Integración y gestión avanzada de datos para inteligencia artificial

Breve descripción:

Este componente profundiza en las técnicas avanzadas de integración y gestión de datos para sistemas de inteligencia artificial. Abarca desde el modelamiento avanzado de bases de datos hasta el análisis exploratorio y la preparación sofisticada de datos, incluyendo metodologías de inteligencia de negocios. Proporciona las herramientas necesarias para implementar soluciones de gestión de datos en contextos empresariales modernos.

Noviembre 2024

Tabla de contenido

Introducción 1

1. Modelamiento avanzado de datos 4

1.1. Reglas de negocio y metodologías 4

1.2. Normalización y diseño de bases de datos 5

1.3. Principios ACID (acrónimo en inglés de atomicidad, consistencia, aislamiento y durabilidad) 6

2. Inteligencia de negocios 8

2.1. Bodegas de datos 8

2.2. Arquitecturas estrella y copo de nieve 9

2.3. Metodologías Kimball e Inmon 11

3. Análisis exploratorio de datos 13

3.1. Estadística descriptiva e inferencial 13

3.2. Métodos de análisis univariable y multivariable 14

3.3. Técnicas de visualización 15

4. Preparación avanzada de datos 17

4.1. Detección y tratamiento de errores 17

4.2. Identificación de variables relevantes 18

4.3. Transformación y validación de datos 19

Síntesis 21

Material complementario 23

Glosario 25

Referencias bibliográficas 28

Créditos 30

Introducción

En la era actual de la inteligencia artificial, la gestión avanzada de datos se ha convertido en un pilar fundamental para el éxito de cualquier iniciativa de transformación digital. La complejidad creciente de los sistemas de información y la necesidad de procesar volúmenes masivos de datos requieren un enfoque sofisticado que va más allá de las técnicas tradicionales de gestión de datos.

Este componente formativo aborda los aspectos más avanzados de la integración y gestión de datos para IA, proporcionando una comprensión de las metodologías, arquitecturas y técnicas necesarias para construir soluciones robustas y escalables. Desde el modelamiento avanzado de datos hasta la implementación de bodegas de datos empresariales, cada tema se explora con un enfoque práctico.

A lo largo del componente, se examinarán las mejores prácticas en inteligencia de negocios, incluyendo las metodologías de Kimball e Inmon, así como las arquitecturas modernas para el almacenamiento y procesamiento de datos. El análisis exploratorio y las técnicas avanzadas de preparación de datos completarán el conjunto de herramientas necesarias para enfrentar los desafíos actuales en la gestión de datos para IA.

La comprensión de estos conceptos avanzados es importante para cualquier persona que busque diseñar e implementar soluciones efectivas de gestión de datos en el contexto de la inteligencia artificial. Como dice un principio fundamental en este campo: "La arquitectura de datos de hoy determina las posibilidades analíticas del mañana".

¡Bienvenido a este viaje por las técnicas avanzadas de gestión de datos para IA!

Video 1. Integración y gestión avanzada de datos para la inteligencia artificial

Enlace de reproducción del video

Síntesis del video: Integración y gestión avanzada de datos para la inteligencia artificial
La gestión avanzada de datos representa uno de los mayores desafíos en la implementación de soluciones de inteligencia artificial. Este componente formativo te guiará a través de las técnicas y metodologías necesarias para enfrentar este reto. El modelamiento avanzado de datos es nuestro punto de partida, exploraremos cómo diseñar estructuras de datos robustas y escalables. Las reglas de negocio y los

principios ACID son fundamentales para garantizar la integridad y consistencia de nuestros datos.

En inteligencia de negocios, abordaremos las bodegas de datos y sus diferentes arquitecturas. Las metodologías de Kimball e Inmon nos proporcionarán marcos de trabajo probados para implementar soluciones empresariales.

El análisis exploratorio de datos cobra especial relevancia en el contexto de la IA. Aprenderemos técnicas estadísticas y métodos de visualización que nos permitirán comprender mejor nuestros datos.

La preparación avanzada de datos cierra el ciclo, con técnicas sofisticadas para detectar y tratar errores, identificar variables relevantes y validar nuestros conjuntos de datos.

Las tendencias actuales apuntan hacia arquitecturas cada vez más complejas que deben manejar datos en tiempo real y a gran escala. La comprensión de estos conceptos avanzados es crucial para mantenerse competitivo.

El componente integra teoría y práctica, permitiéndote desarrollar las habilidades necesarias para implementar soluciones de gestión de datos efectivas en el mundo real.

¡Bienvenido al mundo de la gestión avanzada de datos para IA!

1. Modelamiento avanzado de datos

El modelamiento avanzado de datos constituye la base de la construcción de sistemas de información modernos, especialmente aquellos destinados a soportar aplicaciones de inteligencia artificial. Este capítulo explora las técnicas y metodologías para diseñar e implementar las estructuras de datos que almacenen información de manera eficiente, y que también soporten las complejas necesidades analíticas de los sistemas de IA actuales. Abordaremos desde las reglas de negocio fundamentales hasta los principios de diseño de bases de datos, estableciendo una base sólida para la gestión efectiva de datos en entornos empresariales modernos.

