Modelamiento y gestión de datos para modelos de inteligencia artificial

Breve descripción:

Este componente ofrece una introducción comprehensiva al modelamiento y gestión de datos para sistemas de inteligencia artificial. Abarca los fundamentos de estructuras de datos, técnicas de calidad y tratamiento, sistemas de gestión de bases de datos y conceptos básicos de IA. Proporciona las bases teóricas necesarias para comprender cómo los datos se transforman en información valiosa para modelos de IA.

Noviembre 2024

Tabla de contenido

Introducción

En la era de la inteligencia artificial, los datos se han convertido en uno de los recursos más valiosos para las organizaciones. Sin embargo, para que estos datos sean verdaderamente útiles, necesitan ser adecuadamente modelados, gestionados y procesados. La diferencia entre un proyecto de IA exitoso y uno fallido a menudo radica en la calidad y estructura de los datos que lo alimentan.

Este componente formativo aborda los principios esenciales del modelamiento y gestión de datos en el contexto de la inteligencia artificial. Comienza con los conceptos básicos de estructuras de datos, seguido por técnicas de calidad y tratamiento de datos. Se exploran los sistemas de gestión de bases de datos y, finalmente, se introduce la inteligencia artificial y sus requerimientos de datos.

A lo largo del componente, el aprendiz descubrirá cómo las diferentes piezas del rompecabezas de datos se unen para formar la base de sistemas de IA efectivos. Mediante referencias de estudio, se explorarán las mejores prácticas en el modelamiento y gestión de datos, así herramientas y tecnologías relevantes en el campo.

La comprensión profunda de estos fundamentos es muy importante para toda persona que aspire a trabajar en el campo de la IA. Como dice el refrán en ciencia de datos: "Los modelos son tan buenos como los datos que los alimentan".

¡Bienvenidos a este viaje por los fundamentos del modelamiento y gestión de datos para IA!

Video 1. Modelamiento y gestión de datos para modelos de inteligencia artificial

Enlace de reproducción del video

Síntesis del video: Modelamiento y gestión de datos para modelos de inteligencia artificial
El componente formativo "Modelamiento y gestión de datos para modelos de inteligencia artificial" explora los fundamentos esenciales para trabajar con datos en el contexto de la IA moderna. La calidad y estructura de los datos son esenciales para el éxito de cualquier proyecto de IA. Se estima que los científicos de datos dedican hasta el 70 % u 80 % de su tiempo a tareas relacionadas con la preparación y gestión de datos.

Comenzaremos explorando los fundamentos de las estructuras de datos, desde tipos básicos hasta estructuras complejas. Comprenderemos cómo diferentes estructuras sirven para diferentes propósitos en el procesamiento de datos.

La calidad de datos es nuestro siguiente pilar, donde aprenderemos sobre dimensiones de calidad, técnicas de extracción y métodos de transformación. Un dato limpio y bien estructurado es la base de cualquier modelo de IA exitoso.

En gestión de bases de datos, abordaremos tanto sistemas relacionales como NoSQL, entendiendo cuándo usar cada uno y cómo optimizar su rendimiento para aplicaciones de IA.

Finalmente, nos introduciremos en conceptos básicos de IA, explorando cómo los datos bien gestionados alimentan estos sistemas inteligentes.

Las tendencias actuales apuntan hacia una integración cada vez más estrecha entre la gestión de datos y la IA, con herramientas que facilitan el procesamiento automatizado y el análisis en tiempo real.

¡Les damos la bienvenida al fascinante mundo del modelamiento y gestión de datos para IA!

1. Fundamentos de datos y estructuras

En el contexto actual de la revolución digital, comprender los fundamentos de datos y estructuras se ha vuelto esencial para cualquier persona que se desempeñe en el campo de la tecnología. Este capítulo explora los conceptos básicos que fundamentan el trabajo con datos, desde su naturaleza más elemental hasta las estructuras complejas que permiten su organización y manipulación eficiente.

1.1. Conceptos básicos de datos y análisis

En el mundo actual, los datos se han convertido en el activo más valioso para las organizaciones y la base fundamental para el desarrollo de modelos de inteligencia artificial. Pero ¿qué son realmente los datos? En su forma más básica, los datos son representaciones de hechos, observaciones o mediciones que, por sí solos, carecen de contexto o significado. Es a través del análisis y la transformación que estos datos en bruto se convierten en información valiosa que permite tomar decisiones informadas.

Cuando se analizan los datos en el contexto de la inteligencia artificial, se debe considerar que estos pueden presentarse de diversas formas. Los datos cuantitativos son aquellos que se pueden medir y expresar numéricamente. Ejemplos incluyen la temperatura diaria de una ciudad, el precio de un producto o el número de usuarios que visitan un sitio web. Estos datos permiten realizar cálculos matemáticos y análisis estadísticos precisos.

Por otro lado, los datos cualitativos describen características o cualidades que no pueden medirse numéricamente. Las opiniones de los clientes sobre un servicio, los colores preferidos en un estudio de mercado, o las respuestas a preguntas abiertas en una encuesta son ejemplos de datos cualitativos. Aunque estos datos no son numéricos

por naturaleza, son igualmente valiosos y pueden transformarse en información cuantificable mediante técnicas de procesamiento adecuadas.

Fundamentos de los datos y su análisis en inteligencia artificial

a) **Introducción al análisis de datos**

Los datos son representaciones de hechos, observaciones o mediciones que requieren análisis para convertirse en información útil. En inteligencia artificial, los datos se usan para entrenar algoritmos y hacer predicciones.

b) **¿Qué son los datos?**

Los datos son hechos u observaciones sin contexto. A través del análisis, se transforman en información significativa que guía decisiones en múltiples campos.

c) **Datos cuantitativos: definición y ejemplos**

Los datos cuantitativos son numéricos y medibles. Ejemplos:

- Temperatura diaria
- Precio de productos
- Número de usuarios en un sitio web

Permiten cálculos y análisis estadísticos.

d) **Datos cualitativos: definición y ejemplos**

Los datos cualitativos describen características no numéricas. Ejemplos:

- Opiniones de clientes
- Colores preferidos en encuestas
- Respuestas abiertas

Son útiles para obtener contexto y enriquecer el análisis.

e) **Diferencias clave entre datos cuantitativos y cualitativos**

- **Cuantitativos**: son numéricos y se analizan matemáticamente.
- **Cualitativos**: son descriptivos y no numéricos. Capturan aspectos subjetivos.

