Universidad Tecnológica de Chihuahua
Tecnologías de la Información

Reporte de limpieza de datos

Alumno:

Jatzel Israel Cruz Castruita

Grupo:

IDGS91N

Materia:

Extracción de Conocimiento en Bases de Datos

Docente:

Enrique Mascote

Índice

Introducción

El presente documento tiene como objetivo realizar un análisis integral de un conjunto de datos procedente de un sistema de comercio electrónico, con el propósito de identificar su procedencia, clasificar sus tipos y fuentes, así como aplicar técnicas de limpieza que garanticen la calidad, consistencia y confiabilidad de la información, lo cual es fundamental para la generación de reportes precisos y la toma de decisiones basadas en datos confiables. En la primera sección se aborda la procedencia de los datos, describiendo detalladamente su origen, ya sea generado por humanos, máquina a máquina, desde la web o mediante redes sociales, y se explica la relevancia de cada fuente dentro del contexto del análisis, destacando cómo cada tipo de dato contribuye a obtener una visión integral del comportamiento del consumidor y del funcionamiento del negocio. La segunda sección se centra en la clasificación de los datos según sus características cuantitativas, cualitativas, nominales, ordinales, estructuradas, no estructuradas y semiestructuradas, mostrando ejemplos concretos que permiten comprender mejor su aplicación práctica y cómo cada tipo de dato se relaciona con las distintas fuentes de información, tanto internas como externas. Finalmente, la tercera sección detalla las técnicas de limpieza de datos aplicadas, identificando los principales problemas detectados, como valores nulos, duplicados, errores de formato, inconsistencias y valores atípicos, y explicando las acciones correctivas implementadas para cada caso, enfatizando la importancia de la normalización y la consistencia de los registros para garantizar la fiabilidad del análisis posterior. En conjunto, este documento proporciona una visión completa y estructurada del manejo de datos en el caso de estudio, resaltando la importancia de su correcta gestión y procesamiento para obtener información útil y relevante que apoye la toma de decisiones estratégicas.

Procedencia de los datos

El conjunto de datos utilizado en este caso de estudio proviene de un sistema de comercio electrónico, diseñado para registrar las transacciones realizadas por los clientes en una tienda virtual. Este tipo de información surge a partir de diferentes fuentes digitales que intervienen en el proceso de compra y gestión de usuarios, lo que permite disponer de una base de datos amplia y diversa para el análisis.

En primer lugar, los datos generados por humanos incluyen la información que los clientes proporcionan directamente al interactuar con la plataforma. Ejemplos de ello son los nombres, direcciones, correos electrónicos, métodos de pago y opiniones sobre los productos. Este tipo de datos refleja el comportamiento y las preferencias personales de los consumidores, por lo que resulta útil para analizar patrones de compra, niveles de satisfacción y fidelización.

Por otra parte, los datos máquina a máquina, conocidos como M2M, son aquellos que el sistema genera de manera automática durante el registro de las operaciones. Entre ellos se encuentran los identificadores de transacción, las marcas de tiempo de fecha y hora, los códigos de producto y los registros de actividad del usuario dentro del sitio. Estos datos son estructurados y precisos, y contribuyen a la trazabilidad y al control de las operaciones.

También se recopilan datos provenientes de la web, generados a través del monitoreo de la actividad del usuario en la página. Estos datos, obtenidos mediante herramientas analíticas como Google Analytics, permiten conocer aspectos como el número de visitas, la duración de las sesiones, los clics realizados o los productos más consultados. Este tipo de información facilita la evaluación del rendimiento del sitio y la efectividad de las estrategias de marketing.

Finalmente, se consideran los datos derivados de redes sociales, como los comentarios, valoraciones o reacciones de los usuarios en plataformas como Facebook, Instagram o X. Aunque estos datos son no estructurados, ofrecen una perspectiva valiosa sobre la percepción del público respecto a la marca y los productos, y pueden ser procesados mediante técnicas de análisis de texto o minería de opiniones.

En conjunto, esta diversidad de fuentes ofrece una visión integral del negocio y del comportamiento del consumidor, combinando datos estructurados y no estructurados para realizar análisis más completos y obtener conclusiones sobre tendencias y satisfacción del cliente.