1.1. Reglas de negocio y metodologías

El modelamiento avanzado de datos representa un paso obligado en la gestión moderna de información, especialmente cuando se orienta hacia aplicaciones de inteligencia artificial. Este proceso va más allá del simple diseño de estructuras de almacenamiento, incorporando las reglas de negocio que definen cómo la organización opera y toma decisiones.

Las reglas de negocio son las directrices que gobiernan aspectos del negocio y determinan cómo los datos deben ser capturados, almacenados, transformados y utilizados. Estas reglas pueden variar desde simples validaciones (como "la edad de un empleado debe ser mayor a 18 años") hasta complejas interrelaciones entre diferentes aspectos del negocio (como "un cliente premium debe haber realizado compras por más de $10,000 en los últimos 6 meses y tener un historial de pagos impecable").

Las metodologías de modelamiento actuales tienen un enfoque iterativo y colaborativo, que les permite a los expertos en datos trabajar estrechamente con los

expertos del dominio para asegurar que el modelo final refleje adecuadamente tanto los requisitos técnicos como las necesidades del negocio.

1.2. Normalización y diseño de bases de datos

La normalización es un proceso fundamental en el diseño de bases de datos que busca eliminar redundancias y dependencias problemáticas. Para comprender mejor este proceso, se deben conocer las diferentes formas normales y su impacto en el diseño de bases de datos. Cada forma normal representa un nivel incremental de organización y optimización de la estructura de datos.

La siguiente tabla expone una visión general de las principales formas normales, sus objetivos y consideraciones clave:

Tabla 1. Formas de normalización en las bases de datos

Forma normal	Objetivo	Beneficios	Consideraciones
1FN	Eliminar grupos repetitivos.	Atomicidad de datos.	Puede aumentar el número de tablas.
2FN	Eliminar dependencias parciales.	Mejor integridad de datos.	Requiere identificar dependencias funcionales.
3FN	Eliminar dependencias transitivas.	Reduce redundancia.	Puede afectar el rendimiento de consultas.
BCNF	Eliminar anomalías restantes.	Máxima normalización práctica.	Complejidad aumentada.
4FN	Manejar dependencias multivaluadas.	Mejor manejo de relaciones M:N.	Raramente necesaria.

Fuente. OIT, 2024.

Esta progresión de formas normales ilustra cómo el proceso de normalización evoluciona desde conceptos básicos hasta consideraciones más avanzadas. Sin embargo, es importante notar que no siempre es necesario o deseable alcanzar los niveles más altos de normalización. La decisión de hasta qué nivel normalizar debe basarse en un análisis cuidadoso de los requisitos específicos del sistema, considerando factores como el rendimiento de las consultas, la naturaleza de los datos y las necesidades del negocio.

1.3. Principios ACID (acrónimo en inglés de atomicidad, consistencia, aislamiento y durabilidad)

Los principios ACID constituyen la base de la integridad transaccional en sistemas de bases de datos. Estos principios son especialmente determinantes en entornos donde la precisión y la consistencia de los datos son fundamentales para las operaciones del negocio.

- **Atomicidad**

  Asegura que una transacción se complete en su totalidad o no se realice en absoluto. Por ejemplo, en una transferencia bancaria, tanto el débito de una cuenta como el crédito en otra deben completarse exitosamente, o ninguna operación debe realizarse.

- **Consistencia**

  Garantiza que una transacción lleve la base de datos de un estado válido a otro igualmente válido. Esto significa que todas las reglas de integridad definidas deben cumplirse antes y después de cada transacción.

- **Aislamiento**

  Previene que transacciones concurrentes interfieran entre sí. Cada transacción debe ejecutarse como si fuera la única operación siendo realizada en el sistema, aunque en realidad múltiples transacciones pueden estar ejecutándose simultáneamente.

- **Durabilidad**

  Asegura que una vez una transacción se ha completado, sus efectos son permanentes y sobrevivirán a cualquier falla subsecuente del sistema. Esto típicamente se logra a través de logs de transacciones y mecanismos de recuperación.

En el contexto de la IA, estos principios ACID adquieren una nueva dimensión de importancia. Los modelos de IA dependen claramente de la calidad y consistencia de los datos de entrenamiento, y cualquier violación de estos principios podría resultar en modelos sesgados o poco confiables. Por ejemplo, si los datos de entrenamiento se recopilan durante estados inconsistentes de la base de datos, los patrones aprendidos por el modelo podrían no reflejar la realidad del negocio.

El desafío actual en el modelamiento avanzado de datos radica en encontrar el equilibrio adecuado entre estos principios tradicionales y las necesidades emergentes de sistemas de IA, que a menudo requieren flexibilidad y velocidad en el acceso a los datos.

2. Inteligencia de negocios

La inteligencia de negocios (BI) representa la convergencia entre la gestión de datos empresariales y la toma de decisiones estratégicas. En el contexto de la inteligencia artificial, la BI se ha convertido en un componente elemental que proporciona la infraestructura y los métodos necesarios para transformar datos crudos en información accionable. Este capítulo explora las arquitecturas, metodologías y mejores prácticas que permiten a las organizaciones aprovechar al máximo sus datos para obtener ventajas competitivas.