Ambos son cruciales para el análisis en inteligencia artificial.

f) **Transformación de datos cualitativos a cuantitativos**

Los datos cualitativos se pueden convertir en datos cuantitativos mediante codificación o análisis de sentimiento. Ejemplo: una opinión positiva se transforma en un puntaje.

g) **Aplicaciones en inteligencia artificial**

Los datos cuantitativos permiten crear modelos predictivos, mientras que los cualitativos aportan contexto y comprensión emocional. Ambos tipos son necesarios para entrenar algoritmos en IA.

h) **Resumen**

- **Cuantitativos**: son numéricos, útiles para análisis matemáticos.
- **Cualitativos**: describen características y enriquecen el análisis contextual.

Ambos son cruciales para el análisis en inteligencia artificial.

1.2. Técnicas de modelado de datos

El modelado de datos es un proceso fundamental que permite crear una representación abstracta de cómo se relacionan y organizan los datos en un sistema. Imaginen que están construyendo una casa: antes de comenzar la construcción, necesitan planos detallados que muestren cómo se conectarán las diferentes habitaciones y qué función cumplirá cada espacio. De manera similar, el modelado de

datos nos proporciona un “plano” de cómo se estructurará y fluirá la información en nuestros sistemas.

El proceso de modelado comienza con una visión conceptual de alto nivel, donde se identifican las principales entidades o conceptos sobre los que se necesita almacenar información y cómo se relacionan entre sí. Por ejemplo, en un sistema de comercio electrónico, las entidades principales podrían ser “Cliente”, “Producto” y “Pedido”. Cada una de estas entidades tendrá sus propias características o atributos, y establecerá relaciones con otras entidades del sistema.

A medida que se avanza en el modelado, se llega a un nivel más detallado en el que se define cómo se implementarán estas estructuras en los sistemas de almacenamiento de datos. En este punto, conceptos como la normalización cobran importancia, ya que permiten organizar los datos de manera eficiente y eliminar redundancias innecesarias.

1.3. Estructuras de datos fundamentales

Las estructuras de datos son la columna vertebral de cualquier sistema de procesamiento de información. Son formas específicas de organizar y almacenar datos que nos permiten acceder a ellos y modificarlos de manera eficiente. La elección de la estructura de datos correcta puede tener un impacto significativo en el rendimiento de nuestros algoritmos de inteligencia artificial.

Para comprender mejor las diferentes estructuras de datos y sus aplicaciones, consideremos la siguiente tabla comparativa:

Tabla 1. Principales estructuras de datos, casos de uso y limitaciones.

Estructura	Características principales	Casos de uso ideales	Limitaciones
Arreglos.	Acceso directo a elementos, tamaño fijo.	Datos secuenciales, operaciones de búsqueda frecuentes.	Tamaño inmutable, inserción/eliminación costosa.
Listas enlazadas.	Tamaño dinámico, inserción eficiente.	Datos que cambian frecuentemente.	Acceso secuencial.
Árboles.	Organización jerárquica, búsqueda eficiente.	Datos jerárquicos, índices.	Complejidad de implementación.
Hash tables.	Búsqueda rápida por clave.	Cachés, diccionarios.	Colisiones, uso de memoria.

Fuente. OIT, 2024.

Los arreglos, por ejemplo, son una de las estructuras más básicas y ampliamente utilizadas. Imaginemos un array como una fila de casilleros numerados: cada casillero puede contener un dato, y podemos acceder directamente a cualquier casillero conociendo su número. Esta estructura es ideal cuando conocemos de antemano el tamaño de nuestros datos y necesitamos acceder a elementos específicos rápidamente.

Las listas enlazadas, por otro lado, son como una cadena de elementos donde cada uno conoce la ubicación del siguiente. Esta estructura es más flexible que los arreglos, ya que podemos agregar o eliminar elementos fácilmente, pero requiere más tiempo para encontrar un elemento específico, ya que debemos recorrer la lista desde el principio.

Los árboles y las hash tables son estructuras más complejas que nos permiten representar relaciones jerárquicas y conexiones entre datos. Estas estructuras son

fundamentales en muchos algoritmos de inteligencia artificial, especialmente en áreas como el procesamiento del lenguaje natural y el análisis de redes sociales.

La elección de la estructura de datos adecuada dependerá de varios factores, incluyendo el tipo de operaciones que necesitamos realizar con mayor frecuencia, el volumen de datos que manejaremos, y los requisitos de rendimiento de nuestra aplicación. Es importante recordar que no existe una estructura "perfecta" que sirva para todos los casos: cada una tiene sus propias ventajas y desventajas que debemos considerar cuidadosamente.

Al desarrollar soluciones de inteligencia artificial, la comprensión profunda de estas estructuras de datos fundamentales nos permite diseñar sistemas más eficientes y escalables. Un buen diseño de datos desde el principio puede significar la diferencia entre un sistema que funciona adecuadamente y uno que se vuelve inmanejable a medida que crece el volumen de datos.

Aplicaciones reales de estructuras de datos en la industria

a) **Arreglos: uso en la industria**

   Los arreglos se utilizan en industrias que requieren acceso rápido a datos secuenciales, como en sistemas de procesamiento de pagos o monitoreo de inventarios en tiempo real.

b) **Arreglos en procesamiento de pagos**

   En sistemas de pago, los arreglos permiten almacenar y acceder rápidamente a transacciones o registros de clientes por ID de manera eficiente.

c) **Listas enlazadas: uso en la industria**

Las listas enlazadas son útiles en aplicaciones donde los datos cambian frecuentemente, como en bases de datos dinámicas y sistemas de gestión de tareas.

d) **Listas enlazadas en bases de datos dinámicas**

Las listas enlazadas permiten la inserción y eliminación rápida de registros, como en aplicaciones de bases de datos de clientes donde los datos están en constante cambio.

e) **Árboles: uso en la industria**

Los árboles se utilizan en aplicaciones que requieren organizar datos jerárquicos, como en la gestión de archivos o en algoritmos de búsqueda.