Tipos y fuentes de datos

El conjunto de datos del sistema de comercio electrónico se compone de distintos tipos de información que permiten analizar las operaciones de venta, el comportamiento de los clientes y la eficiencia del negocio. Para su correcta interpretación, los datos se clasifican según sus características y estructura.

Tipos de datos según su naturaleza

Datos cuantitativos:

Son aquellos que pueden medirse y expresarse numéricamente. Permiten realizar cálculos estadísticos y análisis comparativos.

- Ejemplos: monto total de cada compra, cantidad de productos vendidos, número de visitas diarias al sitio, tiempo promedio de sesión, descuentos aplicados.
- Aplicación: sirven para elaborar reportes financieros, medir el rendimiento de ventas y detectar tendencias de consumo.

Datos cualitativos:

Representan características o atributos que no se expresan con números, pero son útiles para comprender aspectos subjetivos del comportamiento del cliente.

- Ejemplos: método de pago tarjeta, PayPal, transferencia, categoría del producto ropa, electrónica, hogar, o comentarios en las reseñas.
- Aplicación: ayudan a identificar patrones de preferencia, percepción de marca y satisfacción del usuario.

Datos nominales:

Corresponden a categorías que no tienen un orden específico.

- Ejemplos: país de origen, género del cliente, tipo de producto o método de envío.
- Aplicación: permiten segmentar la información y agrupar registros para análisis demográficos o logísticos.

Datos ordinales:

Se clasifican en categorías que sí presentan un orden jerárquico.

- Ejemplos: nivel de satisfacción bajo, medio, alto, prioridad de atención al cliente baja, media, alta.
- Aplicación: útiles para medir la calidad del servicio y el nivel de aceptación de los productos.

Tipos de datos según su estructura

Datos estructurados:

Son aquellos organizados en tablas, filas y columnas dentro de bases de datos relacionales o archivos CSV. Tienen un formato definido que facilita su almacenamiento y consulta.

- Ejemplos: registros de ventas con campos como ID de transacción, fecha, monto, producto, cliente y método de pago.
- Fuente: bases de datos internas del sistema de ventas o registros exportados desde el software de gestión.
- Aplicación: se emplean para generar reportes automáticos y análisis estadísticos mediante herramientas como Excel, SQL o Power BI.

Datos no estructurados:

Carecen de un formato fijo y pueden provenir de texto libre, imágenes, audios o publicaciones. Requieren técnicas de procesamiento avanzadas para su análisis.

- Ejemplos: comentarios y reseñas de clientes, mensajes en redes sociales, correos de atención al cliente.
- Fuente: plataformas externas como Facebook, Instagram o formularios abiertos en la web.
- Aplicación: se utilizan para analizar la percepción del cliente, detectar problemas de servicio o identificar oportunidades de mejora.

Datos semiestructurados:

Contienen cierta organización, aunque no tan rígida como una tabla. Se representan comúnmente en formatos como JSON o XML.

- Ejemplos: registros de actividad del usuario, datos de navegación, información capturada por cookies o APIs.
- Fuente: servidores web y herramientas de análisis digital.
- Aplicación: ayudan a estudiar patrones de comportamiento en línea y optimizar la experiencia del usuario

Integración de las fuentes

En este caso de estudio, los datos provienen de fuentes internas, conocidas como sistema de ventas, base de clientes y registros de facturación, así como de fuentes externas, como redes sociales, herramientas de analítica web y encuestas de satisfacción. Esta combinación de información estructurada y no estructurada ofrece una perspectiva más completa del entorno comercial, permitiendo comprender tanto el rendimiento operativo como la experiencia del consumidor.

Técnicas de limpieza de datos

El análisis inicial del conjunto de datos reveló diversos problemas que podían afectar la calidad de la información y, por tanto, la confiabilidad de los resultados. La limpieza de datos consistió en identificar estos problemas y aplicar acciones correctivas adecuadas, siguiendo buenas prácticas de gestión de datos.

Valores nulos o faltantes

Los valores nulos se presentaban en campos importantes, como correo electrónico, número de contacto y dirección de envío.