2.1. Bodegas de datos

Las bodegas de datos (Data Warehouses) son el fundamento de cualquier estrategia moderna de inteligencia de negocios. A diferencia de las bases de datos operacionales tradicionales, las bodegas de datos están diseñadas específicamente para el análisis y el soporte a la toma de decisiones.

Sus características principales incluyen:

- **Orientación a temas**

  Los datos se organizan por áreas temáticas principales del negocio.
- **Integración**

  Datos provenientes de múltiples fuentes se unifican bajo un esquema coherente.
- **No volatilidad**

  Los datos históricos se preservan para análisis temporales.
- **Variación temporal**

  Se mantiene la dimensión temporal de todos los datos.

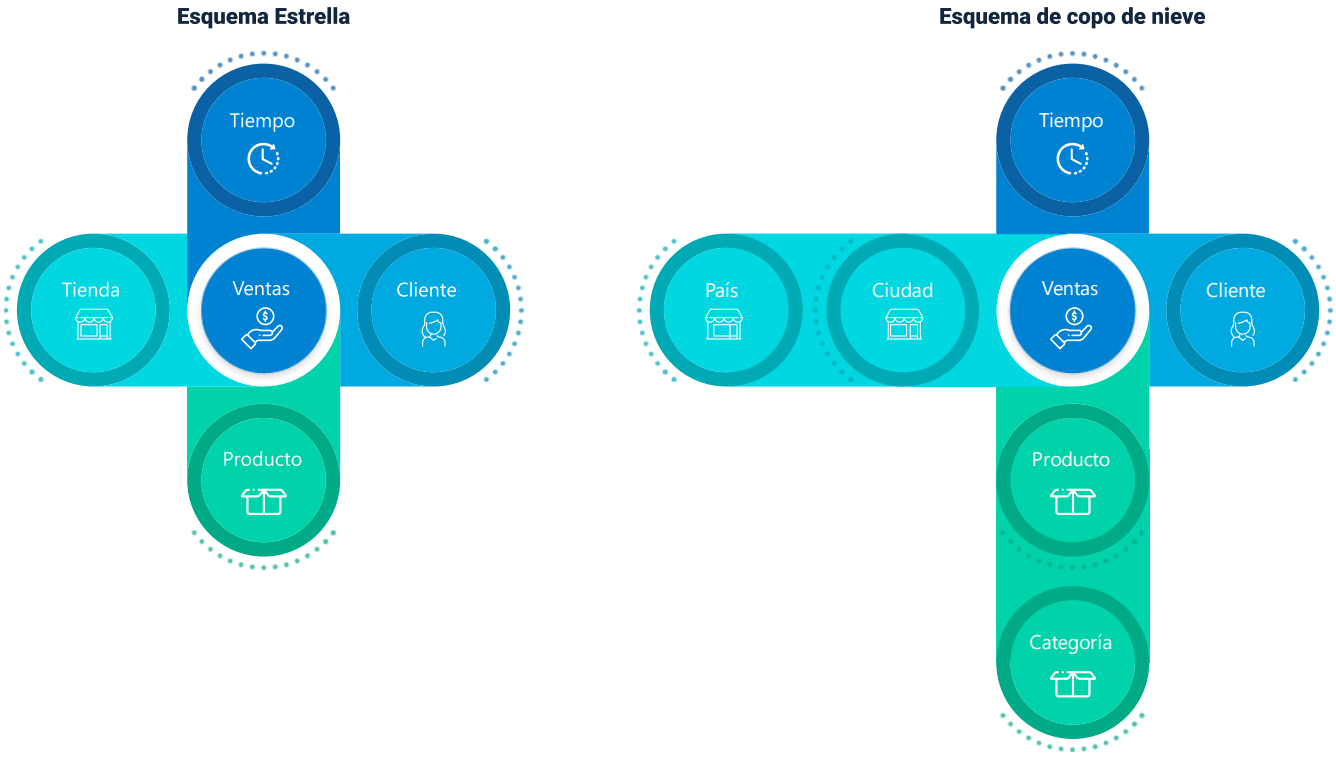
El diseño de una bodega de datos requiere una comprensión profunda tanto de las necesidades analíticas del negocio como de las características de los datos disponibles. Un aspecto crítico es la estrategia de actualización y mantenimiento, que debe balancear la frescura de los datos con la estabilidad del sistema.

2.2. Arquitecturas estrella y copo de nieve

Las arquitecturas estrella y copo de nieve son los dos paradigmas principales para organizar datos en un entorno de bodega de datos. Cada arquitectura tiene sus propias características y casos de uso óptimos. Para comprender mejor las diferencias fundamentales entre estas arquitecturas, consideremos una representación visual de ambos esquemas. La figura que se muestra más adelante ilustra cómo se organizan las tablas y sus relaciones en cada tipo de arquitectura, tomando como ejemplo un caso típico de análisis de ventas.