f) **Árboles en gestión de archivos**

Los árboles permiten organizar directorios y archivos de manera eficiente, facilitando el acceso rápido a documentos en sistemas de almacenamiento.

g) **Hash tables: uso en la industria**

Las tablas hash son fundamentales en aplicaciones que requieren búsqueda rápida por clave, como en sistemas de caché o diccionarios electrónicos.

h) **Hash tables en caché y diccionarios**

Se usan en sistemas de caché de alto rendimiento y en diccionarios electrónicos para almacenar y buscar rápidamente valores asociados a claves.

i) Elección de la estructura adecuada

La elección de la estructura de datos depende de factores como el tipo de operación, el volumen de datos y los requisitos de rendimiento. No hay una solución única.

j) Resumen

- Arreglos: uso en datos secuenciales y accesos rápidos.
- Listas enlazadas: adecuadas para datos dinámicos.
- Árboles: útiles para organizar datos jerárquicos.
- Hash tables: eficaces para búsquedas rápidas por clave.

La selección de la estructura correcta optimiza el rendimiento según el caso de uso.

2. Calidad y tratamiento de datos

La calidad y el tratamiento adecuado de los datos son pilares fundamentales para el éxito de cualquier proyecto de análisis o inteligencia artificial. En esta sección, se abordan las metodologías y técnicas necesarias para garantizar que los datos con los que trabajamos sean confiables, consistentes y útiles para los objetivos planteados.

2.1. Características de calidad de datos

La calidad de los datos es un aspecto fundamental en el desarrollo de modelos de inteligencia artificial. Así como un chef no puede preparar un platillo exquisito con ingredientes en mal estado, un modelo de IA no puede proporcionar resultados confiables si se alimenta con datos de baja calidad. La famosa frase "**garbage in, garbage out**" (entra basura, sale basura) resume perfectamente esta realidad en el campo del análisis de datos.

La calidad de los datos se puede evaluar a través de múltiples dimensiones. La precisión es quizás la más evidente: los datos deben representar fielmente la realidad que intentan capturar. Por ejemplo, si estamos registrando la temperatura de un proceso industrial, necesitamos asegurarnos de que nuestros sensores estén correctamente calibrados y que las mediciones sean exactas.

La completitud es otra dimensión estratégica. Los datos incompletos pueden llevar a conclusiones sesgadas o incorrectas. Imaginemos un estudio de satisfacción del cliente donde solo tenemos respuestas de usuarios muy satisfechos o insatisfechos: esto nos daría una visión distorsionada de la realidad, pues faltaría la información de aquellos clientes con opiniones moderadas.

La consistencia de los datos implica que la información sea coherente a través de diferentes sistemas y registros. Por ejemplo, si un cliente aparece con diferentes direcciones en distintas bases de datos de la empresa, ¿cuál es la correcta? La falta de consistencia puede generar confusión y errores en el procesamiento de la información.

Cinco recomendaciones para garantizar datos de calidad

a) **Precisión en los datos**

   Los datos deben reflejar con exactitud la realidad. Es fundamental utilizar herramientas de calibración y validación para asegurar que las mediciones y registros sean correctos.

b) **Completitud de los datos**

   Los datos incompletos pueden generar conclusiones sesgadas. Asegurarse de registrar toda la información necesaria previene la distorsión en los análisis.

c) **Consistencia entre sistemas**

   La coherencia de los datos debe ser garantizada entre diferentes sistemas y bases de datos. Es esencial realizar procesos de limpieza para evitar errores y discrepancias.

d) **Validación continua de datos**

   Implementar mecanismos automáticos de validación garantiza que los datos nuevos ingresados sigan los estándares de calidad definidos.

e) **Contextualización de los datos**

   Los datos deben ser acompañados del contexto adecuado (fecha, ubicación, evento) para ser realmente útiles en el análisis y la toma de decisiones.

f) Resumen

- Asegurar precisión en los datos.
- Garantizar la completitud de la información.
- Mantener consistencia entre sistemas.
- Validar los datos de forma continua.
- Proporcionar contexto para su correcta interpretación.

La calidad de los datos garantiza el éxito de los modelos de inteligencia artificial y genera análisis rigurosos tanto para información simple como compleja.

2.2. Técnicas de extracción y filtrado

La extracción de datos es como la minería: debemos saber dónde buscar y qué herramientas utilizar para obtener la información valiosa. Este proceso involucra la identificación y recuperación de datos relevantes de diversas fuentes, que pueden incluir bases de datos relacionales, archivos de texto, hojas de cálculo, páginas web o incluso sensores IoT.

Una de las técnicas más comunes de extracción es el web scraping, que permite obtener información estructurada de páginas web. Sin embargo, esta técnica debe utilizarse con responsabilidad, respetando los términos de servicio de los sitios web y las políticas de privacidad aplicables.

El filtrado de datos es el proceso de eliminar o modificar datos que no cumplen con ciertos criterios de calidad o relevancia. Podemos clasificar las técnicas de filtrado en varias categorías según su propósito:

Tabla 2. Tipos de filtrado de datos, objetivos, técnicas comunes y consideraciones.

Tipo de Filtrado	Objetivo	Técnicas Comunes	Consideraciones
Limpieza.	Eliminar errores y anomalías.	Detección de outliers, corrección de formatos.	Puede requerir validación manual.
Reducción.	Disminuir el volumen de datos.	Muestreo, agregación.	Riesgo de pérdida de información.
Transformación.	Convertir datos a formato útil.	Normalización, codificación.	Debe mantener la integridad de los datos.
Enriquecimiento.	Agregar información adicional.	Joins, lookup tables.	Puede aumentar la complejidad.

Fuente. OIT, 2024.

2.3. Transformación e integración de datos

La transformación de datos es como traducir un texto a diferentes idiomas: el significado debe mantenerse, aunque la forma cambie. Este proceso implica convertir los datos de su formato original a uno que sea más adecuado para el análisis o el modelado. Las transformaciones pueden ser tan simples como cambiar el formato de una fecha o tan complejas como aplicar cálculos matemáticos avanzados.

