05/10/2025

# Reporte de limpieza de datos

*Extracción de Conocimiento en Bases de Datos*

Luis Eduardo Aguilar Sarabia

*IDGS91N*

Docente: Luis Enrique Mascote Cano

# Introducción

El rápido crecimiento de los datos en el sector tecnológico ha hecho indispensable comprender su procedencia, estructura y calidad para tomar decisiones informadas. Este caso de estudio simula un proyecto de análisis de incidencias de soporte técnico en una empresa de TI. El objetivo es evaluar cómo se recopilan, clasifican y preparan los datos para su análisis. Se describe el origen de la información (biométrica, máquina a máquina, transaccional, generada por humanos y de redes sociales), los tipos de datos recogidos, los problemas de calidad hallados y las técnicas de limpieza aplicadas para garantizar un conjunto de datos confiable.

# Procedencia de los datos

Los datos utilizados en este estudio provienen de diversas fuentes relacionadas con la operación de un servicio de soporte técnico. La clasificación de Forodatos distingue cinco tipos de datos en el contexto de Big Data: grandes transacciones, redes sociales/páginas web, biométricos, generados por humanos y máquina a máquina (M2M). A continuación, se describe cómo se aplican estas categorías en el caso simulado:

- **Datos transaccionales:** son registros de facturación y de llamadas asociadas a los tickets de soporte. Incluyen información de clientes obtenida de sistemas CRM y ERP, inventarios de incidencias y métricas de respuesta. Estos datos se almacenan en bases relacionales y constituyen la parte estructurada del conjunto.
- **Datos de redes sociales y web:** se recogen comentarios publicados en las redes sociales de la empresa (Twitter, LinkedIn) y en foros de soporte. Este tipo de información se utiliza para analizar tendencias de satisfacción y detectar incidentes emergentes.
- **Datos biométricos:** la compañía utiliza sistemas de control de acceso basados en huellas dactilares para sus centros de datos. Los registros de entradas y

salidas de personal permiten cruzar la disponibilidad de técnicos con las incidencias, y forman parte de los datos biométricos.

- **Datos generados por humanos:** corresponden a correos electrónicos, documentos internos, notas de voz y pagos con tarjeta que realizan los empleados al adquirir software interno. También se incluyen encuestas de satisfacción donde los usuarios describen su experiencia con el soporte.
- **Datos máquina a máquina (M2M):** se obtienen de sensores conectados a los servidores y a la red de la empresa. El Internet de las Cosas ha multiplicado este tipo de datos; los dispositivos registran temperatura de los equipos, uso de CPU, ancho de banda y alertas de hardware. Estos registros se reciben en formato JSON y se combinan con las incidencias para identificar patrones predictivos.

## Tipos y fuentes de datos

La naturaleza de los datos determina cómo deben ser tratados. Según su naturaleza, los datos pueden ser cuantitativos o cualitativos. También se clasifican por su estructura en no estructurados, semiestructurados y estructurados. El cuadro siguiente resume la tipología de las variables del caso simulado:

1. **Datos cuantitativos (numéricos)**:
    - *Continuos*: variables como el tiempo de resolución de un ticket (horas), la temperatura de un servidor (°C) o el uso de CPU (%). Se miden en escalas continuas y pueden adoptar valores decimales
    - *Discretos*: número de tickets por día, número de técnicos asignados o número de sensores activos. Solo admiten valores enteros.
2. **Datos cualitativos (categóricos)**:
    - *Nominales*: variables sin orden, como el departamento que reporta la incidencia (desarrollo, operaciones, soporte) o el tipo de incidente (hardware, software, red). Son atributos sin jerarquía.

- *Ordinales*: variables con un orden lógico, como el nivel de severidad de la incidencia (bajo, medio, alto) o la clasificación del nivel socioeconómico del cliente. Para estas variables es importante respetar el orden durante el análisis.

3. **Datos estructurados**: provienen de bases de datos relacionales y hojas de cálculo; son registros con una longitud y formato definidos. En el caso estudiado, los tickets de soporte, los registros de control de acceso y las métricas de sensores se almacenan en tablas SQL, lo que facilita su consulta mediante SQL.
4. **Datos semiestructurados**: son datos parcialmente organizados, como archivos JSON provenientes de sensores, mensajes XML o correos electrónicos que contienen campos estructurados (remitente, asunto) y texto libre. Estas fuentes permiten un mayor grado de flexibilidad que una base de datos relacional y requieren herramientas como Pandas para su procesamiento.
5. **Datos no estructurados**: incluyen textos de encuestas, descripciones de incidentes, comentarios en redes sociales y grabaciones de voz. Este tipo de datos carece de una estructura rígida y representa aproximadamente el 80–90 % de los datos generados actualmente. Se requieren técnicas de procesamiento del lenguaje natural para extraer información útil.