Como se puede observar, mientras que el esquema estrella mantiene una estructura más simple y directa con todas las dimensiones conectadas directamente a la tabla de hechos central, el esquema copo de nieve introduce niveles adicionales de normalización en las dimensiones. Esta diferencia estructural tiene implicaciones significativas tanto en el rendimiento de las consultas como en la mantenibilidad del sistema. Por ejemplo, en el esquema copo de nieve, la información sobre productos está normalizada con una tabla separada para categorías, lo que reduce la redundancia de datos, pero aumenta la complejidad de las consultas.

Figura 1. Comparación de arquitecturas de bodegas de datos



Fuente. OIT, 2024.

El esquema estrella se caracteriza por su simplicidad y eficiencia en las consultas. En el centro se encuentra la tabla de hechos, que contiene las métricas del negocio, rodeada por tablas de dimensiones desnormalizadas. Esta estructura facilita la navegación intuitiva de los datos y generalmente proporciona mejor rendimiento en consultas analíticas. Por otro lado, el esquema copo de nieve introduce normalización adicional en las tablas de dimensiones, creando una estructura más compleja, pero más eficiente en términos de almacenamiento. Esta arquitectura es particularmente útil cuando:

a) Las dimensiones tienen múltiples niveles de jerarquía.
b) El almacenamiento es una preocupación primordial.

c) La integridad referencial es crítica.

d) Las dimensiones son compartidas entre múltiples esquemas.

2.3. Metodologías Kimball e Inmon

Las metodologías de Kimball e Inmon representan dos filosofías fundamentales pero diferentes para el diseño y la implementación de bodegas de datos. Aunque ambas buscan crear un repositorio de datos empresariales eficiente, sus enfoques difieren significativamente.

La metodología Kimball, también conocida como el enfoque "bottom-up", comienza con la identificación de procesos de negocio específicos y construye data marts interconectados que eventualmente forman una bodega de datos empresarial. Las características principales de este enfoque incluyen:

a) Desarrollo incremental por áreas de negocio.

b) Uso consistente de dimensiones conformadas.

c) Enfoque en la usabilidad y el rendimiento analítico.

d) Mayor flexibilidad y tiempo más rápido hasta el primer resultado.

Por su parte, la metodología Inmon, conocida como el enfoque "top-down", aboga por la construcción de una bodega de datos empresarial normalizada desde el principio, a partir de la cual se pueden derivar data marts específicos. Este enfoque se distingue por:

a) Visión empresarial integral desde el inicio.

b) Datos altamente normalizados en el nivel empresarial.

c) Énfasis en la consistencia y la integración de datos.

d) Mayor inversión inicial pero mejor escalabilidad a largo plazo.

La selección de la metodología más apropiada requiere una evaluación cuidadosa del contexto organizacional. Las empresas deben considerar su nivel de madurez en la gestión de datos, pues organizaciones más experimentadas pueden estar mejor preparadas para implementar el enfoque más estructurado de Inmon. Los recursos disponibles también juegan un papel medular; el tiempo, presupuesto y personal técnico necesarios varían significativamente entre metodologías. Además, la urgencia por obtener resultados analíticos puede inclinar la balanza hacia el enfoque más ágil de Kimball, mientras que la complejidad del ambiente de datos y los requisitos de gobierno podrían favorecer la estructura más rigurosa de Inmon.

En el panorama actual, la dicotomía entre estas metodologías se ha difuminado considerablemente. Las organizaciones modernas tienden a adoptar enfoques híbridos, al seleccionar y adaptar elementos de ambas metodologías según sus circunstancias particulares. Lo verdaderamente determinante no es la adherencia estricta a una metodología específica, sino mantener la consistencia en el enfoque elegido y asegurar que la implementación se alinee efectivamente con los objetivos analíticos de la organización. Este equilibrio pragmático permite a las empresas aprovechar las fortalezas de ambas metodologías mientras mitigan sus respectivas limitaciones.

3. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA, por sus siglas en inglés) constituye una etapa fundamental en el proceso de comprensión y preparación de datos para modelos de inteligencia artificial. Esta fase inicial de investigación ayuda a los analistas a comprender las características fundamentales de los conjuntos de datos, identificar patrones significativos y detectar anomalías que podrían afectar análisis posteriores. A diferencia de los métodos estadísticos confirmatorios, el EDA adopta un enfoque más flexible y orientado al descubrimiento, permitiendo que los datos revelen sus secretos de manera orgánica.

3.1. Estadística descriptiva e inferencial

La estadística descriptiva proporciona el conjunto de herramientas fundamental para comenzar cualquier análisis exploratorio. A través de medidas de tendencia central, dispersión y forma, los analistas pueden obtener una primera aproximación a la naturaleza de sus datos. Sin embargo, el verdadero poder de la estadística descriptiva radica en su capacidad para revelar aspectos sutiles de los datos que podrían pasar desapercibidos en un examen superficial.

Las medidas de tendencia central, como la media, mediana y moda, ofrecen diferentes perspectivas sobre el centro de los datos. Por ejemplo, en distribuciones altamente sesgadas, la diferencia significativa entre la media y la mediana puede revelar la presencia de valores extremos que requieren atención especial. La dispersión de los datos, medida a través de la varianza, desviación estándar y rango intercuartílico, proporciona información clave sobre la variabilidad y la confiabilidad de las mediciones.