Una transformación común es la normalización, que ajusta los valores numéricos a una escala común. Por ejemplo, si tenemos datos de ventas en diferentes monedas, necesitaremos convertirlos todos a una moneda común para poder compararlos adecuadamente. Otro ejemplo sería la estandarización de textos, asegurando que todas las cadenas de caracteres sigan el mismo formato (mayúsculas/minúsculas, tratamiento de espacios, etc.).

La integración de datos es el proceso de combinar datos de diferentes fuentes en una vista unificada y coherente. Este proceso puede ser particularmente desafiante cuando las fuentes tienen diferentes estructuras, formatos o niveles de calidad. Es como armar un rompecabezas donde las piezas provienen de diferentes cajas: necesitamos encontrar la forma de que encajen correctamente.

Un aspecto clave de la integración es la resolución de conflictos. ¿Qué hacemos cuando diferentes fuentes proporcionan información contradictoria sobre el mismo elemento? Por ejemplo, si un cliente aparece con diferentes números de teléfono en distintas bases de datos, necesitamos establecer reglas claras para determinar cuál es la información correcta o más actualizada.

La integración también debe considerar la temporalidad de los datos. Los datos históricos pueden ser valiosos para identificar tendencias y patrones, pero deben integrarse de manera que mantengan su contexto temporal. Por ejemplo, al analizar el rendimiento de ventas, necesitamos asegurarnos de que estamos comparando períodos equivalentes y considerando factores estacionales.

El éxito en la transformación e integración de datos requiere un equilibrio entre la automatización y el criterio humano. Si bien muchos procesos pueden automatizarse, el conocimiento del dominio y el juicio experto son fundamentales para garantizar que las transformaciones sean significativas y que la integración produzca resultados útiles para el análisis posterior.

Errores comunes y riesgos de datos contradictorios

- **Confusión en la toma de decisiones**

  Los datos contradictorios pueden llevar a tomar decisiones erróneas, ya que diferentes fuentes de información proporcionan resultados opuestos. Esto afecta la confiabilidad de los análisis y puede generar estrategias mal orientadas.

- **Impacto en el análisis de tendencias**

  Cuando los datos contienen contradicciones, los modelos predictivos y el análisis de tendencias se ven gravemente afectados. Esto puede dar lugar a conclusiones erróneas sobre comportamientos futuros o el desempeño de una estrategia.

- **Problemas en la integración de datos**

  La combinación de datos provenientes de diferentes fuentes, sin un proceso de validación y reconciliación adecuado, puede generar incoherencias. Esto es especialmente crítico cuando se integran bases de datos de sistemas no compatibles o mal gestionados.

- **Desconfianza de los usuarios**

  La presencia de datos contradictorios puede minar la confianza de los usuarios o stakeholders en los resultados obtenidos. La falta de confianza afecta la aceptación y el uso de los sistemas de inteligencia empresarial o modelos de IA.

- **Dificultad en la escalabilidad**

  A medida que los volúmenes de datos crecen, las contradicciones en los datos se vuelven más difíciles de detectar. Esto complica la escalabilidad de

los sistemas de análisis y puede generar errores en el procesamiento a gran escala.

Cómo mitigar estos problemas

Para evitar que los datos contradictorios afecten el análisis y la toma de decisiones, es necesario implementar procesos robustos de validación, limpieza y unificación de datos. Asegurar la consistencia y la verificación continua es clave para obtener resultados precisos y confiables.

3. Gestión de bases de datos

La gestión efectiva de bases de datos representa un componente central en la infraestructura de cualquier sistema de información moderno. En este capítulo, se explorarán los diferentes tipos de bases de datos, sus características distintivas y las mejores prácticas para su administración y optimización.

3.1. Bases de datos relacionales y no relacionales

En el ecosistema moderno de gestión de datos, nos encontramos principalmente con dos grandes familias de bases de datos: las relacionales y las no relacionales (NoSQL). Cada una tiene sus propias fortalezas y casos de uso ideales, y comprender sus diferencias es fundamental para tomar decisiones acertadas en el diseño de sistemas de datos.

Para visualizar mejor las diferencias fundamentales entre estos dos tipos de bases de datos, observemos la siguiente representación gráfica. Como se ilustra en la infografía, mientras las bases de datos relacionales se caracterizan por su estructura rígida y garantías ACID, las bases de datos NoSQL destacan por su flexibilidad y escalabilidad. Esta distinción es muy importante al momento de seleccionar el tipo de base de datos más adecuado para nuestras necesidades específicas. Por ejemplo, si necesitamos mantener la integridad referencial en transacciones financieras, una base de datos relacional sería la elección óptima. En cambio, si estamos manejando datos de redes sociales con estructuras variables y alto volumen de escrituras, una base de datos NoSQL podría ser más apropiada.

Figura 1. Tipos de bases de datos

Fuente. OIT, 2024.

Ahora bien, las bases de datos relacionales han sido el pilar de la gestión de datos durante décadas. Se basan en el modelo relacional, donde los datos se organizan en tablas con filas y columnas claramente definidas. La fortaleza de este modelo radica en su capacidad para mantener la integridad de los datos y establecer relaciones claras entre diferentes entidades. Cuando necesitamos garantizar la consistencia de las transacciones y mantener relaciones complejas entre datos, las bases de datos relacionales son la opción más segura.

Por otro lado, las bases de datos NoSQL surgieron como respuesta a las necesidades de la era del big data y la web moderna. Ofrecen mayor flexibilidad en cuanto a la estructura de los datos y mejor escalabilidad horizontal. No requieren un esquema fijo, lo que las hace ideales para datos semi-estructurados o cuando la estructura de los datos puede cambiar con frecuencia.

3.2. Operaciones CRUD y consultas básicas

Las operaciones CRUD (Create, Read, Update, Delete) son de alta relevancia para interactuar con los datos almacenados en cualquier sistema de base de datos. La implementación de estas operaciones varía según el tipo de base de datos, pero los principios básicos permanecen constantes.

En el contexto de bases de datos relacionales, estas operaciones se realizan mediante SQL (Structured Query Language). Las consultas SELECT permiten recuperar datos específicos, mientras que INSERT, UPDATE y DELETE se utilizan para modificar los registros existentes. La potencia de SQL radica en su capacidad para realizar consultas complejas que combinan datos de múltiples tablas y aplican criterios de filtrado y agregación.

