



Tecnológico
de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos

Daniel Otero Fadul

*Departamento de Ciencias
Escuela de Ingeniería y Ciencias*

¿Qué es Ciencia de Datos?



Figura: Según IBM, "la ciencia de datos es un enfoque multidisciplinario que permite extraer información útil de los grandes y crecientes volúmenes de datos recopilados y creados por las organizaciones actuales. La ciencia de datos abarca la preparación de los datos para su análisis y procesamiento, la realización de análisis de datos avanzados y la presentación de los resultados para revelar patrones y permitir a las partes interesadas sacar conclusiones fundamentadas".

Imagen tomada de <https://www.wired.com/story/the-matrix-code-sushi-recipe/>.

¿Por qué es importante?

En general, un científico de datos puede demostrar y comunicar el valor de los productos analíticos de la institución para ayudar a crear un proceso de toma de decisiones mejorado en los diferentes niveles de la organización, mediante el registro, la medición y el seguimiento de todas las métricas de rendimiento. Esto puede impactar a la organización de distintas maneras, algunas de estas son:

- Identificar oportunidades.
- Tomar decisiones basadas en evidencia cuantificable.
- Poner a prueba decisiones que tome la institución.
- Identificar audiencias o clientes potenciales.

¿Por qué es importante?

La ciencia de datos combina el método científico, matemáticas y estadística, programación, analítica de datos, IA, e incluso la capacidad de relatar una historia fácil de entender que muestre la información relevante que se encuentra oculta en los datos.

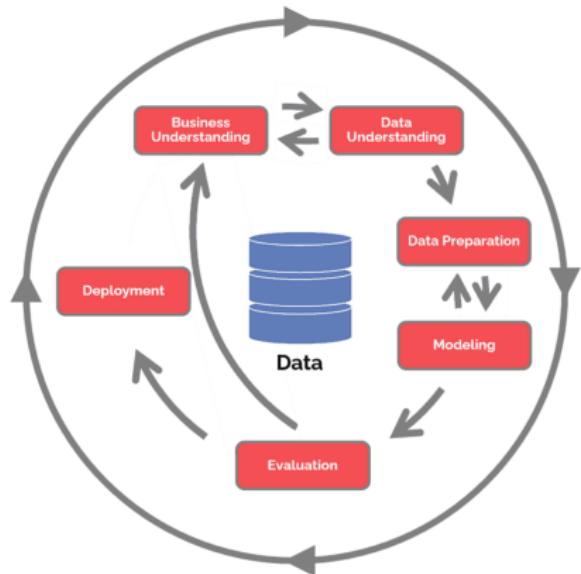


Figura: Diagrama del estándar CRISP-DM. Imagen tomada de <https://www.datascience-pm.com/crisp-dm-2/>.

El estándar “**CRoss-Industry Standard Process for Data Mining**” es una metodología que describe los enfoques típicos usados por los expertos en minería de datos. Es más, esta es una de las metodologías más empleadas para guiar el ciclo de un proyecto de ciencia de datos.

“Business Understanding”

La fase de **comprensión del negocio** se centra en entender los objetivos y requisitos del proyecto. En esta fase es común llevar a cabo las siguientes tareas:

- **Determinar los objetivos empresariales:** En primer lugar, hay que lograr un buen entendimiento de lo que el cliente realmente quiere conseguir.
- **Evaluar la situación:** Determinar la disponibilidad de recursos, los requisitos del proyecto, evaluar los riesgos y las contingencias y realizar un análisis de costes y beneficios.
- **Determinar los objetivos de la extracción de datos:** Además de definir los objetivos del proyecto, es necesario definir qué objetivos se deben alcanzar desde una perspectiva técnica de la minería de datos.
- **Elaborar el plan del proyecto:** Selección de las tecnologías y herramientas que se utilizarán y definir un plan detallado para cada fase del proyecto.

Vale la pena mencionar que esta fase es esencial.