Por su parte, la estadística inferencial permite extender las conclusiones obtenidas de una muestra a la población general. Este proceso requiere una comprensión profunda de conceptos como intervalos de confianza, pruebas de hipótesis y significancia estadística. En el contexto del análisis exploratorio, las técnicas inferenciales ayudan a validar patrones observados y a determinar si las relaciones descubiertas son estadísticamente significativas.

3.2. Métodos de análisis univariable y multivariable

El análisis univariable representa el primer paso en la exploración detallada de cada variable individual dentro de un conjunto de datos. Este análisis fundamental examina la distribución, centralidad, dispersión y forma de cada variable por separado, proporcionando una base sólida para análisis más complejos. Los principales aspectos que se deben considerar en el análisis univariable incluyen:

a) Distribución de frecuencias y visualizaciones básicas.
b) Detección de valores atípicos y anomalías.
c) Análisis de valores faltantes.
d) Evaluación de la normalidad.
e) Identificación de patrones temporales o secuenciales.

El análisis multivariable, por su parte, examina las relaciones entre múltiples variables simultáneamente. Este enfoque es particularmente relevante en el contexto de la inteligencia artificial, donde los modelos frecuentemente deben procesar numerosas variables interrelacionadas. **La complejidad de estas relaciones puede apreciarse mejor a través de la siguiente tabla comparativa**:

Tabla 2. Tipos de análisis: técnicas, aplicaciones y consideraciones clave

Tipo de Análisis	Técnicas Principales	Aplicaciones	Consideraciones clave
Bivariado.	Correlación de Pearson, Chi-cuadrado.	Relaciones entre pares de variables.	Asunciones de linealidad.
Regresión múltiple.	Mínimos cuadrados, Ridge, Lasso.	Predicción de variables continuas.	Multicolinealidad.
Análisis factorial.	PCA, Factor Analysis.	Reducción de dimensionalidad.	Interpretabilidad de factores.
Análisis de clústeres.	K-means, jerárquico.	Segmentación de datos.	Selección de número de grupos.
Análisis discriminante.	LDA, QDA.	Clasificación supervisada.	Separabilidad de clases.

Fuente. OIT, 2024.

3.3. Técnicas de visualización

La visualización de datos representa una herramienta indispensable en el análisis exploratorio, transformando números abstractos en representaciones visuales intuitivas que facilitan la identificación de patrones, tendencias y anomalías. Las técnicas de visualización modernas van más allá de los gráficos básicos, incorporando elementos interactivos y múltiples dimensiones de información en una sola representación.

La selección de la técnica de visualización adecuada depende tanto de la naturaleza de los datos como del objetivo del análisis. Por ejemplo, los diagramas de dispersión resultan invaluables para examinar relaciones entre variables continuas, mientras que los gráficos de calor pueden revelar patrones complejos en matrices de correlación. Los gráficos de caja (box plots) combinan múltiples aspectos de la

distribución de datos, mostrando simultáneamente la mediana, cuartiles y valores atípicos.

En el contexto del big data y la inteligencia artificial, las técnicas de visualización han evolucionado para manejar volúmenes masivos de datos. Las visualizaciones interactivas permiten a los analistas explorar diferentes niveles de detalle, mientras que las técnicas de reducción de dimensionalidad como t-SNE y UMAP facilitan la visualización de datos altamente dimensionales en espacios bidimensionales o tridimensionales.

La efectividad de una visualización no solo depende de su precisión técnica, sino también de su capacidad para comunicar información de manera clara y convincente. Los principios de diseño visual, como el uso apropiado del color, la gestión del espacio y la jerarquía visual, desempeñan un papel destacado en la creación de visualizaciones que sean tanto informativas como accesibles

El análisis exploratorio de datos, con sus múltiples facetas y técnicas, es una fase crítica en cualquier proyecto de análisis de datos o inteligencia artificial. La combinación de estadística rigurosa, análisis multivariable y visualización efectiva proporciona los cimientos necesarios para construcción de modelos robustos y confiables.

4. Preparación avanzada de datos

La preparación avanzada de datos representa la culminación del proceso de transformación de datos crudos en información lista para alimentar modelos de inteligencia artificial. Este proceso va más allá de la limpieza de datos; implica utilizar técnicas sofisticadas de detección de errores, selección de variables y validación que aseguran la calidad y relevancia de los datos para su uso en análisis avanzados. La complejidad de esta etapa frecuentemente determina el éxito o fracaso de los proyectos de IA.

4.1. Detección y tratamiento de errores

La detección y tratamiento de errores en datos constituye una fase crítica que requiere una combinación de automatización inteligente y criterio experto. Los errores en datos pueden manifestarse de múltiples formas, desde inconsistencias obvias hasta anomalías sutiles que solo se revelan a través de análisis detallados. El proceso de detección debe ser sistemático y exhaustivo, considerando tanto la calidad individual de cada variable como la coherencia global del conjunto de datos.