- Problema: registros incompletos que podían dificultar el contacto con el cliente o la validación de transacciones.
- Acción correctiva: Se reemplazaron valores nulos con la etiqueta “Desconocido” cuando la información no era obligatoria para el análisis general.
  En campos críticos como monto de compra o ID de transacción, los registros incompletos se eliminaron para evitar errores en los reportes y cálculos estadísticos.

Datos duplicados

Se detectaron registros repetidos de transacciones debido a errores en la integración del sistema o reenvíos de formularios.

- Problema: inflaban artificialmente los resultados y generaban inconsistencias.
- Acción correctiva: se implementó un proceso de deduplicación utilizando el ID de transacción, la fecha y la hora, conservando solo el registro original y eliminando las repeticiones.

‘

Errores de formato

Al revisar la base de datos, se encontraron inconsistencias en la forma de registrar fechas, métodos de pago y nombres de productos.

- Problema: formatos diferentes dificultaban el análisis automático y la generación de reportes.
- Acción correctiva:
    - Se unificó el formato de fechas a AAAA-MM-DD.
    - Se corrigieron las inconsistencias en los nombres de productos y formas de pago, asegurando que todas las entradas sigan un mismo formato uniforme, por ejemplo, todas las referencias a “tarjeta” se registraron en minúsculas.
    - Se estandarizó la notación de montos monetarios, eliminando símbolos o espacios innecesarios.

Valores atípicos

Se identificaron montos de compra extremadamente altos o negativos, que no correspondían al comportamiento habitual del negocio.

- Problema: podían distorsionar el cálculo de promedios, totales y tendencias.
- Acción correctiva:
    - Se revisaron los registros sospechosos de forma manual para confirmar si eran errores.
    - Los registros incorrectos se corrigieron cuando fue posible; en casos dudosos, se eliminaron del análisis.

Inconsistencias en datos categóricos

Algunos campos categóricos, como país de residencia o nivel de satisfacción, contenían errores de escritura o categorías mal definidas.

- Problema: dificultaban el agrupamiento y análisis de los datos.
- Acción correctiva: se aplicó un proceso de normalización, unificando categorías y corrigiendo errores ortográficos.

Conclusiones

Creo que este análisis me ayudó a entender mejor de dónde vienen los datos y por qué es tan importante conocer su origen para poder trabajar con ellos correctamente. Al clasificar los tipos y fuentes, me di cuenta de la cantidad de información diferente que existe y de cómo cada tipo, ya sea cuantitativo, cualitativo, estructurado o no estructurado, tiene su propio valor y forma de ser analizado. También aprendí que aplicar técnicas de limpieza no es solo eliminar errores, sino asegurarse de que los datos sean confiables y útiles, corrigiendo valores nulos, duplicados, errores de formato o inconsistencias. Personalmente, este ejercicio me hizo ver lo importante que es mantener la información organizada y correcta, ya que de eso depende que los resultados de un análisis sean precisos y que podamos tomar decisiones acertadas.

Bibliografía

Mucci, T. (2024, 23 de julio). ¿Qué es la procedencia de los datos? IBM. https://www.ibm.com/mx-es/think/topics/data-provenance

Grupo Winecta. (2022, 20 de junio). Tipos de datos en Big Data. https://winecta.com/tipos-de-datos-en-big-data/

Escuela de Postgrado de la Universidad Católica San Pablo. (2019, 5 de marzo). ¿Qué es el big data y cuáles son sus beneficios? https://postgrado.ucsp.edu.pe/articulos/que-es-big-data/

QuestionPro. (2021, 16 de diciembre). Tipos de fuentes de datos. https://www.questionpro.com/blog/es/fuentes-de-datos

Great Learning Editorial Team. (2024, 15 de agosto). 4 Tipos de Datos: Nominal, Ordinal, Discreto y Continuo. https://www.mygreatlearning.com/blog/tipos-de-datos

IBM. (2024, 23 de julio). ¿Qué es la limpieza de datos? https://www.ibm.com/mx-es/think/topics/data-cleaning

OBS Business School. (2024, 15 de abril). Técnicas de data cleaning para garantizar datos de calidad. https://www.obsbusiness.school/blog/tecnicas-de-data-cleaning-para-garantizar-datos-de-calidad

ISOL. (2024, 30 de julio). Limpieza de datos: definición, técnicas y mejores prácticas para 2024. https://www.isol.mx/mentoring/limpieza-de-datos