UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

REPORTE DE LIMPIEZA DE DATOS

IDGS91N

PRESENTA:

REGINA CHÁVEZ TAMAYO

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 05 de octubre de 2025

Índice

Introducción 3
Procedencia de los datos 3
Tipos y fuentes de datos 3
Técnicas de limpieza de datos 4
Fundamentación y estructura 5
Conclusión 5
Referencias 6

Introducción

El análisis de datos es una herramienta esencial para las empresas digitales que buscan comprender el comportamiento de sus clientes y optimizar sus operaciones. En una tienda en línea, los datos provienen de diversas fuentes, como transacciones, reseñas, navegación web o redes sociales. Sin embargo, para que sean útiles y confiables, deben clasificarse correctamente y someterse a procesos de limpieza que garanticen su calidad.

El presente reporte analiza la procedencia, tipos y fuentes de los datos, así como las técnicas de limpieza más usadas, aplicadas a un caso realista de comercio electrónico.

Procedencia de los datos

En este caso de estudio, los datos provienen de una tienda en línea que registra las compras y el comportamiento de sus usuarios. La procedencia es variada:

- **Datos de transacciones:** generados automáticamente por el sistema cada vez que un cliente realiza una compra, incluye monto, productos, método de pago y fecha.
- **Datos generados por humanos:** provienen de reseñas, calificaciones y comentarios escritos por los usuarios en la plataforma.
- **Datos web:** capturados mediante cookies y registros de navegación, como páginas visitadas, tiempo de permanencia o clics.
- **Datos de redes sociales:** recopilados a través de integraciones con plataformas como Facebook o Instagram, donde se analizan interacciones con campañas publicitarias.
- **Datos máquina a máquina (M2M):** generados por los sistemas de inventario o pasarelas de pago que intercambian información automáticamente.

Estos conjuntos de datos permiten construir una visión completa del comportamiento del cliente y del rendimiento del negocio.

Tipos y fuentes de datos

En la tienda en línea se manejan distintos tipos de datos según su naturaleza y estructura:

- **Datos estructurados:** organizados en tablas y bases de datos, como registros de ventas, catálogos de productos o datos de clientes.

- **Datos no estructurados:** textos libres de reseñas o comentarios, imágenes de productos y publicaciones en redes sociales.
- **Datos semiestructurados:** archivos JSON o XML usados para intercambiar información entre sistemas.

De acuerdo con su tipo de medición o variable, pueden clasificarse en:

- **Cuantitativos:** precios, cantidades vendidas, montos de compra, número de visitas.
- **Cualitativos:** opiniones, categorías de productos, métodos de envío.
- **Nominales:** nombres de usuarios o marcas.
- **Ordinales:** valoraciones de 1 a 5 estrellas.

Las principales fuentes de datos son los sistemas internos (base de datos del e-commerce), herramientas de análisis web como Google Analytics, y redes sociales que aportan información sobre la interacción del consumidor.

Técnicas de limpieza de datos

Durante la recopilación pueden surgir problemas de calidad que deben corregirse mediante un proceso de limpieza de datos:

- **Eliminación o imputación de valores nulos:** cuando faltan datos en campos como dirección o teléfono, se completan con información de respaldo o se eliminan los registros incompletos.
- **Detección y corrección de duplicados:** se identifican usuarios o transacciones repetidas mediante claves únicas o comparaciones de campos.
- **Normalización de formatos:** unificar fechas, monedas o formatos de texto para mantener consistencia.
  Ejemplo: convertir todas las fechas a formato ISO (YYYY-MM-DD).
- **Tratamiento de valores atípicos:** revisar precios o cantidades inusualmente altos o bajos que puedan ser errores de registro.
- **Validación de integridad:** asegurar que los datos cumplan con las reglas de negocio (por ejemplo, que el total de la compra coincida con la suma de los productos).

Estas técnicas garantizan la fiabilidad del análisis y evitan conclusiones erróneas.

Fundamentación y estructura

El análisis y limpieza de datos en el comercio electrónico se fundamentan en los principios de gestión de calidad de datos y en el ciclo de vida de la información, que garantizan que los datos sean precisos, completos y confiables.

De acuerdo con IBM (2023), la limpieza de datos es una etapa crítica dentro de la preparación de datos, ya que su correcta ejecución influye directamente en la validez de los análisis.

Laureano-Cruz y Méndez (2021) destacan que, en el ámbito del e-commerce, los datos bien gestionados permiten personalizar la experiencia del cliente y mejorar la fidelización.

La estructura de este trabajo sigue el flujo lógico del procesamiento de datos: primero se identifica su origen (procedencia), luego se clasifican (tipos y fuentes), se depuran (técnicas de limpieza) y finalmente se justifican con fundamentos teóricos.

En síntesis, la correcta administración de los datos no solo es una práctica técnica, sino un proceso estratégico que impacta en la competitividad y eficiencia de las empresas digitales.

Conclusión

Los datos son un activo estratégico para cualquier tienda en línea, pero su valor depende directamente de su calidad. Conocer su procedencia permite entender su contexto, mientras que clasificarlos adecuadamente facilita su análisis. Las técnicas de limpieza de datos aseguran que la información sea confiable y útil para la toma de decisiones.

A través de una buena gestión de datos, las empresas pueden ofrecer experiencias personalizadas, identificar oportunidades de venta y aumentar su competitividad en el mercado digital.

Referencias

Actian. (s. f.). Calidad de datos: una guía detallada. Recuperado de https://www.actian.com/es/data-quality/

Bismart. (s. f.). ¿Qué es data quality y cómo mejorar la calidad de tus datos? Blog Bismart. Recuperado de https://blog.bismart.com/que-es-data-quality-casos-de-uso

Captain Verify. (2025, 13 de agosto). Calidad de datos: ¿por qué es tan importante? Recuperado de https://captainverify.com/es/blog/importancia-calidad-datos.html

OpenAI. (2025). *ChatGPT – Consulta sobre procedencia, tipos y limpieza de datos en e-commerce.* Recuperado el 5 de octubre de 2025, de https://chat.openai.com

IBM. (s. f.). Limpieza de datos (uso en SPSS Modeler). Recuperado de https://www.ibm.com/docs/es/spss-modeler/saas?topic=preparation-cleaning-data

IBM. (s. f.). ¿Qué es la limpieza de datos? IBM Think. Recuperado de https://www.ibm.com/mx-es/think/topics/data-cleaning