# ANALISIS EXPLORATORIO

Alejandro Medina Diaz

José Canto Peral

**1. Introducción**

1.1. Objetivo del análisis
1.2. Descripción general del conjunto de datos
1.3. Contexto del proyecto y relevancia del estudio

---

**2. Análisis Descriptivo Inicial**

2.1. Estructura y dimensiones del conjunto de datos
2.2. Tipología de variables
- Variables categóricas y logicas
- Variables numéricas
- Variables de fecha
- Variables de texto libre

2.3. Revisión de valores nulos, vacíos y desconocidos
2.4. Análisis de duplicados y consistencia de registros
2.5. Identificación de outliers (valores atípicos)
2.6. Estadísticas descriptivas básicas
- Medidas de tendencia central (media, mediana, moda)
- Medidas de dispersión (desviación estándar, rango, percentiles)
- Distribución de variables clave (edad, estancia, coste, etc.)

2.7. Visualización inicial de los datos
- Histogramas y diagramas de caja
- Gráficos de barras y pastel (variables categóricas)
- Correlaciones entre variables numéricas

---

**3. Ingeniería de Características**

3.1. Limpieza y transformación de datos
3.2. Creación de nuevas variables derivadas
- Edad en ingreso (verificada)
- Duración del contacto (Fecha de Fin – Fecha de Inicio)
- Agrupaciones por tipo de diagnóstico o servicio

3.3. Codificación de variables categóricas
- Columnas obsoletas y variables obsoletas

3.4. Normalización y escalado de variables numéricas
3.5. Tratamiento de fechas (extracción de mes, año, estación, etc.)
3.6. Selección preliminar de características relevantes para análisis posterior

---

**4. Conclusiones del Análisis Exploratorio**

4.1. Hallazgos clave
4.2. Limitaciones del conjunto de datos
4.3. Recomendaciones para fases siguientes (modelado o inferencia)

---

**5. Anexos**

5.1. Tablas descriptivas completas
5.2. Gráficos adicionales
5.3. Implementacion en Oracle

## Introducción

### 1.1 Objetivo del análisis

El objetivo principal es realizar un Análisis Exploratorio de Datos (AED) exhaustivo sobre el conjunto completo de registros clínicos proporcionado, con foco en caracterizar la estructura y calidad de los datos, identificar patrones relevantes y preparar la base para la fase de modelado y los entregables del hito. Este AED debe detectar errores y valores atípicos, cuantificar datos faltantes, validar coherencias entre campos clínicos y administrativos, y generar un conjunto reproducible de variables derivadas y reglas de limpieza que permitan análisis posteriores (cohortes, predicción de reingreso, análisis de costes, etc.). El resultado esperado es: (a) un paquete de artefactos reproducibles (scripts R, CSVs resumidos, gráficos), (b) un documento con hallazgos y recomendaciones, y (c) una lista priorizada de variables derivadas para el siguiente hito.

### 1.2 Descripción general del conjunto de datos

El dataset es un registro administrativo-clínico que contiene, por paciente y episodio, variables administrativas, demográficas, fechas de eventos, diagnósticos codificados (CIE), procedimientos, medidas de severidad y costes. Tipos de columnas presentes y su función principal:

- Identificadores y contexto: Centro recodificado, CIP SNS, Número de registro anual, Nombre (identificativo, debe anonimizarse).
- Geografía y procedencia: Comunidad Autónoma, CCAA Residencia, País Nacimiento, País Residencia.
- Fechas: Fecha de nacimiento, Fecha de Ingreso, Fecha de Inicio contacto, Fecha de Fin contacto, Fecha de Intervención.
- Demografía y estado: Sexo, Edad, Edad en ingreso, Mes de ingreso.
- Codificación clínica y administrativa: Diagnóstico Principal, Diagnóstico 2..20, POA Diagnóstico 1..20, CIE, Servicio, Categoría, Tipo Alta, Régimen Financiación, Procedencia.
- Intervenciones y procedimientos: Procedimiento 1..20, Procedimiento Externo 1..6, CDM/GRD/AP/IR y sus variantes.
- Resultados y recursos: Estancia Días, Ingreso en UCI, Días UCI, Reingreso, Coste APR, Valor Peso Español/Americano, Nivel Severidad APR, Riesgo Mortalidad APR.
  Observaciones de calidad preliminares detectadas en la muestra: formatos de fecha mixtos (mm/dd/yyyy y dd/mm/yyyy), identificadores con notación científica o cadenas no uniformes, textos descriptivos en columnas donde

