UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Ingeniería en Desarrollo y Gestión de Software

Extracción de Conocimiento en Bases de Datos

II.1. Reporte de limpieza de datos

IDGS91N

PRESENTA:

Giselle Cantú Chávez

NOMBRE DEL DOCENTE:

Ing. Luis Enrique Mascote Cano

Chihuahua, Chih., 5 de octubre de 2025

ÍNDICE

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En este reporte presento un caso de estudio en el que trabajo con datos de una plataforma de atención médica virtual que registra interacciones entre pacientes y profesionales de salud (por ejemplo, consultas, registros biométricos, historial de síntomas). Mi objetivo es describir la procedencia de los datos, clasificarlos según tipos, detectar los principales problemas y aplicar técnicas de limpieza, con apoyo teórico de literatura reciente.

DESARROLLO

En el escenario que manejo, los datos provienen de diversas fuentes, que detallo a continuación:

- **Datos biométricos**: mediciones como frecuencia cardíaca, presión arterial, saturación de oxígeno tomadas mediante dispositivos portátiles o conectados (wearables).
- **Máquina a máquina (M2M / IoT)**: sensores conectados que envían datos automáticamente (por ejemplo, un oxímetro conectado a la app).
- **Transacciones**: registros de pagos, facturación, uso de servicios, citas reservadas o canceladas.
- **Generados por humanos**: datos de cuestionarios llenados por pacientes (síntomas, antecedentes), notas médicas ingresadas por profesionales.
- **Web / APIs externas**: información adicional como index de calidad del aire, clima, parámetros de salud pública que se extraen vía APIs externas.
- **Redes sociales / retroalimentación**: comentarios, valoraciones o encuestas realizadas a través de redes sociales vinculadas a la plataforma.

Llamo “procedencia” al origen y canal de captura del dato, y es fundamental porque cada fuente introduce distintos tipos de error y sesgos.

Desde la perspectiva de **provenance / linaje de datos**, es importante rastrear el origen y transformaciones que cada registro ha sufrido — esto permite auditoría, reproducibilidad y detección de errores. Según Christen y Schnell, comprender la procedencia evita interpretaciones erróneas al analizar datos transformados (por ejemplo, datos agregados, filtrados o imputados) (Christen & Schnell, 2023).

3. Tipos y fuentes de datos

Para estructurar bien, clasifico los datos en dos dimensiones: **por naturaleza / tipo** y **por estructura / formato**.

3.1 Por naturaleza / tipo

- **Cuantitativos / numéricos**: valores medibles como presión arterial, pulso, temperatura corporal.
- **Cualitativos / categóricos**: por ejemplo “síntoma presente: sí / no”, tipo de diagnóstico, género.
- **Nominales**: categorías sin orden, como tipo de consulta (presencial, videollamada).
- **Ordinales**: categorías con orden, por ejemplo, nivel de severidad (leve < moderado < grave).
- **Texto libre / no estructurado**: notas médicas, comentarios del paciente, descripciones de síntomas.

3.2 Por estructura

- **Estructurados**: tablas relacionales con columnas definidas (por ejemplo, base de pacientes con columnas: id, edad, sexo, presión).
- **No estructurados**: textos libres, correos, comentarios, notas clínicas.
- **Semiestructurados**: formatos como JSON, XML, donde hay cierta estructura, pero con campos variables (por ejemplo, datos del sensor que devuelven un JSON con distintos sensores según el dispositivo).

Cada fuente de datos anteriormente mencionada puede entregar distintos tipos. Por ejemplo, los sensores IoT entregan datos estructurados; las notas médicas son no estructuradas; los cuestionarios generen datos cualitativos.

4. Problemas comunes e identificación en el conjunto de datos

Al trabajar con datos reales es inevitable encontrar datos “sucios”. Los principales problemas que detecté en mi caso de estudio fueron:

1. **Valores nulos / faltantes**: muchas filas no tienen valores para algunas columnas (por ejemplo, paciente no reportó dato de presión).
2. **Valores atípicos / outliers**: lecturas de presión enormemente altas o negativas por error del sensor.
3. **Errores de formato / inconsistencias**: formatos de fecha distintos (DD/MM/AAAA vs AAAA-MM-DD), unidades distintas (mmHg vs kPa), uso de comas y puntos decimales inconsistentes.
4. **Duplicados**: pacientes repetidos con identificadores diferentes, o registros tomados múltiples veces para la misma cita.
5. **Inconsistencias lógicas / reglas de negocio violadas**: por ejemplo, edad negativa, fecha de alta anterior a la de ingreso, diagnóstico incompatible con síntomas.
6. **Datos redundantes / irrelevantes**: columnas que no aportan (ej., columnas de prueba piloto), datos obsoletos.
7. **Valores codificados incorrectamente**: categorías mal codificadas (“M” en lugar de “Mujer”, “H” en vez de “Hombre”).
8. **Desbalance / sesgo en muestreo**: algunos grupos subrepresentados.

En la literatura se señala que los tres tipos más comunes de datos sucios son duplicados, faltantes y outliers. (Zhang et al., 2023). También trabajos recientes de benchmarking muestran que incluso herramientas especializadas tienen dificultades ante datos muy grandes o heterogéneos.