Para las bases de datos NoSQL, las operaciones CRUD se implementan de manera diferente según el tipo específico de base de datos. Por ejemplo, en una base de datos documental como MongoDB, se trabaja con documentos JSON que pueden tener estructura variable, y las operaciones se realizan mediante métodos específicos de la API de la base de datos.

3.3. Gestión y optimización de bases de datos

La gestión efectiva de bases de datos va más allá de simplemente almacenar y recuperar información. Implica un conjunto de prácticas y técnicas para asegurar el rendimiento, la disponibilidad y la integridad de los datos. La optimización es un proceso continuo que debe adaptarse a las cambiantes necesidades del sistema.

Un aspecto esencial de la optimización es el diseño e implementación de índices. Los índices son estructuras de datos adicionales que mejoran la velocidad de las operaciones de búsqueda, pero tienen un costo en términos de espacio de almacenamiento y rendimiento de escritura. La decisión de qué campos indexar debe basarse en un análisis cuidadoso de los patrones de acceso a los datos.

El monitoreo del rendimiento es una parte fundamental de la gestión de bases de datos. Es necesario prestar atención a métricas como el tiempo de respuesta de las consultas, el uso de recursos del sistema y los patrones de acceso a los datos. Las herramientas de monitoreo identifican cuellos de botella y oportunidades de optimización.

La seguridad es otro aspecto que no puede descuidarse. Esto incluye la gestión de usuarios y permisos, la encriptación de datos sensibles y la implementación de políticas de respaldo y recuperación. En el contexto de la inteligencia artificial, donde frecuentemente se trabaja con datos sensibles o personales, la seguridad adquiere una importancia aún mayor.

La escalabilidad es un factor clave en la gestión de grandes volúmenes de datos. Los sistemas de bases de datos deben ser diseñados para adaptarse al crecimiento futuro, ya sea a través de escalabilidad vertical (ampliando los recursos de un único

servidor) o horizontal (distribuyendo la carga entre varios servidores). Las estrategias de particionamiento y replicación son fundamentales para manejar grandes cantidades de datos de forma eficiente.

Cinco pasos para crear índices eficientes en bases de datos

- **Identificar los campos más consultados**

  Para crear índices efectivos, es necesario analizar las consultas frecuentes y determinar qué campos son utilizados con mayor regularidad en los filtros y las cláusulas JOIN. Esto permite enfocarse en las columnas que realmente mejorarán el rendimiento de las consultas.

- **Evaluar el costo de actualización**

  Aunque los índices aceleran las búsquedas, también pueden ralentizar las operaciones de inserción, actualización y eliminación. Es importante equilibrar el beneficio de la velocidad de lectura con el costo adicional en las operaciones de escritura, priorizando índices en columnas que se consultan más que las que se actualizan con frecuencia.

- **Seleccionar el tipo adecuado de índice**

  Dependiendo de las necesidades del sistema, se deben elegir entre diferentes tipos de índices, como los índices B-tree, hash, o índices compuestos. La elección debe basarse en la estructura de los datos y las operaciones de búsqueda que se realizarán con mayor frecuencia.

- **Monitorear el uso de índices**

  Es fundamental realizar un seguimiento del uso de los índices en las consultas. Algunos índices pueden volverse innecesarios con el tiempo, especialmente si las consultas o patrones de acceso cambian. Monitorear

su rendimiento permite eliminar índices redundantes y mantener la base de datos optimizada.

- **Considerar la fragmentación de índices**

  A medida que se insertan, actualizan o eliminan datos, los índices pueden fragmentarse, lo que reduce su eficiencia. Se debe realizar un mantenimiento periódico para reorganizar o reconstruir los índices y asegurarse de que sigan funcionando de manera óptima.

Estas prácticas permiten crear índices bien diseñados que no solo optimizan las consultas, sino que también equilibran el uso de recursos y el rendimiento general de la base de datos.

4. Introducción a la Inteligencia Artificial

La Inteligencia Artificial ha emergido como una de las tecnologías más transformadoras del siglo XXI, revolucionando la manera en que procesamos información y resolvemos problemas complejos. Este capítulo proporciona una visión general de los conceptos fundamentales de la IA, sus aplicaciones prácticas y las herramientas esenciales para su implementación.

4.1. Conceptos fundamentales de IA

La Inteligencia Artificial representa uno de los avances más significativos en la historia de la computación, permitiendo a las máquinas emular ciertas capacidades cognitivas humanas. En su núcleo, la IA busca crear sistemas que puedan aprender de los datos, identificar patrones complejos y tomar decisiones con cierto grado de autonomía.

Por su parte, el aprendizaje automático (**machine learning**) es un subconjunto de la IA que se centra en desarrollar algoritmos que pueden mejorar automáticamente a través de la experiencia. A diferencia de la programación tradicional, donde las reglas se establecen explícitamente, en el aprendizaje automático los sistemas aprenden estas reglas a partir de los datos proporcionados.

