La importancia de los datos en la inteligencia artificial

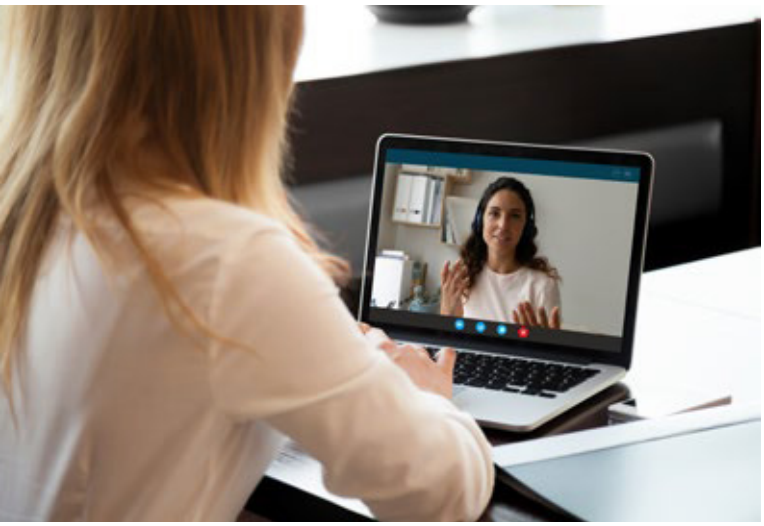
La importancia de los datos en la inteligencia artificial

Introducción

En este tema comprenderás como se generan y clasifican los datos, aprenderás cuáles son las diversas herramientas con las que se pueden gestionar los mismos y la metodología más popular que utilizan las empresas *data-drive* para obtener la información relevante que las distingue de su competencia.

Explicación

Según Domo (2020), después del año 2020 todas las estimaciones relacionadas con el crecimiento de los datos quedaron prácticamente obsoletas. Por primera vez en la historia una aplicación de videoconferencia como *Zoom*, ocupaba el 20% del tráfico total de internet, llegando a gestionar en promedio, más de 200 000 usuarios conectados por minuto. La cantidad de información prevista que un usuario iba a generar para el 2023 aumentó, dos años antes, en 4.5 veces, provocando una explosión de datos que todavía se mantiene en ascenso.



La mayor parte de la información generada se almacena de forma digital. La unidad básica de información digital es el bit, el cual no es más que un dígito binario, que puede tomar el valor de 0 y 1. El surgimiento del sistema binario se remonta al siglo III a.n.e. y fue evolucionando a lo largo del tiempo. El destacado investigador George Boole publicó, en el año 1984, un artículo donde se describían los detalles de un sistema de análisis lógico basado en los números binarios. Estos estudios posteriormente adoptarían el nombre de álgebra de Boole, cuyas bases influyeron considerablemente en el desarrollo de los circuitos electrónicos y en los sistemas de almacenamiento modernos. Normalmente, cuando queremos expresar la cantidad de datos almacenados en unidades de memoria, utilizamos el término *Byte*, el cual está conformado por 8 Bits. En la tabla 1 se muestran algunas de las unidades básicas de información, con algunos ejemplos que nos permiten visualizar de forma más clara la dimensión de estas:

Unidad	Valor	Caso de uso
1 kilobyte	1000 bytes	Documento de texto
1 megabyte	1000 kilobytes	Libro digital
1 gigabyte	1000 megabytes	230 canciones en formato MP3
1 terabyte	1000 gigabytes	500 hora en video
1 petabyte	1000 terabytes	65536 películas en 4K
1 exabyte	1000 petabytes	6900 millones de discos de música
1 zottabyte	1000 exabyte	17.200 millones de iPhones de 64 GB
1 yottabyte	1000 zottabyte	Un Centro de Datos con un tamaño equivalente a 15 veces el área de la Ciudad de Monterrey, N.L.

Tabla 1. Unidades de medida de la información

Los datos digitales se generan de diferentes formas:

- **Redes sociales**: son los datos generados en aplicaciones sociales como Facebook, Twitter, Instagram, entre otras.
- **Datos biométricos**: se obtienen a partir del reconocimiento facial, dispositivos de seguimiento y sensores de varios tipos.
- **Dispositivos inteligentes**: teléfonos y relojes inteligentes, monitores de salud y asistentes virtuales.
- **Bases de datos empresariales**: contienen información de usuarios, ventas, inventarios, incidentes y otras fuentes de información.
- **Transacciones**: datos generados por operaciones comerciales, sistemas de punto de venta, aplicaciones bancarias, entre otras.
- **Comunicación entre dispositivos**: son datos que intercambian los dispositivos entre sí de forma automática, pueden ser etiquetas, protocolos, información de sensores, entre otros.

