EXTRACCIÓN DE CONOCIMIENTOS EN BASES DE DATOS

ING. LUIS ENRIQUE MASCOTE CANO.

Universidad Tecnológica de Chihuahua

REPORTE DE LIMPIEZA DE DATOS

Lic. Ricardo Hernández Martínez

Fecha de Entrega: 5/OCTUBRE/2025

Índice

Introducción

En la actualidad, la recopilación, análisis y gestión de datos constituyen una de las actividades más estratégicas para las organizaciones. El sector retail (comercio minorista) es un claro ejemplo de cómo los datos, provenientes de múltiples fuentes, permiten comprender mejor a los clientes, optimizar operaciones y diseñar estrategias de mercado más efectivas. El presente reporte expone un caso de estudio simulado basado en una cadena minorista de supermercados que implementa un sistema de análisis de datos para la mejora de su servicio al cliente y la eficiencia en la gestión de inventario. Se describirá la procedencia de los datos, los tipos y fuentes empleados, así como las técnicas de limpieza aplicadas para garantizar la calidad y confiabilidad del análisis.

Procedencia de los datos

La procedencia de los datos hace referencia al origen de la información, es decir, a las fuentes desde donde se genera y recopila. En el caso de estudio de la cadena de supermercados, se identifican las siguientes:

- **Datos transaccionales**
  Provienen de los sistemas de punto de venta (POS) y reflejan las operaciones comerciales realizadas diariamente. Estos datos incluyen precios de productos, cantidades adquiridas, medios de pago y horarios de compra. Su importancia radica en que permiten identificar patrones de consumo, segmentar clientes y evaluar la rentabilidad de las categorías de productos (Kimball & Ross, 2013).
- **Datos máquina a máquina (M2M)**
  Los supermercados modernos utilizan sensores en refrigeradores y estantes inteligentes que reportan en tiempo real la cantidad de productos disponibles, la temperatura de los equipos y alertas por fallas. Este tipo de datos permite reducir pérdidas por caducidad y mejorar la logística del reabastecimiento (Romero & Ventura, 2021).
- **Datos generados por humanos**
  Los clientes aportan información a través de encuestas de satisfacción, formularios digitales y comentarios enviados a los canales de atención. Este tipo de datos es esencial porque aporta una visión cualitativa sobre las percepciones, expectativas y emociones de los consumidores.
- **Datos web**
  Al contar con un sitio de comercio electrónico, la empresa obtiene información sobre navegación, clics, carritos de compra abandonados, búsquedas de productos y tiempo de permanencia en la página. Estos datos permiten analizar la usabilidad del portal y diseñar estrategias de marketing digital más efectivas.
- **Datos de redes sociales**
  Opiniones, reseñas y publicaciones en plataformas como Facebook, Twitter

e Instagram constituyen una fuente de datos no estructurados. Su análisis mediante técnicas de *sentiment analysis* brinda a la empresa información sobre la reputación de la marca y la aceptación de campañas promocionales.

En conjunto, estas fuentes conforman un ecosistema de datos complejo y heterogéneo, cuyo aprovechamiento adecuado genera ventajas competitivas.

Tipos y fuentes de datos

Una vez identificada la procedencia, resulta fundamental clasificar los datos según sus características. Para el caso de la cadena de supermercados, se distinguen los siguientes:

