Pontificia Universidad
JAVERIANA
Cali

2021

Big Data Processing - Project

JUAN SEBASTIAN REYES

ALEXANDER CASTRO CONTRERA

JHAN CARLOS DEL RIO

23-9-2021

Contenido

Introducción

El conjunto de datos base para la elaboración del proyecto se toma directamente de reportes generados por la Alcaldía de Medellín. Es un conjunto de datos muy completo, contiene 1’725.243 registros de las instituciones médicas de Medellín donde se trataron distintas enfermedades a pacientes en el periodo 2009 – 2017.

Características:	Multivaluado
Número de Registros:	1'725.243
Área:	Atenciones, Morbilidad, Urgencias, Salud, Servicio, prestación
Tipo de Datos:	String, Char, Integer, Float
Numero de Atributos:	18
Idioma:	Español
Autor:	Secretaría de Salud
Nivel de Acceso Público:	Público
Fecha de modificación:	20/04/2028
Fecha de publicación:	19/09/2024
Licencia:	https://opendefinition.org/licenses/cc-by-sa/

Información de los Atributos

- **Identificadas:**
  1. **consecutivo:** Número único (**PK**) de cada registro.
  2. **año:** Año en que se realizó el registro.
  3. **cod_eas:** Identificador EPS (**FK** referencia ‘**nombre_eas**’).
  4. **nombre_eas:** Nombre EPS.
  5. **edad:** Edad del conjunto de personas atendidas asociadas al ‘**consecutivo**’.
  6. **sexo:** Sexo común del conjunto de personas atendidas asociadas al ‘**consecutivo**’.
  7. **cod_departamento:** Identificador Departamento. En este caso todos los registros están asociados al **ID:**5 = ‘**Antioquia**’.
  8. **cod_municipio:** Identificador Municipio. En este caso todos los registros están asociados al **ID:**1 = ‘**Medellín**’.
  9. **zona:** Identifica el tipo de zona del conjunto de personas atendidas asociadas al ‘**consecutivo**’.
  10. **cod_ips:** Identificador Institución (**FK** referencia ‘**nombre_institucion**’).
  11. **nombre_institucion:** Nombre Institución.
  12. **cod_dx_salida:** Identificador Enfermedad (**FK** referencia ‘**nombre_dx**’).
  13. **nombre_dx:** Nombre Enfermedad

14. **servicio:** Identificador del tipo de servicio. En este caso todos los registros están asociados a '**URGENCIAS**'.
15. **total_atenciones:** Representa el numero de pacientes atendidos en la misma institución, por la misma enfermedad, del mismo sexo asociadas al '**consecutivo**'.

- **Desconocidas:**
    1. **tipo_usuario**
    2. **tipo_edad**
    3. **causa_externa.**

Clasificación de los Atributos

Evidenciamos que el conjunto de datos maneja datos con variables cualitativas y cuantitativas. A continuación, mostramos la clasificación:

Cualitativa (Nominal)	Cualitativa (Ordinal)	Cuantitativa (Intervalo)	Cuantitativa (Razón)
cod_eas	consecutivo	edad	total_atenciones
nombre_eas	año		
sexo	tipo_usuario		
zona	tipo_edad		
cod_ips	cod_departamento		
nombre_institucion	cod_municio		
cod_dx_salida	causa_externa		
nombre_dx			
servicio			

Valores Únicos

Como podemos observar se tiene que los atributos como servicio, **cod_municipio** y **cod_departamento** solo presentan 1 elemento dentro de su rango de opciones. Por otro lado, se tiene que el atributo **zona**, presenta un problema en la distinción entre mayúsculas y minúsculas, específicamente con la letra **U/u**. Además, se puede evidenciar que los datos de edad presentan algunos valores anormales.

Pre-Procesamiento de la Data

Este es el proceso inicial más importante, la idea es que al finalizar este proceso las siguientes preguntas claves estén resueltas: **¿Tienen sentido los datos? ¿Los datos siguen las reglas apropiadas para su campo? ¿Los datos tienen problemas de calidad?**
Para ello, seguimos una serie de recomendaciones estrictas que permitieron un mayor desempeño del conjunto de datos a la hora de encaminarnos hacia la variable objetivo.
A continuación, detallamos el proceso:

1. **Eliminamos observaciones duplicadas o irrelevantes:**
   Las columnas de servicio-cod_municipio-cod_departamento.

2. **Analizamos y corregimos errores estructurales:**
   - Columna zona, caso de mayúsculas-minusculas.
   - Columna Edad en algunos registros no cumplía con el formato.

3. **Filtramos valores atípicos y evaluamos la importancia en el conjunto de datos:**
   El hecho de que exista un valor atípico no significa que sea incorrecto. Este paso es necesario para determinar la validez de ese dato.

4. **Etapa de QA:**
   En esta etapa de prueba lo que buscamos fue identificación de patrones de errores para reducir estos mismos en el futuro. Después de esta etapa hicimos una iteración mas por estos mismos 4 pasos.