Características de los datos

Un primer análisis de los datos se puede hacer mediante el modelo de las 5 Vs (IBM, 2020):

- **Volumen**: define la enorme cantidad de datos que se generan constantemente.
- **Variedad:** se relaciona a los tipos de datos heterogéneos que se producen (Estructurados, No Estructurados, Semi-Estructurados).
- **Velocidad**: está relacionado con la velocidad con que se generan los nuevos datos y la rapidez con que se mueven de un sistema a otro.
- **Veracidad:** se refiere a la confiabilidad de los datos y la veracidad de las fuentes que los generan.
- **Valor**: tener acceso a grandes cantidades de datos no es garantía de éxito, a menos que podamos sacar información valiosa de los mismos.

IBM (2020) indica que el 80% de los datos actuales forman parte de un segmento llamado la **Data Obscura** (*Dark Data*), lo cual significa que todavía no se conoce la forma de utilizarlos para obtener información valiosa de ellos, dejando abiertas varias áreas de oportunidad a las nuevas técnicas de inteligencia artificial que se están desarrollando para procesarla.

Variedad de datos

Según Taulli (2019) los datos se pueden clasificar en cuatro grupos principales:

Datos estructurados: En tu trabajo cotidiano, puedes encontrar datos que están etiquetados, organizados o clasificados de alguna manera. A este tipo de información se le denomina como **datos estructurados**, y la mayoría

de estos provienen de bases de datos, hojas de cálculo, información de productos, números telefónicos, datos financieros, entre otros.

Datos no estructurados: Muchos de los datos con los que interactúas, son generados de forma aleatoria y no poseen una organización u orden definido, a este conjunto se le conoce como **datos no estructurad**os se obtienen mediante diversas fuentes como: noticias, videos, imágenes, libros, canciones, blogs, entre otras. Los datos no estructurados son complejos de almacenar de forma tradicional.

Datos semiestructurados: Este tipo de datos tiene características híbridas, ya que en su contenido logra definir algún tipo de estructura, pero manejan valores diversos y representan solamente del 5% al 10% de toda la información disponible. Como ejemplos de este tipo de datos podemos mencionar: los ficheros XML (*Extensible Markup Language*), HTML (*Hyper Text Markup Language*) y o JSON (*JavaScrip Object Notation*), los cuales son muy populares para compartir información en internet a través de las API.

Datos seriados: Los datos seriados son una novedosa forma de clasificar los datos, utilizando marcas de tiempo. Este concepto va de la mano con la interconexión de sistemas y se basa en la relación que sostiene el usuario por diversas vías: páginas *web*, aplicación móvil, redes sociales o visita a tiendas físicas con una plataforma o servicio durante el transcurso del tiempo. Probablemente, en un futuro la inteligencia artificial sea la clave para procesar este tipo de información, pero en la actualidad solamente se están dando los primeros pasos.

Big Data: Es uno de los elementos principales de la inteligencia artificial y su empleo es ampliamente difundido en los sectores de finanzas, marketing, ventas, atención al cliente, entre otros. Constituye una de las mejores formas en que una empresa puede sacar beneficio de todos los datos que genera para favorecer su propio crecimiento. Entre las aplicaciones en las que se puede implementar podemos mencionar:

- Satisfacer las necesidades de los clientes de forma puntual, aumentando su nivel de satisfacción.
- Advertir intentos de fraudes financieros.
- Detectar novedosas oportunidades comerciales.
- Identificar oportunidades de reducción de costos de procesos.
- Tomar decisiones más asertivas.

Según Silva (2021), para comprender los detalles del Big Data es fundamental identificar los 4 tipos de análisis que se pueden implementar con este:

- **Análisis descriptivo**: su propósito es aportar información sobre la situación actual a partir de datos de estado anteriores. En forma general, este tipo de análisis apoya la selección de las decisiones que deben tomarse en tiempo real.
- **Análisis diagnóstico**: este tipo de estudio permite examinar los resultados y progreso de determinadas acciones. Gracias a esto, es posible realizar modificaciones a las estrategias que se están implementando.

- **Análisis predictivo**: también conocido como ciencia de datos, el análisis predictivo realiza un pronóstico sobre posibilidades futuras, apoyándose en patrones encontrados en los datos que fueron analizados.
- **Análisis prescriptivo**: el objetivo del análisis prescriptivo es mostrar las posibles consecuencias que cada operación puede generar para el negocio. Esto ayuda a distinguir las estrategias óptimas, que generarán nuevos y mejores resultados para la compañía.

Dichos análisis utilizan modelos estadísticos basados en técnicas tales como regresión logística, regresión lineal, redes neuronales y sistemas de decisión de árbol. Estas son algunas de las técnicas que son utilizadas en los modelos para los algoritmos de la inteligencia artificial.