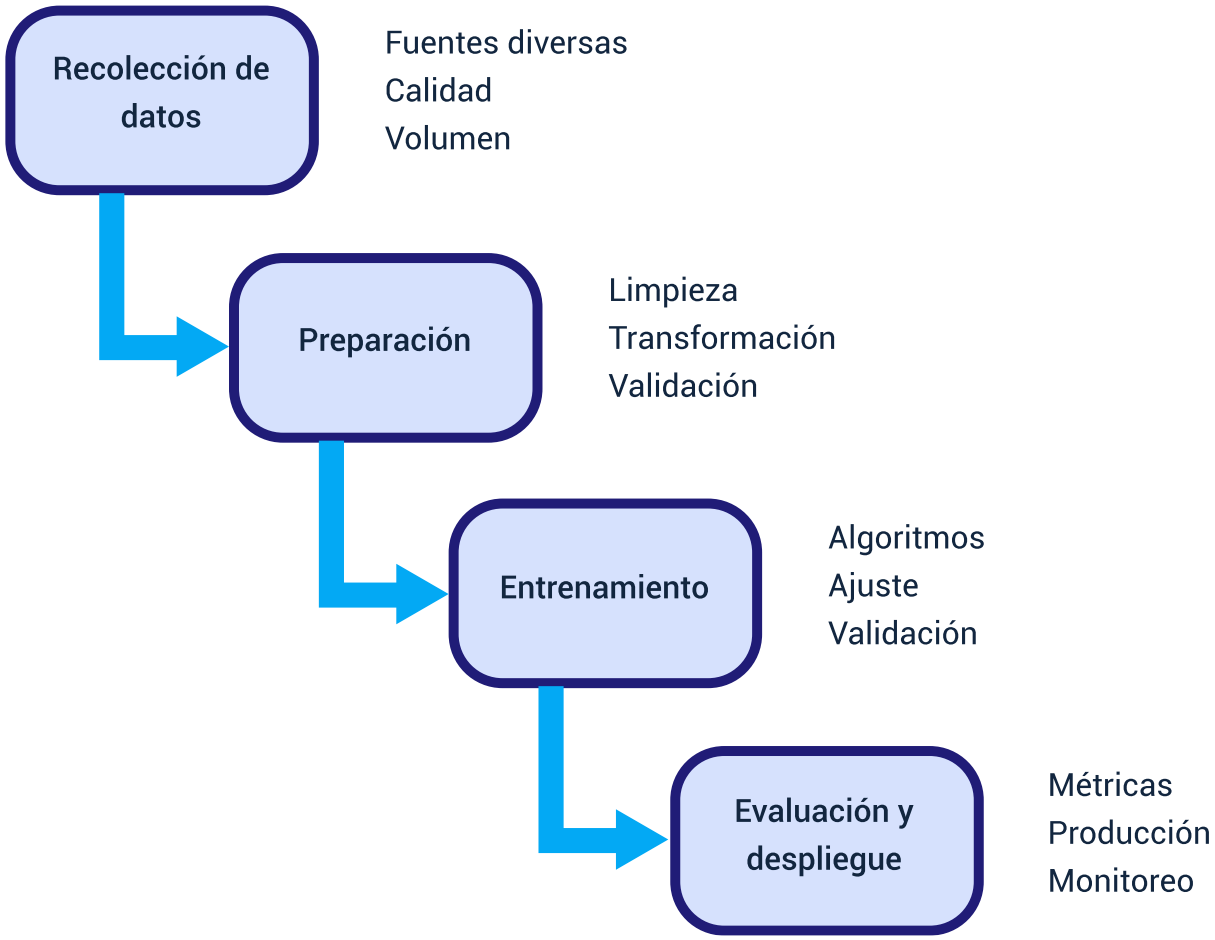
Un concepto fundamental en IA es el de modelo, que bien se puede entender como una representación matemática de un proceso del mundo real. Los modelos de IA aprenden patrones en los datos de entrenamiento y utilizan este conocimiento para hacer predicciones o tomar decisiones sobre nuevos datos.

4.2. Aplicaciones y casos de uso

Las aplicaciones de la IA son vastas y continúan expandiéndose a medida que la tecnología evoluciona. En el ámbito empresarial, la IA se utiliza para optimizar procesos, predecir tendencias de mercado y mejorar la experiencia del cliente. Los sistemas de recomendación, por ejemplo, utilizan IA para analizar el comportamiento de los usuarios y sugerir productos o contenido relevante.

Para comprender mejor cómo se desarrollan los proyectos de IA en la práctica, es útil examinar el flujo de trabajo típico que siguen estos proyectos, como se ilustra en la siguiente infografía. Este proceso iterativo comienza con la recolección de datos de diversas fuentes, seguido de una fase de preparación donde los datos se limpian y transforman. Posteriormente, se desarrolla la fase de entrenamiento donde los modelos aprenden de los datos procesados, y finalmente se llega a la etapa de evaluación y despliegue, donde se mide el rendimiento del modelo y se implementa en producción.

Figura 2. Flujo de trabajo en proyectos de inteligencia artificial



Fuente. OIT, 2024.

Cada etapa de este flujo presenta sus propios desafíos y requerimientos específicos. Por ejemplo, en la fase de recolección de datos se debe asegurar no solo la cantidad, sino también la calidad y representatividad de los datos. La fase de preparación suele ser la más intensiva en tiempo y recursos, ya que de ella depende en gran medida el éxito del modelo. Durante el entrenamiento, la selección del algoritmo adecuado y el ajuste de hiperparámetros son medulares, mientras que en la fase de evaluación y despliegue, el enfoque está en garantizar que el modelo funcione de manera confiable en un entorno de producción.

En el campo de la medicina, la IA está revolucionando el diagnóstico por imagen, al permitir la detección temprana de enfermedades mediante el análisis automatizado de radiografías, resonancias magnéticas y otros tipos de imágenes médicas. Los sistemas de IA también están ayudando en el descubrimiento de nuevos medicamentos, analizando grandes cantidades de datos genéticos y moleculares.

La industria manufacturera está aprovechando la IA para implementar mantenimiento predictivo, donde los algoritmos pueden predecir cuándo una máquina necesitará mantenimiento antes de que ocurra una falla. Esto no solo reduce los costos de mantenimiento, sino que también minimiza el tiempo de inactividad.

4.3. Herramientas básicas para IA

El desarrollo de soluciones de IA requiere un conjunto específico de herramientas y tecnologías. Python se ha convertido en el lenguaje de programación dominante en este campo, gracias a su simplicidad y a la rica ecosistema de bibliotecas especializadas. Entre las bibliotecas más importantes encontramos:

- Scikit-learn para aprendizaje automático tradicional.
- TensorFlow y PyTorch para aprendizaje profundo.
- Pandas para manipulación y análisis de datos.
- NumPy para computación numérica.
- Matplotlib y Seaborn para visualización de datos.

La elección de las herramientas adecuadas dependerá de varios factores, incluyendo la naturaleza del problema a resolver, el volumen de datos a procesar, y los requisitos de rendimiento del sistema. Es importante mencionar que las herramientas

son solo un medio para un fin: el verdadero valor radica en comprender los principios subyacentes y saber cuándo y cómo aplicar cada técnica.

El futuro de la IA promete ser aún más emocionante, con avances en áreas como el aprendizaje por refuerzo, la IA explicable y los sistemas de IA más eficientes en términos de recursos computacionales. Sin embargo, también debemos ser conscientes de los desafíos éticos y de privacidad que surgen con el uso cada vez más generalizado de la IA.