Los errores más comunes incluyen valores fuera de rango, inconsistencias lógicas entre variables relacionadas, y patrones temporales imposibles. Sin embargo, la verdadera complejidad radica en identificar errores que son técnicamente válidos pero contextualmente incorrectos. Por ejemplo, un valor de temperatura podría estar dentro del rango permitido, pero ser improbable dado el contexto geográfico y temporal.

El tratamiento de errores una vez detectados requiere un enfoque matizado. La simple eliminación de registros problemáticos puede introducir sesgos en los datos, mientras que la corrección automática puede crear artificios que afecten análisis

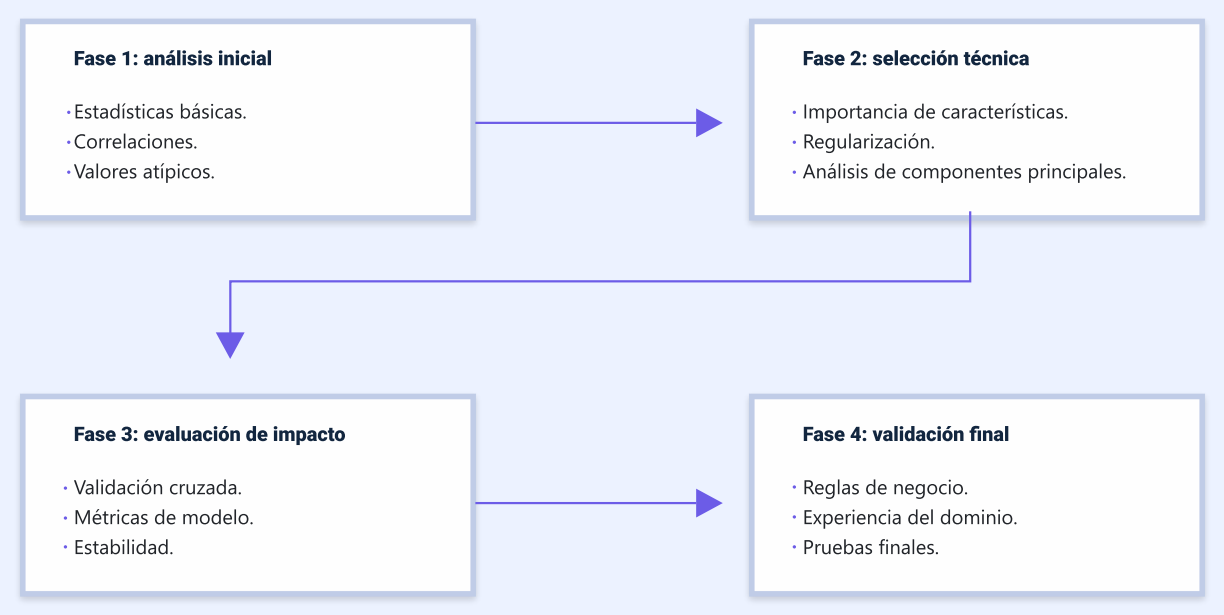
posteriores. Es fundamental documentar todas las decisiones de tratamiento de errores y mantener la trazabilidad de las modificaciones realizadas.

4.2. Identificación de variables relevantes

La identificación y selección de variables relevantes constituye uno de los desafíos más significativos en la preparación avanzada de datos. Como se ilustra en la infografía anterior, este proceso sigue una secuencia metodológica que combina análisis estadístico, técnicas de selección automatizada y validación experta. Cada fase del proceso contribuye a la identificación de las variables que realmente aportan valor al modelo final.

La identificación de variables relevantes sigue un proceso estructurado que combina análisis estadístico con conocimiento del dominio. La siguiente figura ilustra las cuatro fases principales de este proceso y sus componentes clave.

Figura 2. Proceso de selección de variables



Fuente. OIT, 2024.

El proceso comienza con un análisis inicial exhaustivo que examina las características estadísticas de cada variable, sus relaciones con otras variables y su completitud. Esta fase establece la base para decisiones informadas sobre qué variables merecen consideración adicional. La fase de selección aplica técnicas avanzadas como análisis de importancia de características y métodos de regularización para identificar las variables más prometedoras.

La evaluación de impacto y la validación final son estratégicas para asegurar que las variables seleccionadas no solo son estadísticamente significativas, sino también relevantes desde una perspectiva del negocio. Este enfoque holístico ayuda a evitar la trampa común de seleccionar variables basándose únicamente en criterios estadísticos.

4.3. Transformación y validación de datos

La transformación y validación de datos representa la última línea de defensa antes de que los datos sean utilizados en modelos de IA. Esta fase combina técnicas de transformación sofisticadas con procesos rigurosos de validación para asegurar que los datos cumplan con todos los requisitos necesarios para su uso en modelado.

Las transformaciones pueden incluir codificación de variables categóricas, normalización de variables numéricas, y creación de características derivadas. Cada transformación debe ser cuidadosamente documentada y validada para asegurar que preserve la integridad de la información mientras la hace más adecuada para el análisis automatizado.

La validación debe ser un proceso continuo que ocurre en múltiples niveles. A nivel técnico, se verifica que las transformaciones mantengan las relaciones importantes entre variables y no introduzcan sesgos indeseados. A nivel de negocio, se

confirma que los datos transformados sigan representando fielmente la realidad del dominio.