Manejar tanta información es un tema complejo, por tanto, en la actualidad solo una pequeña parte de las compañías puede trabajar directamente con ella. Gracias al aumento del poder de cómputo actual y a las iniciativas de grandes compañías tecnológicas como IBM, la facilidad de acceder a este tipo de herramientas es cada vez más sencilla, poniendo en nuestras manos la increíble capacidad que posee la **Big Data**.

Herramientas para trabajar con datos

Bases de datos y procesamiento distribuido

Una tecnología que se ha mantenido en evolución durante décadas para el trabajo con datos son las denominadas **bases de datos**, de las cuales incluso, algunas de las primeras variantes han perdurado en el tiempo y todavía se utilizan. Los principios de las bases de datos se remontan a 1963, cuando el termino fue acuñado por primera vez, pero no fue hasta la década de los setenta con los trabajos de Edgar Frank Codd, que se presenta un punto de quiebre en la línea de investigación que se venía desarrollando, dando pie a la introducción de la estructura de **Modelo Relacional** que conocemos en la actualidad.

Durante más de 20 años las **bases de datos relacionales** reinaron como el estándar en las plataformas cliente – servidor, pero con el advenimiento de Big Data varios factores tecnológicos influyeron de forma significativa en la disminución de su desempeño:

- Aumento exponencial en la cantidad de datos.
- Surgimiento de nuevos entornos de desarrollo.
- Altos costes de mantenimiento.
- Desafíos de implementación.

Estas condiciones prepararon el terreno para el desarrollo de otras tecnologías que pudieran solucionar estos problemas con los datos.

Almacén de datos

Un almacén de datos se puede considerar como un escalón superior en el procesamiento de datos estructurados y es esencial en los procesos de analítica y toma de decisiones, en especial para los

proyectos de inteligencia empresarial.

Lagos de datos

Cuando abordamos el tema de las grandes cantidades de **datos no estructurados** que manejan las empresas e incluso la Big Data, una óptima forma de representar su sistema de almacenamiento es con la introducción del concepto de los **lagos de datos**.

Los lagos de datos requieren un mantenimiento permanente y un plan que permita definir el acceso a los datos y su uso. Sin este mantenimiento, se corre el riesgo de que los datos se tornen inaccesibles, muy difíciles de manejar, costosos e inservibles. Los lagos de datos a los que los usuarios no pueden entrar se les conoce como "pantanos de datos" (IBM, 2020).

Análisis de los datos en el diseño de sistemas inteligentes

Una vez recolectados los datos, es necesario realizar el procesamiento de estos. Muchas empresas invierten en Big Data y soluciones de analítica, pensando incorrectamente que esto solucionará de manera directa sus problemas. En ocasiones, gran gasto no significa buenos resultados. Estudios estiman que 85% de los proyectos basados en datos, son abandonados en las primeras etapas.

Las razones más comunes son:

- Análisis desenfocado.
- Datos contaminados.
- Inversión en herramientas equivocadas.
- Problemas con la recolección.
- Desmotivación o agotamiento de los desarrolladores y directivos.

La mejor forma de enfrentar este tipo de situaciones radica en la aplicación de técnicas para procesamiento de datos; estas metodologías estas constituyen los fundamentos de la **minería de datos** y forman una simbiosis perfecta entre la **ciencia de datos** y la **inteligencia artificial**.

Cierre

La inteligencia artificial se apoya en la ciencia de datos para realizar la comprensión, el razonamiento y el aprendizaje de las soluciones a los problemas. Tener muchos datos no garantiza un mejor resultado, pues todo depende del tipo de modelo y de la solución que desee encontrar, pero recíprocamente, no hay forma de generar un buen modelo sin la cantidad suficiente de datos.

Checkpoint

Asegúrate de:

- Distinguir las fuentes de generación de datos.
- Comprender el modelo de las 5 vs.
- Diferenciar los tipos de datos que existen.
- Comprender la definición de Big data y Ciencia de datos.
- Determinar las fases que componen el modelo CRISP-DM.

Referencias bibliográficas

- DOMO. (2020). *Learn Center / Data Never Sleeps 8.0*. Recuperado de https://www.domo.com/learn/data-never-sleeps-8
- IBM (2020). *Lecture 1 - Data Science Landscape Notes*. Recuperado de https://skills-academy.comprehend.ibm.com/
- Silva, D. (2021). *¿Qué es el Big Data y para qué sirve?* Recuperado de https://www.zendesk.com.mx/blog/big-data-que-es/
- Taulli T. (2019). *Artificial Intelligence Basics: A Non-Technical Introduction*. Estados Unidos: Apress.