La implementación exitosa de soluciones de IA requiere no solo conocimientos técnicos, sino también una comprensión profunda del dominio del problema y consideraciones éticas sólidas. A medida que la tecnología continúa evolucionando, la capacidad de adaptarse y aprender nuevas herramientas y técnicas será cada vez más importante para quienes se desempeñan en este campo.

5. Conclusiones

El procesamiento de datos para modelos de inteligencia artificial requiere una comprensión profunda y holística que va desde los fundamentos más básicos hasta las aplicaciones más avanzadas. A lo largo de este componente formativo, se exploró cómo los datos, en su forma más elemental, se transforman en estructuras complejas que alimentan los sistemas de IA modernos.

La gestión efectiva de estos datos, desde su recolección hasta su procesamiento final, resulta fundamental para el éxito de cualquier proyecto de inteligencia artificial.

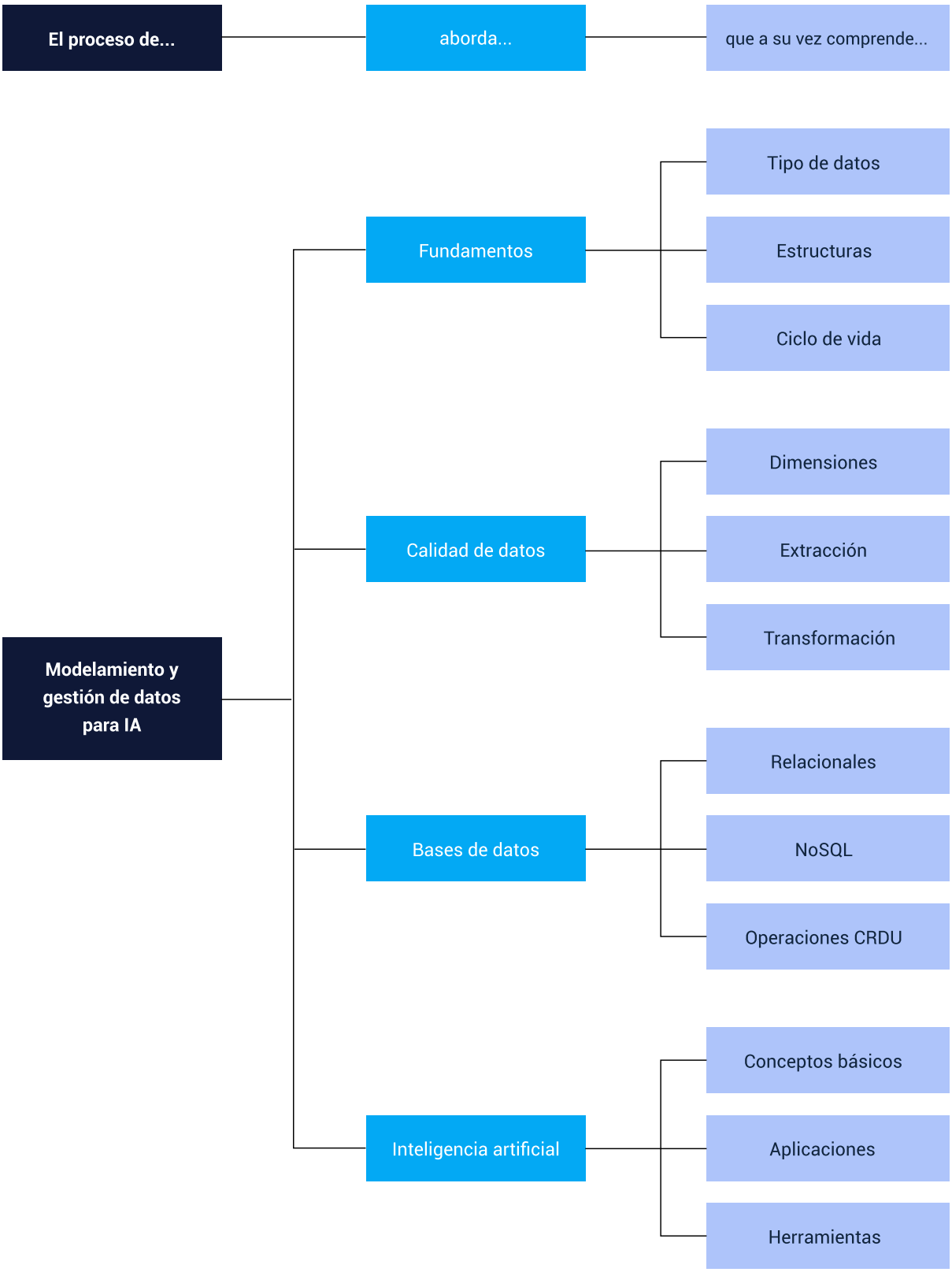
La convergencia entre las estructuras de datos tradicionales, los sistemas de gestión de bases de datos y las tecnologías emergentes de IA nos muestra un panorama en constante evolución. La calidad de los datos, su tratamiento adecuado y la selección de las herramientas apropiadas no son simplemente pasos en un proceso, sino elementos críticos que determinan la eficacia de los modelos de IA.

Síntesis

El siguiente diagrama ilustra los cuatro pilares fundamentales que conforman este componente formativo. Partiendo del concepto central de modelamiento y gestión de datos para IA, se ramifica en áreas esenciales: fundamentos de datos, calidad de datos, bases de datos e inteligencia artificial. Cada una de estas áreas se desglosa en subtemas específicos que reflejan la progresión lógica del aprendizaje.

La estructura presentada permite visualizar cómo los conceptos básicos de datos y estructuras sirven como base para comprender los procesos más complejos de calidad y transformación. Estos, a su vez, se integran con los sistemas de gestión de bases de datos, culminando en las aplicaciones prácticas en inteligencia artificial. Esta organización jerárquica facilita la comprensión de cómo cada elemento contribuye al objetivo final de preparar y gestionar datos efectivamente para su uso en modelos de IA.

El diagrama sirve como una guía de referencia rápida para navegación y repaso, permitiendo al aprendiz visualizar la interconexión entre los diferentes conceptos y su progresión lógica. Se recomienda utilizarlo como un mapa conceptual complementario al contenido detallado del componente, facilitando la identificación de relaciones entre temas y la comprensión integral del material.