## Problemas de calidad y técnicas de limpieza de datos

El análisis de datos de soporte técnico reveló varios problemas comunes en los conjuntos de datos, que debían abordarse antes de aplicar modelos predictivos:

- **Datos duplicados**: se identificaron tickets y registros de sensores repetidos, lo que podría sesgar las estadísticas. Una de las técnicas comunes de data cleaning es comprobar si existen filas duplicadas y eliminarlas, ya que los duplicados generan ruido y afectan los resultados. Para ello se usó un

procedimiento de deduplicación basado en identificadores únicos y comparación de campos clave.

- **Valores faltantes (nulos)**: muchas entradas carecían de información sobre la satisfacción del cliente o no incluían la temperatura registrada por los sensores en determinados momentos. Para tratarlos, se puede eliminar las filas incompletas o imputar valores utilizando la media, la mediana u otros algoritmos avanzados. En este caso, se imputó la mediana para variables numéricas y se creó una categoría “No responde” para variables cualitativas.
- **Inconsistencias en formatos**: los datos de fecha y hora se almacenaban en formatos distintos (por ejemplo, DD/MM/AAAA y MM-DD-AAAA). Establecer un formato común es esencial para la coherencia. Se normalizó la estructura de fechas al estándar ISO 8601 y se unificaron los nombres de los departamentos en minúsculas.
- **Valores atípicos (outliers)**: se detectaron tiempos de resolución extremadamente altos o temperaturas de servidores fuera de rango. Identificar y eliminar valores extremos evita distorsiones en los análisis estadísticos. Se aplicaron métodos basados en el rango intercuartil y la desviación estándar para filtrar estas observaciones.
- **Errores tipográficos y conversión de tipos**: algunas entradas presentaban errores de escritura o se registraron como cadena cuando debían ser numéricas. Convertir cada columna al tipo de dato correcto (fechas, enteros, booleanos) facilita el procesamiento posterior. Se diseñaron reglas de validación para corregir edades fuera de los rangos plausibles y para asegurar que las edades de los técnicos no superaran los límites fisiológicos.

Además de estas acciones, se consideró la validación de reglas de negocio, consistente en comprobar que los valores cumplan criterios lógicos (por ejemplo, que un ticket resuelto tenga un tiempo de resolución mayor a cero). También se combinaron herramientas de software como Python (pandas) para operaciones avanzadas, Excel para tareas básicas y Power BI para integrar la limpieza con la visualización de datos.

# Conclusiones

Este caso de estudio simulado demuestra la importancia de conocer la procedencia y la estructura de los datos antes de analizarlos. Los datos utilizados provienen de múltiples fuentes: transaccionales, redes sociales, biométricos, generados por humanos y máquina a máquina que requieren tratamientos diferenciados. Clasificar las variables como cuantitativas o cualitativas, y como estructuradas, semiestructuradas o no estructuradas, permite seleccionar herramientas y técnicas de análisis adecuadas.

Los problemas de calidad identificados – duplicados, valores nulos, formatos inconsistentes, outliers y errores tipográficos – son comunes en proyectos de ciencia de datos. Aplicar técnicas de limpieza como la eliminación de duplicados, imputación de valores faltantes, normalización de formatos, detección de outliers y validación de reglas de negocio mejora la precisión de los modelos y refuerza la confianza en los resultados. De este modo, la empresa puede realizar análisis fiables para optimizar la asignación de técnicos, prever picos de incidencias y mejorar la satisfacción de sus clientes.

# Referencias


Forodatos. (2025). *¿Qué son los datos? ¿Cómo se clasifican?* Recuperado de https://forodatos.com/data-science/que-son-los-datos/

The Bridge. (s.f.). *Diferencias entre datos estructurados, no estructurados y semiestructurados*. Recuperado de https://thebridge.tech/blog/diferencias-entre-datos-estructurados-no-estructurados-y-semiestructurados/

García Muñoz, J. L. (2025). *Data cleaning: qué es, técnicas y cómo aplicarlo correctamente*. ESEID AI Business School. Recuperado de https://eseid.com/data-cleaning-que-es-tecnicas-y-como-aplicarlo-correctamente/

DataCamp. (2025). *Cómo limpiar datos en Excel: Guía para principiantes*. Recuperado de https://www.datacamp.com/es/tutorial/data-cleaning-in-excel-a-beginners-guide