El proceso de validación también debe incluir pruebas de robustez para asegurar que las transformaciones sean estables y reproducibles en diferentes condiciones. Esto es particularmente importante en sistemas de producción donde los datos se procesan de manera continua.

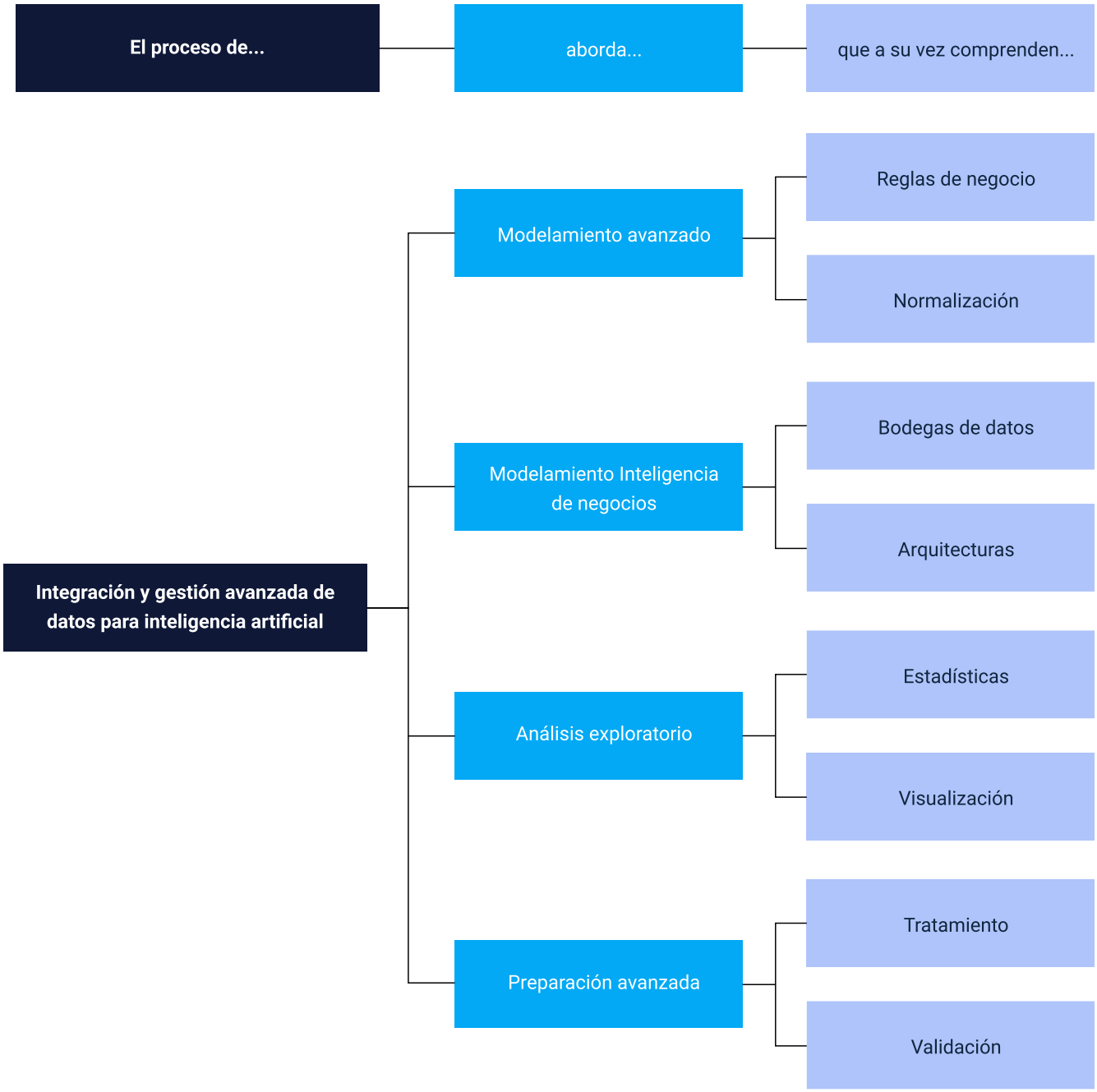
El éxito en la preparación avanzada de datos requiere un balance delicado entre automatización y supervisión humana. Mientras que las herramientas automatizadas pueden manejar eficientemente grandes volúmenes de datos, el juicio experto sigue siendo indispensable para tomar decisiones estratégicas y validar resultados críticos.

Síntesis

El diagrama siguiente representa la estructura integral del componente formativo, donde se parte del concepto central de gestión avanzada de datos para IA y se ramifica en cuatro áreas fundamentales: modelamiento avanzado, inteligencia de negocios, análisis exploratorio y preparación avanzada. Cada una de estas áreas se desglosa en subtemas específicos que constituyen los pilares del proceso de integración y gestión de datos para IA.

Esta organización refleja la progresión lógica del aprendizaje, desde los fundamentos del modelamiento hasta las técnicas más avanzadas de preparación de datos. La interconexión entre las diferentes áreas ilustra cómo cada concepto contribuye al objetivo final de preparar datos de alta calidad para su uso en sistemas de inteligencia artificial.