“Data Understanding”

En la fase de comprensión de los datos se identifican, recopilan y analizan los conjuntos de datos que pueden ayudarnos a lograr los objetivos del proyecto. Esta fase puede incluir las siguientes tareas:

- **Recoger los datos iniciales:** Adquisición de los datos necesarios.
- **Describir los datos:** Examinar los datos y documentar sus propiedades superficiales: formato de los datos, el número de registros o las identidades de los campos.
- **Explorar los datos:** Analizar más detenidamente la información contenida en los datos, por ejemplo, haciendo visualizaciones e identificando relaciones entre diferentes variables.
- **Verificar la calidad de los datos:** Determinar esto es fundamental para saber hasta dónde podemos llegar con la información que tenemos.

“Data Preparation”

En esta etapa se procesan los datos para dejarlos listos para que puedan ser interpretados de manera correcta por los modelos. Las siguientes tareas son típicas de esta fase:

- **Seleccionar los datos:** Determinar qué conjuntos de datos se utilizarán y documentar los motivos de su inclusión o exclusión.
- **Limpiar los datos:** Suele ser la tarea más larga. Es convencional llevar a cabo durante esta tarea la corrección o eliminación de valores erróneos.
- **Construir datos:** Creación de nuevas variables predictoras que puedan mejorar el desempeño de los modelos.
- **Integrar datos:** Integrar datos de múltiples fuentes.
- **Formatear los datos:** Cambiar el formato de los datos según sea necesario.

“Modeling”

En esta fase se implementan varios modelos para los datos. Esta etapa tiene cuatro tareas:

- **Selección de modelos:** Determinar qué algoritmos se probarán.
- **Generar el diseño de la prueba:** División de los datos en conjuntos de entrenamiento, validación y prueba.
- **Construir el modelo:** Esta tarea consiste en la implementación y entrenamiento del modelo.
- **Evaluar el modelo:** Por lo general, varios modelos compiten entre sí, así que el científico de datos tiene que interpretar los resultados de los modelos con base a unas métricas de evaluación que se hayan definido previamente.

“Evaluation”

En esta fase se va más allá de la evaluación técnica de los modelos y se observa cómo estos se ajustan a las expectativas del cliente o de la empresa, lo cual puede determinar qué se debe hacer a continuación. Las siguientes son tareas comunes de esta etapa:

- **Evaluar los resultados:** Se determina si los modelos cumplen con los criterios de éxito del cliente o de la empresa. Es más, es necesario determinar cuál o cuáles modelos se eligen según estos criterios.
- **Revisar el proceso:** Revisión del trabajo realizado hasta este punto con el fin de saber si se ha pasado algo por alto y se han ejecutado correctamente todas las fases.
- **Determinar los siguientes pasos:** Con base en lo observado en las tareas anteriores, se debe concluir si se siguen iterando anteriores pasos o se procede a la siguiente fase.

“Deployment”

Un modelo no es especialmente útil si el cliente no puede acceder a sus resultados. La complejidad de esta fase es muy variable. Esta etapa final tiene cuatro tareas:

- **Planificar el despliegue:** Desarrollar y documentar un plan de despliegue del modelo.
- **Planificar la supervisión y el mantenimiento:** Desarrollar un plan de seguimiento y mantenimiento exhaustivo para evitar problemas durante la fase operativa (o posterior al proyecto) de un modelo.
- **Revisar el proyecto:** El equipo del proyecto documenta un resumen del proyecto que puede incluir una presentación final de los resultados.
- **Elaborar el informe final:** Se realiza una recopilación de lo realizado en el proyecto que incluya lo que ha salido bien, lo que podría haber sido mejor y sugerencias sobre cómo mejorar en el futuro.

Es posible que el trabajo del equipo no concluya aquí, puede ocurrir que se requiera una supervisión constante y un ajuste ocasional del modelo.

Moraleja: No todos los proyectos siguen la metodología CRISP-DM ya que esto está sujeto a las preferencias del científico de datos, la empresa y la naturaleza del proyecto en el que se trabaja. Algunas empresas pueden exigir que se siga un protocolo estricto, mientras que otras tienen una forma de trabajar más informal. En general, sí se necesita un enfoque estructurado cuando se trabaja en un proyecto complejo o cuando hay muchas personas o recursos involucrados.