también hay códigos (p.ej. texto del diagnóstico junto a su código CIE), y presencia de placeholders o códigos especiales (ej. ZZZ) que requieren tratamiento.

### 1.3 Contexto del proyecto y relevancia del estudio

Este análisis se enmarca en un proyecto que requiere transformar registros hospitalarios en información utilizable para evaluación clínica, gestión sanitaria y modelado predictivo. Relevancia clave:

- Mejora de la calidad de datos: identificar y corregir inconsistencias reducirá sesgos en modelos y errores en métricas operativas (estancia media, tasas de UCI, costes).
- Soporte a decisiones clínicas y administrativas: variables derivadas (comorbilidades, flags UCI, duración real de estancia) permiten segmentar pacientes por riesgo y consumo de recursos.
- Preparación para modelado: la ingeniería de características propuesta facilitará tareas de clasificación/regresión (p. ej. predecir reingreso, estancia larga, coste elevado) y garantizará compatibilidad entre equipos al entregar scripts reproducibles.
- Cumplimiento y privacidad: dado que el dataset contiene identificadores directos, el AED debe incluir pasos obligatorios para anonimización y tratamientos de datos sensibles antes de cualquier compartición o difusión.
  Impacto esperado: con un AED y un conjunto de variables derivadas validadas, el equipo podrá construir modelos más robustos, generar dashboards operativos y elaborar informes sanitarios confiables que apoyen intervenciones de gestión y mejora de calidad

## 2. Análisis Descriptivo Inicial

### 2.1. Estructura y dimensiones del conjunto de datos

Antes de proceder con los análisis, es fundamental conocer la estructura y dimensiones del conjunto de datos. Esto implica identificar el número de filas (registros) y columnas (variables), así como comprender la naturaleza de cada variable. Esta visión general permite planificar adecuadamente los procesos de limpieza, transformación y análisis, asegurando que el tratamiento de los datos sea el más adecuado según su tipología y volumen.

### 2.2. Tipología de variables

- Variables categóricas y lógicas

```
-- categorico --
Comunidad Autónoma
Categoría
Procedimiento 6
Procedimiento 7
Procedimiento 8
Procedimiento 9
Procedimiento 10
Procedimiento 11
Servicio
Centro Recodificado
Régimen Financiación
Procedencia
Continuidad Asistencia
Ingreso en UCI
Días UCI
Diagnóstico 17
Diagnóstico 18
Diagnóstico 19
Diagnóstico 20
POA Diagnóstico Princi
POA Diagnóstico 2
POA Diagnóstico 3
POA Diagnóstico 4
POA Diagnóstico 5
POA Diagnóstico 6
POA Diagnóstico 7
POA Diagnóstico 8
POA Diagnóstico 9
POA Diagnóstico 10
POA Diagnóstico 11
POA Diagnóstico 12
POA Diagnóstico 13
POA Diagnóstico 14
POA Diagnóstico 15
POA Diagnóstico 16
POA Diagnóstico 17
POA Diagnóstico 18
POA Diagnóstico 19
POA Diagnóstico 20
Tipo GRD APR
Mes de Ingreso
```

```
-- logico --
CCAA Residencia
Procedimiento 12
Procedimiento 13
Procedimiento 14
Procedimiento 15
Procedimiento 16
Procedimiento 17
Procedimiento 18
Procedimiento 19
Procedimiento 20
GDR AP
CDM AP
Tipo GDR AP
Valor Peso Español
Tipo GDR APR
Valor Peso Americano APR
Reingreso
GDR IR
Tipo GDR IR
Tipo PROCESO IR
Procedimiento Externo 1
Procedimiento Externo 2
Procedimiento Externo 3
Procedimiento Externo 4
Procedimiento Externo 5
Procedimiento Externo 6
```