Fuente. OIT, 2024.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
1. Fundamentos de datos y estructuras	Ecosistema de Recursos Educativos Digitales SENA. (2023c, septiembre 5). Ejemplo problemas en la recolección de la información.	Video	https://www.youtube.com/watch?v=LOlsg6ZkdcA
2. Calidad y tratamiento de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023b, septiembre 5). Datos sucios.	Video	https://www.youtube.com/watch?v=qf6MR4o58cs
2. Calidad y tratamiento de datos	Limpiar datos de Excel, CSV, PDF y Hojas de cálculo de Google con el intérprete de datos. (s. f.). Tableau.	Portal web	https://help.tableau.com/current/pro/desktop/es-es/data_interpreter.htm
2. Calidad y tratamiento de datos	Ecosistema de Recursos Educativos Digitales SENA. (2022a, agosto 31). Procesos y procedimientos para la gestión de calidad de la información.	Video	https://www.youtube.com/watch?v=PeVlTP8qLhE
2. Calidad y tratamiento de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023c, julio 25). Procesamiento y análisis de datos.	Video	https://www.youtube.com/watch?v=8OSIN2kdU5o
3. Gestión de bases de datos	Ecosistema de Recursos Educativos Digitales SENA. (2022b, octubre 11). Conceptos y estructuras de las bases de datos.	Video	https://www.youtube.com/watch?v=xUpr20u9dmc

Tema	Referencia	Tipo de material	Enlace del recurso
3. Gestión de bases de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023a, marzo 24). Administración de bases de datos: Introducción.	Video	https://www.youtube.com/watch?v=GL7CHwwPlKM
4. Introducción a la Inteligencia Artificial	Ecosistema de Recursos Educativos Digitales SENA. (2023b, marzo 24). Inteligencia artificial en los datos.	Video	https://www.youtube.com/watch?v=-hYXrGAUYAE

Glosario

ACID: acrónimo de Atomicidad, Consistencia, Aislamiento y Durabilidad; propiedades que garantizan que las transacciones en una base de datos sean fiables.

Algoritmo: conjunto ordenado y finito de operaciones sistemáticas que permite hacer un cálculo y hallar la solución de un problema específico.

Base de datos: sistema organizado para recopilar, almacenar y gestionar datos de manera estructurada y eficiente.

CRUD: Acrónimo de `Create`, `Read`, `Update`, `Delete`; operaciones básicas que se pueden realizar sobre datos almacenados.

Dataset: conjunto de datos organizados y formateados de manera específica para su uso en análisis o entrenamiento de modelos.

Datos estructurados: información que está organizada en un formato predefinido y fácilmente procesable por máquinas, típicamente en tablas con filas y columnas.

Estructura de datos: forma particular de organizar datos en una computadora para que puedan ser utilizados de manera eficiente.

ETL: `Extract`, `Transform`, `Load` (Extraer, Transformar, Cargar); proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos.

Indexación: proceso de crear estructuras de datos adicionales que mejoran la velocidad de recuperación de información en una base de datos.

Inteligencia Artificial: campo de la informática que busca crear sistemas capaces de aprender y resolver problemas de manera similar a como lo haría un ser humano.

JSON: JavaScript Object Notation; formato ligero de intercambio de datos, fácil de leer y escribir para humanos y máquinas.

Machine Learning: rama de la inteligencia artificial que se centra en el desarrollo de técnicas que permiten que las computadoras aprendan y mejoren a partir de la experiencia.

Metadata: datos que proporcionan información sobre otros datos, describiendo su contenido, calidad, condición y otras características.

Normalización: proceso de organizar los datos en una base de datos para reducir la redundancia y mejorar la integridad de los datos.

NoSQL: tipo de base de datos que no utiliza el esquema tradicional de tablas relacionales, permitiendo mayor flexibilidad y escalabilidad.

Pipeline de datos: conjunto de procesos y herramientas que permiten mover datos desde una fuente hacia un destino, realizando transformaciones en el camino.

Query: consulta o petición específica para recuperar información de una base de datos.

Schema: estructura que define cómo se organizan los datos en una base de datos, incluyendo tablas, campos y relaciones.

SQL: Structured Query Language; lenguaje estándar para gestionar y manipular bases de datos relacionales.

Validación de datos: proceso de asegurar que los datos cumplan con ciertos criterios de calidad y formato antes de ser utilizados en análisis o procesamiento posterior.

Referencias bibliográficas

Antonio, P. P. (2022). Gestión de bases de datos. Ediciones Paraninfo, S.A.

Díaz, C. O., Soler, P., Pérez, M. & Mier, A. (2024). OMASHU: La ciencia detrás del éxito; Big Data e IA en los eSports. Revista SISTEMAS, 170, 61-79.

Díez, R. P., Gómez, A. G., & De Abajo Martínez, N. (2001). Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva. Universidad de Oviedo.

Guardelli, E. (2024). Minería de Procesos: Convertir Datos en Valor. MedTechBiz.

Jones, H. (2018). Analítica de Datos: Una guía esencial para principiantes en minería de datos, recolección de datos, análisis de Big Data para negocios y conceptos de inteligencia empresarial. Independently Published.

Leyva, D. S. (2024). Domina Machine Learning: Guía completa para principiantes. Independently Published.

McKinsey, W. (2023). Python para análisis de datos. Anaya Multimedia.

Orlandi, M. A. M. (2024). Tecnologías Big Data, Minería de Datos y Analítica aplicada a la gestión de Recursos Humanos: contiene: un caso de estudio. Editora Dialética.

Peraza, E. A. C. (2012). Estructuras y Fundamentos de Datos. Guía de ejercicios prácticos. Lulu.com.

Shovic, J. C. & Simpson, A. (2019). Python All-in-One For Dummies. John Wiley & Sons.

Subirats Maté, L., Pérez Trenard, D. O., Calvo González, M. & Isabel Guitart Hormigo. (2019). Introducción a la limpieza y análisis de los datos. https://openaccess.uoc.edu/bitstream/10609/148647/1/IntroduccionALaLimpiezaYAnalisisDeLosDatos.pdf

Créditos

Elaborado por:

Organización Internacional del Trabajo