Tipos de Datos

Hay, básicamente, tres grandes categorías para el tipo de datos con que trabaja un científico de datos:

- Estructurados.
- Semi-estructurados.
- No estructurados.

En general, cada categoría requiere herramientas diferentes para ser procesada.

Datos Estructurados

Los datos estructurados son el tipo de información que tiene una estructura bien definida, sigue un orden consistente y puede ser interpretada fácilmente por un computador o una persona. Esta información se pueden almacenar de manera sencilla en una base de datos o una tabla. “Structure Query Language”, más popularmente conocido como “SQL”, es una de las herramientas más utilizadas para manejar esta información cuando se encuentra almacenada en una base de datos. Otras herramientas como Excel también son populares, pero esta en particular tiene limitaciones en la cantidad de datos que puede manipular. En general, los datos con que normalmente se trabajan no son estructurados.

Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence II
2	214390830 Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833 Aged 18-44 years	2008	59.4%		58.0%
4	214390831 Aged 18-24 years	2008	37.4%		34.6%
5	214390832 Aged 25-44 years	2008	66.9%		65.5%
6	214390834 Aged 45-64 years	2008	88.6%		87.7%
7	214390834 Aged 45-54 years	2008	86.3%		85.1%
8	214390835 Aged 55-64 years	2008	91.5%		90.4%
9	214390840 Aged 65 years and over	2008	94.6%		93.8%
10	214390837 Aged 65-74 years	2008	93.6%		92.4%
11	214390833 Aged 75-84 years	2008	95.6%		94.4%
12	214390839 Aged 85 years and over	2008	96.0%		94.0%
13	214390841 Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842 Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843 White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844 Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845 American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846 Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847 Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848 2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figura: Una tabla de Excel. Imagen tomada de [1].

Datos Semi-estructurados

Los datos semi-estructurados son datos que tienen cierta estructura, pero carecen de un esquema fijo o rígido. Son los datos que no residen en una base de datos pero que tienen algunas propiedades que facilitan su análisis. Normalmente, con algo de procesamiento, podemos almacenarlos en la base de datos relacional. Algunos ejemplos de esta categoría son los correos, archivos .zip, páginas web, entre otros.

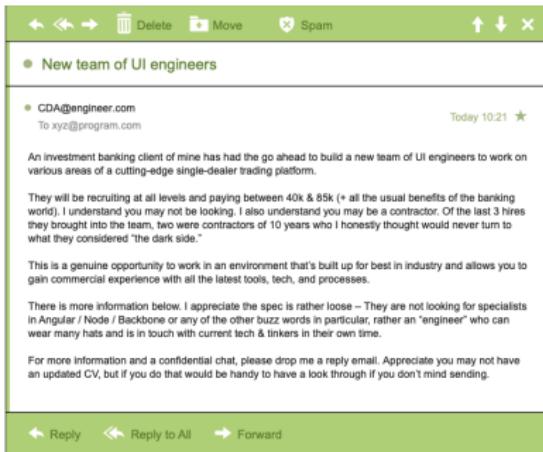


Figura: Un correo que podría enviar algún buscador de talento. Imagen tomada de [1].

Datos no estructurados

Los datos no estructurados son los datos que no tienen una estructura fácilmente identifiable, de manera que no pueden ser utilizados por un computador con facilidad. Este tipo de información no está organizada de manera predefinida, razón por la cual no se adapta a una base de datos relacional convencional. Imágenes, videos, archivos de audio son ejemplos de datos no estructurados.



Figura: Una de las famosas litografías de Escher. Imagen tomada de <https://mcescher.com/gallery/most-popular/>.

Datos Públicos

Los datos públicos, conocidos como “open data” en inglés, son grandes conjuntos de datos que son de libre acceso para cualquiera que tenga una conexión a Internet. Estos datos proceden de fuentes externas de todo el mundo. Puede tratarse de cualquier cosa, desde los datos públicos recogidos por los organismos gubernamentales, hasta los resúmenes de las tendencias económicas de los bancos y los conglomerados financieros.