El diagrama sirve como una guía visual para navegar por los conceptos presentados en el texto, permitiendo al aprendiz comprender rápidamente la estructura del componente y las relaciones entre sus diferentes elementos. Se recomienda utilizarlo como referencia para organizar el estudio y comprender la integración de los diferentes aspectos de la gestión avanzada de datos.



Fuente. OIT, 2024.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
1. Modelamiento avanzado de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023e, marzo 27). Modelos y metodologías de analítica.	Video	https://www.youtube.com/watch?v=96pohadjEWE
2. Inteligencia de negocios	Ecosistema de Recursos Educativos Digitales SENA. (2023d, marzo 27). Bodegas de datos	Video	https://www.youtube.com/watch?v=SsP1tA6hAdg
2. Inteligencia de negocios	Ecosistema de Recursos Educativos Digitales SENA. (2023a, marzo 23). Modelos y esquemas de bodega de datos.	Video	https://www.youtube.com/watch?v=Uq6WxfzaroM
3. Análisis exploratorio 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
3. Análisis exploratorio de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023c, marzo 24). Introducción a la aplicación de herramientas estadísticas en la presentación de datos.	Video	https://www.youtube.com/watch?v=M9q9zxX8Evc%3C
4. Preparación avanzada de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023c, julio 25). Procesamiento y análisis de datos.	Video	https://www.youtube.com/watch?v=8OSIN2kdU5o
4. Preparación avanzada de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023e, diciembre 30).	Video	https://www.youtube.com/watch?v=HjJpqHD6sV0

Tema	Referencia	Tipo de material	Enlace del recurso
	Modelamiento, análisis y preparación de datos.		

Glosario

Arquitectura estrella: modelo de diseño de bases de datos dimensionales donde una tabla de hechos central se conecta con múltiples tablas de dimensiones desnormalizadas.

Bodega de datos: sistema de almacenamiento diseñado específicamente para el análisis y reporteo, que integra datos de múltiples fuentes en un modelo unificado.

Copo de nieve: variante de la arquitectura estrella donde las dimensiones están normalizadas, creando una estructura más compleja pero con mejor eficiencia de almacenamiento.

Data mart: subconjunto de una bodega de datos enfocado en un área específica del negocio o departamento.

Dimensiones conformadas: tablas de dimensiones estandarizadas que se comparten entre diferentes data marts, asegurando consistencia en el análisis.

ETL avanzado: procesos sofisticados de Extracción, Transformación y Carga que incluyen validaciones complejas y transformaciones avanzadas de datos.

Feature importance: medida que indica la relevancia o contribución de cada variable en un modelo predictivo o análisis estadístico.

Metadatos empresariales: información que describe el contenido, formato, estructura y uso de los datos en un contexto empresarial.

Metodología Inmon: enfoque "top-down" para el diseño de bodegas de datos, que comienza con una visión empresarial completa y luego deriva en data marts específicos.

Metodología Kimball: enfoque "bottom-up" para el diseño de bodegas de datos, que construye data marts incrementalmente que luego se integran en una solución empresarial.

Normalización avanzada: proceso de diseño de bases de datos que va más allá de la tercera forma normal, incluyendo BCNF y formas normales superiores.

Prueba de hipótesis: método estadístico para tomar decisiones sobre poblaciones basándose en muestras de datos.

Reglas de negocio: políticas, condiciones y restricciones que definen cómo se deben gestionar y validar los datos en un contexto empresarial.

Tabla de hechos: tabla central en un modelo dimensional que contiene las métricas o medidas del negocio y las claves foráneas a las dimensiones.

Tablas de dimensiones: tablas que contienen los atributos descriptivos utilizados para analizar los datos en las tablas de hechos.

Transformación de datos: proceso de convertir datos de un formato o estructura a otro, incluyendo limpieza, normalización y agregación.

Validación cruzada: técnica estadística para evaluar modelos analíticos dividiendo los datos en conjuntos de entrenamiento y prueba.

Variables categóricas: tipos de datos que representan categorías o grupos discretos, que pueden ser nominales u ordinales.

Visualización avanzada: técnicas sofisticadas para representar datos complejos de manera visual, incluyendo gráficos interactivos y multidimensionales.

Workflow ETL: flujo de trabajo que define la secuencia y dependencias de los procesos de extracción, transformación y carga de datos.

Referencias bibliográficas

Aguilar, L. J. (2020). Inteligencia de negocios y analítica de datos. Marcombo.

De Pablos Heredero, C., Agius, J. J. L. H., Romero, S. M., & Salgado, S. M. (2019). Organización y transformación de los sistemas de información en la empresa. ESIC.

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.

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.

Maldonado, L. (2012). Data Analysis Using Regression and Multilevel/Hierarchical Models. Persona y Sociedad, 26(1), 191. https://doi.org/10.53689/pys.v26i1.12

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

Wilke, C. O. (2019). Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. O'Reilly Media.

Créditos

Elaborado por:

Organización Internacional del Trabajo