UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

Extracción de Conocimiento en Bases de Datos

II.1. Reporte de limpieza de datos

IDGS91N

PRESENTA:

JORGE ALEJANDRO HERNANDEZ CONTRERAS

DOCENTE:

Enrique Mascote

Chihuahua, Chih.,5 Octubre de 2025

Introducción

Hoy en día estamos rodeados de datos por todos lados: cuando compramos en línea, cuando usamos redes sociales o incluso cuando una máquina se comunica con otra. Sin embargo, no todos esos datos se pueden usar tal cual, porque muchos tienen errores, están incompletos o simplemente no tienen el mismo formato. Por eso es importante conocer de dónde vienen los datos, qué tipo de información representan y cómo se pueden limpiar para que realmente sirvan. En este reporte se hablará sobre la procedencia de los datos, sus tipos y fuentes, así como algunas técnicas básicas de limpieza que ayudan a que la información sea más confiable y útil para tomar decisiones.

Procedencia de los datos

Para este caso de estudio, se analizará un conjunto de datos relacionados con las compras en línea de una tienda de comercio electrónico ficticia llamada JHShop y los datos provienen de distintas fuentes como:

- Datos biométricos:

  No aplican directamente en este caso, pero podrían usarse en autenticación.

- Máquina a máquina:

  Los registros generados por sensores en los servidores que monitorean el tráfico web (tiempos de respuesta, fallos, disponibilidad).

- Datos de transacciones:

  Historial de compras de los clientes.

- Datos generados por humanos:

  Comentarios de clientes sobre los productos, calificaciones y solicitudes de soporte.

- Datos web:

  Clics en el sitio web, páginas visitadas, tiempo de permanencia.

- Redes sociales:

  Menciones y reseñas que los usuarios publican en plataformas como Facebook o Instagram sobre la tienda y los productos.

Tipos y fuentes de datos

Los datos se clasifican de la siguiente manera:

- Cuantitativos:

  Montos de compra, número de artículos adquiridos, tiempos de sesión, clics.

- Cualitativos:

  Comentarios de clientes y opiniones en redes sociales.

- Nominales:

  Métodos de pago, categorías de producto.

- Ordinales:

  Calificación de productos.

- Estructurados:

  Historial de transacciones almacenadas en bases de datos relacionales.

- No estructurados:

  Comentarios en redes sociales, reseñas en texto libre, imágenes de productos compartidas.

Técnicas de limpieza de datos

Durante la exploración inicial del conjunto de datos, se identificaron los siguientes problemas y soluciones:

- Valores nulos:

  en algunos registros faltaba el método de pago.
  Solución:
  imputar con el valor más frecuente o eliminar registros según el impacto.

- Datos duplicados:

  existían compras registradas dos veces.
  Solución:
  aplicar deduplicación con base en ID de transacción.

- Errores de formato:

  fechas en distintos formatos.
  Solución:
  estandarizar al formato ISO 8601.

- Valores atípicos:

  montos de compra extremadamente altos por errores de captura.
  Solución:
  detección con métodos estadísticos y corrección manual.

- Texto desordenado en opiniones:

  comentarios con emojis y caracteres especiales que afectan el análisis.
  Solución:
  aplicar técnicas de limpieza de texto.

Fundamentación

El manejo adecuado de la procedencia, clasificación y limpieza de los datos es esencial para obtener resultados confiables en el análisis. Como señalan Han, Pei y Kamber, la calidad de los datos determina la calidad de los resultados en minería de datos. Además, autores como Provost y Fawcett destacan que el 80% del trabajo en ciencia de datos consiste en preparar y limpiar los datos antes de analizarlos. Finalmente, Davenport y Harris enfatizan la importancia de integrar datos de múltiples fuentes para obtener una visión estratégica del cliente.

Conclusión

Al analizar este tema me di cuenta de que los datos por sí solos no son tan valiosos si no se trabajan correctamente. Es necesario entender de dónde vienen, ya que no es lo mismo un dato generado por una transacción que uno sacado de un comentario en redes sociales. También es importante clasificarlos porque cada tipo de dato necesita un trato distinto. Finalmente, la limpieza de datos es fundamental, ya que corrigiendo errores, eliminando duplicados y dando un formato uniforme, la información se vuelve mucho más útil. En conclusión, este proceso es básico para que los análisis sean confiables y no se tomen decisiones equivocadas por información incorrecta.

Bibliografía

IBM. (s.f.). What Is Data Cleaning? Recuperado de https://www.ibm.com/think/topics/data-cleaning
Tableau. (s.f.). Data Cleaning: Definition, Benefits, And How-To. Recuperado de https://www.tableau.com/learn/articles/what-is-data-cleaning
Acceldata. (2024). Essential Data Cleaning Techniques for Improved Data Quality. Recuperado de https://www.acceldata.io/blog/data-cleaning-made-easy-with-tools-techniques-and-best-practices