• Variables numéricas

```
-- numerico --
Sexo
Circunstancia de Contacto
Tipo Alta
Estancia Días
GRD APR
CDM APR
Nivel Severidad APR
Riesgo Mortalidad APR
Edad
Coste APR
CIE
Peso Español APR
Edad en Ingreso
```

• Variables de fecha

```
-- fecha --
Fecha de nacimiento
Fecha de Ingreso
```

• Variables de texto libre

```
-- texto libre --
Nombre
Fecha de Fin Contacto
Diagnóstico Principal
Diagnóstico 2
Diagnóstico 3
Diagnóstico 4
Diagnóstico 5
Diagnóstico 6
Diagnóstico 7
Diagnóstico 8
Diagnóstico 9
Diagnóstico 10
Diagnóstico 11
Diagnóstico 12
Diagnóstico 13
Diagnóstico 14
Fecha de Intervención
Procedimiento 1
Procedimiento 2
Procedimiento 3
Procedimiento 4
Procedimiento 5
Número de registro anual
CIP SNS Recodificado
País Nacimiento
País Residencia
Fecha de Inicio contacto
Diagnóstico 15
Diagnóstico 16
```

**2.3. Revisión de valores nulos, vacíos y desconocidos**

**2.4. Análisis de duplicados y consistencia de registro**

**2.5. Identificación de outliers (valores atípicos)**

Estos tres puntos hemos tratado de unificarlos lo máximo posible, hemos utilizado programas con librerías ya existentes para realizar una criba de datos en 2D.

*¿2D? -> FILAS Y COLUMNAS*

Con la limpieza de datos hemos buscado columnas obsoletas, como es el caso, por ejemplo, de:

```
GDR AP > Grupo Relacionado con el Diagnóstico All Patient
 Obsoleto
CDM AP -> Categoría Diagnóstica Mayor All Patient
 Obsoleto
Tipo GDR AP -> Tipo de GRD (médico / quirúrgico) All Patient
 Obsoleto
```

a. **Estos campos se calculaban con software 3M™ All Patient DRG (AP-DRG), usado hasta 2015.**

```
Coste APR -> Coste nuestro no el medio de la tabla, pero lo podemos consultar en la tabla TABLA GRD
GDR IR -> Tipo de GRD (médico / quirúrgico) All Patient
 Obsoleto
Tipo GDR IR-> Tipo de GRD (médico / quirúrgico) All Patient
 Obsoleto
Tipo PROCESO IR -> Tipo de GRD (médico / quirúrgico) All Patient
 Obsoleto
```

b. **Con la implantación del APR-DRG (All Patient Refined DRG), el Ministerio retiró oficialmente los tres campos anteriores.**

Respecto a las filas (**pacientes**) que hemos limpiado, también hemos utilizado librerías externas, como por ejemplo:

- ***readr***: Lectura y escritura eficiente de archivos de datos (CSV, delimitados).
- ***dplyr:*** Manipulación y filtrado de datos mediante gramática de datos.
- ***stringr***: Operaciones y limpieza de cadenas de texto.
- ***janitor***: Limpieza de nombres de columnas y tablas de datos.
- ***naniar*****:** Detección y gestión de valores faltantes (NA).

- ***outliers***: Identificación y tratamiento de valores atípicos (outliers) en variables numéricas.

Gracias a la flexibilidad de R, es posible aplicar criterios estadísticos robustos, para identificar anomalías y asegurar que los datos finales sean consistentes y representativos.

De esta manera, los datos resultantes son más fiables para el análisis, permitiendo obtener conclusiones más **realistas** y **útiles** para la investigación.

**2.6. Estadísticas descriptivas básicas**

- Medidas de tendencia central (media, mediana, moda)
- Medidas de dispersión (desviación estándar, rango, percentiles)
- Distribución de variables clave (edad, estancia, coste, etc.)