Datos Públicos

Algunas de las más relevantes fuentes de datos públicos son las siguientes:

- <https://www.data.gov/> Desde la ciencia y la investigación hasta la fabricación y el clima, data.gov es una de las fuentes de datos más completas del mundo. Los conjuntos de datos están disponibles en formatos típicos como CSV, JSON y XML.
- <https://www.opendatanetwork.com/> Esta fuente permite a los usuarios buscar datos mediante un sólido motor de búsqueda. Aplique filtros avanzados a sus búsquedas y obtenga datos sobre todo tipo de temas, desde seguridad pública, finanzas, infraestructuras, vivienda y desarrollo, etc.
- <https://data.unicef.org/> Estos valiosos conjuntos de datos abiertos supervisan e informan sobre la situación de los niños y las mujeres en todo el mundo. Las últimas actualizaciones sobre brotes de enfermedades, género y educación, actitudes sobre normas sociales y otros conjuntos de datos están ampliamente disponibles a través de UNICEF, así como las visualizaciones de datos.

Datos Públicos

- <https://www.who.int/data/gho/> La Organización Mundial de la Salud, o la WHO en inglés, posee uno de los repositorios de datos abiertos más completos sobre tasas de mortalidad mundial, brotes de enfermedades, enfermedades mentales, financiación de la sanidad, etc.
- <https://scholar.google.com/> En forma de motor de búsqueda, Google Scholar permite a los usuarios buscar conjuntos de datos como se haría con cualquier otra búsqueda de Google. Encuentra fuentes de datos educativas y revisadas por expertos sobre casi cualquier tema.
- <https://www.ncdc.noaa.gov/cdo-web/datasets> Para los conjuntos de datos climáticos históricos y casi en tiempo real de todo el mundo, "Climate Data Online" (CDO) actúa como una gran fuente de datos abiertos. Busca resúmenes diarios, datos marinos, radares meteorológicos y mucho más.

Datos Públicos

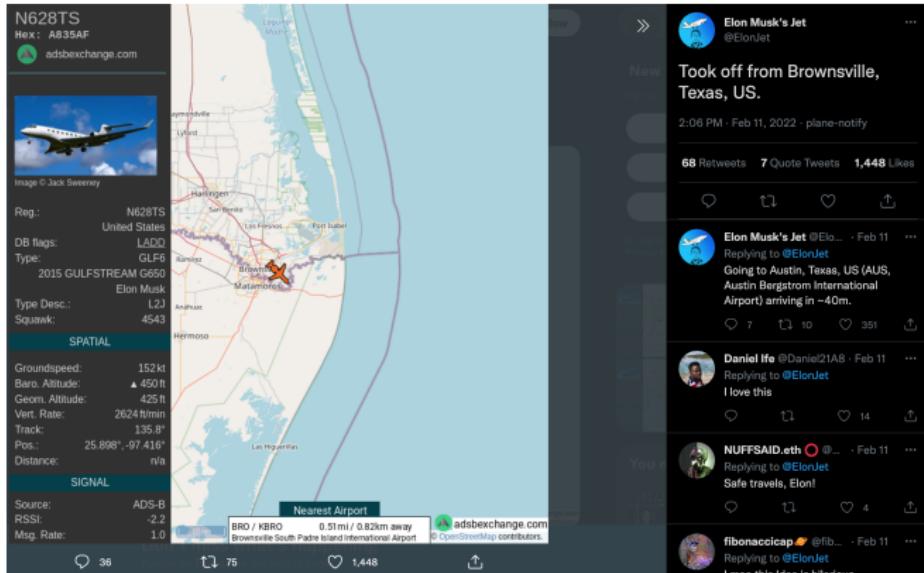


Figura: Cuenta de Twitter que le hace seguimiento del avión privado de Elon Musk (N628TS) con un “bot” usando datos públicos de ADS-B Exchange. Imagen tomada de <https://twitter.com/ElonJet/status/1492228548198809603/photo/1>.

BIBLIOGRAFÍA

- 1 D. Cielen, A. Meysman, *Introducing data science: big data, machine learning, and more, using Python tools*, Simon and Schuster, 2016.
- 2 <https://www.datascience-pm.com/crisp-dm-2/>