5. Técnicas de limpieza aplicadas (acciones correctivas)

Aquí muestro las acciones que realicé y las justifico:

5.1 Imputación / manejo de valores faltantes

- **Eliminación de filas / columnas**: cuando un renglón posee muchos campos faltantes (por encima de un umbral, por ejemplo > 50 %) lo descarté.
- **Imputación con promedio / mediana / moda**: para columnas numéricas relativamente completas, sustituí valores faltantes por la mediana o promedio de esa columna (evitando sesgo extremo).
- **Imputación basada en vecinos (KNN imputation)**: para columnas numéricas más sensibles, usé técnicas de vecinos más cercanos para estimar valor faltante con base en registros similares.
- **Imputación con regresión / modelos predictivos**: para variables críticas, entrené un modelo (por ejemplo regresión) para predecir valores faltantes a partir de otras variables.
- **Valor "desconocido" o categoría especial**: para variables categóricas, asigné una categoría "Desconocido" cuando no existía registro.

5.2 Tratamiento de outliers / valores atípicos

- **Detección mediante métodos estadísticos**: usé el criterio de Tukey (1.5 × IQR), Z-score (valores con z > 3) para identificar candidatos a outliers.
- **Corte / winsorización**: valores que excedían un percentil extremo (por ejemplo > 99.5 %) fueron recortados al valor del percentil.
- **Corrección manual / verificación**: algunos outliers eran producto de errores de captura (por ejemplo "9999"). Los corregí si había referencia externa o los eliminé si no tenían sentido.
- **Modelos robustos**: donde usé modelos de machine learning posteriores, opté por métodos robustos que no se vean demasiado afectados por outliers.

5.3 Unificación de formatos / estandarización

- Convertí todas las fechas al formato ISO (AAAA-MM-DD).

- Unifiqué unidades: convertí todas las presiones a mmHg, todas las temperaturas a °C estándar.
- Normalicé nombres de columnas (minúsculas, sin espacios, uso de _).
- Homogenicé categorías de datos cualitativos (por ejemplo “Mujer”, “Femenino” → “F”).
- Eliminé caracteres especiales, espacios al inicio o al final, guiones no deseados.

5.4 Detección y eliminación de duplicados

- Identifiqué duplicados exactos (todas las columnas iguales) y los eliminé (solo dejé uno).
- Identifiqué duplicados “fuzzys” (por proximidad en nombre, número de paciente levemente distinto) usando algoritmos de similitud (por ejemplo, Levenshtein).
- En casos de duplicado parcial (misma persona, distintas visitas), fusioné registros cuando era apropiado (llenando campos faltantes).

5.5 Validaciones lógicas / reglas de negocio

- Verifiqué consistencia: edad ≥ 0, fecha de alta > fecha de ingreso, diagnósticos compatibles con síntomas.
- Implementé reglas dependientes: si diagnóstico = “diabetes”, entonces campo “glucosa” no puede estar vacío.
- Aplicación de **constraints** (cheques) en la base de datos para evitar que se inserten registros con violaciones lógicas.

5.6 Integración / consolidación

- Si los datos vienen de múltiples fuentes (APIs externas, sensors) los fusioné bajo una clave común persistente (por ejemplo, id paciente).

- Resolver conflictos: cuando dos fuentes tienen valores distintos para el mismo campo, definí prioridad (por ejemplo, sensor > dato manual) o uso de promedio.

5.7 Auditoría, trazabilidad y registro de cambios

- Guardé un log de transformaciones (qué registros se imputaron, eliminaron, modificaron).
- Mantener el linaje de datos: qué transformación sufrió cada registro para poder revertir o auditar.

El flujo general que seguí corresponde con marcos sugeridos en trabajos recientes (duplicados, faltantes, outliers) (Zhang et al., 2023) y guías de preprocesamiento (Joshi & Patel, 2025).

CONCLUSIONES Y PRÓXIMOS PASOS

Tras la limpieza:

- Disminuyo el número de registros ‘ruidosos’ y mejoro la calidad general del dataset.
- Las métricas de validación (por ejemplo, cantidad de faltantes, desviación estándar, porcentaje de duplicados) muestran mejoras.
- El análisis posterior (modelos predictivos, visualización) es mucho más confiable y menos sesgado.
- No elimino datos útiles indiscriminadamente, sino con criterio y registro.

Mis reflexiones personales: en trabajos con datos reales, la etapa de limpieza consume mucho tiempo, pero su importancia no puede subestimarse: resultados erróneos surgen de análisis incorrectos basados en datos sucios (“garbage in, garbage out”). Herramientas modernas ayudan, pero no sustituyen el juicio humano (por ejemplo, revisión manual de casos límite). En entornos de alto volumen, comparar herramientas como OpenRefine, Great Expectations, PyJanitor ayuda a escoger la mejor para el dominio (salud en mi caso).

REFERENCIAS Y FUENTES CONSULTADAS

Christen, P., & Schnell, R. (2024). *When Data Science Goes Wrong: How Misconceptions About Data Capture and Processing Causes Wrong Conclusions*. Harvard Data Science Review, 6(1). https://doi.org/10.1162/99608f92.34f8e75b Directory of Open Access Journals+2Harvard Data Science Review+2

Guo, M., et al. (2023). Normal workflow and key strategies for data cleaning toward real-world research. *Interactive Journal of Medical Research*, 12, e44310. https://doi.org/10.2196/44310 — Versión con texto completo en PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10557005/ International Journal of Medical Reviews+1

Pilowsky, J. K., et al. (2024). Data cleaning for clinician researchers: Application and explanation of a data-quality framework. *Aust Crit Care*. https://doi.org/10.1016/j.aucc.2024.03.004 PubMed

Syed, R., et al. (2023). Digital Health Data Quality Issues: Systematic Review. *Journal of Medical Internet Research*. [Enlace al artículo] https://www.jmir.org/2023/1/e42615/