- Cuantitativos estructurados: Incluyen precios de productos, volúmenes de venta, montos de transacciones y niveles de inventario. Se almacenan en bases de datos relacionales y permiten análisis estadísticos y modelado predictivo.
- Cualitativos no estructurados: Comprenden las opiniones de clientes en redes sociales, reseñas en línea y respuestas abiertas en encuestas. Estos datos, aunque difíciles de procesar, son clave para comprender el sentimiento del consumidor y sus motivaciones (Han, Kamber & Pei, 2012).
- Nominales: Se refieren a categorías como tipo de producto (frutas, verduras, lácteos, bebidas, limpieza) o formas de pago (efectivo, tarjeta, transferencia).
- Ordinales: Son los que presentan un orden, por ejemplo, la valoración de satisfacción en una escala de 1 a 5 o la clasificación de productos según frescura (alto, medio, bajo).
- Datos estructurados: Se concentran principalmente en las transacciones y en la información de inventarios reportada por sensores. Tienen un formato definido y se integran fácilmente en sistemas de gestión empresarial.
- Datos no estructurados: Corresponden a textos libres, imágenes o videos publicados por los clientes en redes sociales. Su análisis requiere técnicas de minería de texto y algoritmos de inteligencia artificial.

La diversidad de datos hace evidente la necesidad de herramientas tecnológicas capaces de integrarlos. Por ejemplo, un sistema de *Business Intelligence* puede correlacionar la venta de productos frescos (dato estructurado) con la opinión de clientes en Twitter sobre su calidad (dato no estructurado).

Técnicas de limpieza de datos

La limpieza de datos es el proceso de detección y corrección de errores o inconsistencias en el conjunto de información. Este paso resulta esencial porque la calidad de los datos influye directamente en la precisión de los análisis y modelos predictivos (Han et al., 2012). En el caso del supermercado, se identifican los siguientes problemas y técnicas de solución:

1. Valores nulos o incompletos
   Algunas encuestas presentan campos sin respuesta, y ciertos registros de inventario reportados por sensores llegan incompletos. La acción correctiva consiste en la imputación de valores promedio para variables numéricas y, en el caso de datos cualitativos, la exclusión de registros sin información relevante.

2. Datos atípicos (outliers)
   Se detectaron transacciones con montos desproporcionadamente altos en comparación con el promedio, lo cual puede indicar errores de captura o intentos de fraude. Para corregirlos, se aplican algoritmos de detección de outliers y reglas de negocio que validan el rango esperado de valores.

3. Errores de formato
   Se encontraron inconsistencias en el registro de fechas (ejemplo: 05/07/24 puede interpretarse como 5 de julio o 7 de mayo). La solución fue estandarizar el formato a aaaa-mm-dd durante el proceso de integración ETL.

4. Registros duplicados
   La base de datos de clientes de programas de lealtad contenía usuarios registrados con correos electrónicos diferentes pero con la misma

información personal. Se aplicaron algoritmos de deduplicación basados en identificadores únicos (ID del cliente).

5. Normalización de texto no estructurado
   En las reseñas de redes sociales se detectaron abreviaciones, errores ortográficos y uso de emoticonos. Para tratarlos, se emplearon técnicas de *text mining* y procesamiento de lenguaje natural (PLN), con el fin de estandarizar la escritura y clasificar los sentimientos expresados.

Estas técnicas garantizan que los datos tengan integridad, consistencia y calidad, lo que permite a la empresa generar reportes confiables y realizar predicciones acertadas.

Conclusiones

La gestión adecuada de los datos en el sector retail representa una ventaja competitiva significativa. El presente caso de estudio simulado muestra cómo los datos provienen de diversas fuentes: biométricos indirectos (sensores), máquina a máquina, transacciones, generados por humanos, web y redes sociales. Asimismo, evidencia la importancia de clasificar los datos según su naturaleza para aplicar las técnicas de análisis adecuadas. Finalmente, se destacó que la limpieza de datos es un paso crítico, ya que la calidad de la información incide directamente en la precisión de los modelos analíticos y en la confiabilidad de las decisiones empresariales. El proceso de integración, clasificación y depuración de datos convierte la información bruta en conocimiento útil, alineado con la necesidad de innovación en la gestión empresarial actual.

Referencias

García, A., & López, M. (2020). *Gestión de datos y analítica en el sector retail*. Revista Iberoamericana de Tecnología, 12(3), 45-59.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.

Romero, D., & Ventura, S. (2021). *Big Data y analítica avanzada en entornos empresariales*. Editorial UOC.