eficacia en datos nuevos [5]. La interpretación de los resultados y la selección adecuada del modelo, tal como se ve en la metodología CRISP-DM, son cruciales para obtener insights valiosos y tomar decisiones informadas.

5. Datos Masivos (Big Data)

El término **Datos Masivos** o **Big Data** es un concepto que genera confusión respecto a la Ciencia de Datos, no obstante están estrechamente relacionados, ya que ambos campos se complementan y apoyan mutuamente. Big Data se refiere al procesamiento y manejo de infraestructura para trabajar con conjuntos de datos extremadamente grandes y complejos que no pueden ser gestionados, procesados o analizados mediante las herramientas y técnicas tradicionales de gestión de datos [6][7][8][9]. Por otro lado, la Ciencia de Datos utiliza técnicas de análisis, modelos de aprendizaje automático para analizar datos que podrían ser proporcionados por Big Data.

5.1 Las tres V del Big Data

El análisis de grandes cantidades de datos es crucial para las principales empresas tecnológicas. Amazon, Microsoft, Google, IBM y Oracle definen Big Data en términos de las "Tres V": *Volumen, Velocidad y Variedad*. Amazon y Microsoft destacan cómo el crecimiento en estas áreas supera las capacidades de las bases de datos tradicionales, afectando la gestión y el análisis de datos. Google añade una cuarta "V", la *Variabilidad*, refiriéndose a cómo cambia el significado de los datos con el tiempo. IBM y Oracle también

[1] Unidad de cálculo que intenta modelar el comportamiento de una neurona natural.
Fuente: https://es.wikipedia.org/wiki/Neurona_de_McCulloch-Pitts

destacan las "Tres V" del Big Data. En el caso de IBM, además subrayan que el Big Data presenta una o más de las siguientes características: gran volumen, alta velocidad o gran variedad[2].

5.2 Componentes empleados en el manejo y procesamiento de Big Data

Para poder manejar y procesar correctamente la gran cantidad de datos involucrada en los proyectos de Big Data, es necesario contar con diferentes componentes que permitan un flujo de trabajo sencillo y eficaz. Con este objetivo, existen diferentes componentes que se utilizan comúnmente para abordar la gestión, el procesamiento, el análisis y el almacenamiento de los datos con éxito.

Estos componentes son:

- *Lago de datos:* Es un espacio de almacenamiento centralizado capaz de albergar datos estructurados y no estructurados, sin procesar, provenientes de diversas fuentes, en grandes cantidades.

- *Data Warehouse (almacén de datos):* Al igual que un lago de datos, es un espacio de almacenamiento centralizado, que puede albergar grandes cantidades de datos de diversas fuentes. A diferencia de un lago de datos, aquí los datos almacenados ya pasaron por las etapas de extracción y transformación.

- *Data Pipeline (canalización de datos):* Es un proceso mediante el cual se integran datos en bruto de diversas fuentes, los cuales son procesados aplicando una serie de transformaciones, como filtrado, enmascaramiento o agregación, con el fin de estandarizarlos.

- *Business intelligence (Inteligencia empresarial): El Business intelligence (BI)* es un conjunto de técnicas y herramientas de software que consume grandes cantidades de datos estructurados y no estructurados, para presentarlos de manera amigable y fácil de entender, empleando paneles de control, reportes y visualizaciones, permitiendo así el análisis de la información [10].

6. Herramientas de software

En Ciencia de Datos, se utilizan diversas herramientas y bibliotecas para facilitar el procesamiento y análisis de datos. Python y R son lenguajes populares debido a sus bibliotecas como Pandas, NumPy y Scikit-learn, TensorFlow y PyTorch para Python, y ggplot2 y dplyr para R. Además, plataformas de Big Data como Apache Spark y bases de datos como MongoDB y PostgreSQL son esenciales para manejar grandes volúmenes de datos. Estas herramientas permiten realizar análisis avanzados y obtener insights valiosos.

En la Figura 2 se presentan algunas de estas herramientas.

Figura 2. Herramientas utilizadas para Ciencia de Datos[3].

2 https://www.ibm.com/analytics/es/es/hadoop/big-data-analytics/
3 Fuente: https://www.decisivedge.com/blog/analytics-at-the-speed-of-open-source/

7. Conclusiones

La Ciencia de Datos se ha establecido como una disciplina esencial en la actualidad, capacitando a las organizaciones para extraer información valiosa de grandes volúmenes de datos. Este artículo ha examinado temas clave en la Ciencia de Datos, incluyendo la metodología CRISP-DM y la relevancia de la estadística descriptiva e inferencial en la interpretación y validación de resultados.

Hemos destacado la importancia de los modelos de aprendizaje automático, como la regresión lineal, las redes neuronales y las máquinas de soporte vectorial, entre otros. Además, lenguajes y herramientas como Python, R, TensorFlow y Apache Spark son fundamentales para gestionar y analizar datos masivos de manera eficaz.

Finalmente, es importante recordar que la Ciencia de Datos no se limita a la tecnología y los algoritmos; también implica entender el contexto del problema a resolver y comunicar los insights de manera efectiva. Con el avance continuo de la tecnología, la Ciencia de Datos seguirá presentando nuevas oportunidades y desafíos.

8. Referencias

[1] Vijay Kotu and Bala Deshpande. Data Science: Concepts and Practice. Elsevier Inc., Cambridge, MA, USA, 2019.

[2] Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science. Cham (CH): Springer; 2019. PMID: 31314217.

[3] Kumar, A. N., Raj, R. K., & et al. (2023). Computer science curricula 2023. ACM Press, IEEE Computer Society Press, and AAAI Press.

[4] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0: Step-by-step data mining guide,” SPSS Inc., 2000.

[5] Dangeti, P. (2017). Statistics for Machine Learning. Packt Publishing.

[6] N. Marz and J. Warren, Big Data: Principles and Best Practices of Scalable Real-Time Data Systems. Shelter Island, NY, USA: Manning Publications, 2015.

[7] V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston, MA, USA: Houghton Mifflin Harcourt, 2013

[8] Information Resources Management Association, Big Data: Concepts, Methodologies, Tools, and Applications. Hershey, PA, USA: IGI Global, 2016.

[9] Genís Roca y Albert Solana, “Big Data para directivos”, Editorial Empresa Activa, 2019.

[10] Juan Gabriel Gomila Salas, Kirill Eremenko, y otros, “Inteligencia Artificial aplicada a Negocios”, Editorial Kindle, 2023.