Documentación del Proceso de Limpieza de Datos

Dataset: Medical Cost Personal

1. Descripción del Dataset Original

El conjunto de datos **Medical Cost Personal** contiene información sobre costos médicos individuales facturados por seguros de salud en Estados Unidos. Las variables incluidas son:

- **age**: Edad del beneficiario
- **sex**: Género del beneficiario (masculino/femenino)
- **bmi**: Índice de masa corporal (IMC)
- **children**: Número de hijos cubiertos por el seguro
- **smoker**: Estado de fumador (sí/no)
- **region**: Región de residencia en EE.UU. (noreste, noroeste, sureste, suroeste)
- **charges**: Costos médicos facturados por el seguro

2. Problemas Identificados en el Dataset Original

Durante la exploración inicial, se identificaron los siguientes problemas:

Valores Faltantes

- Valores nulos en todas las columnas, con mayor incidencia en `region` y `smoker`.

Inconsistencias en Datos Categóricos

- Diferentes formatos para `sex`: 'male', 'female', 'Male', 'Female', 'm', 'f'.
- Diferentes formatos para `smoker`: 'yes', 'no', 'Yes', 'No', 'Y', 'N', '1', '0'.
- Diferentes formatos para `region`: nombres completos y abreviaturas ('ne', 'nw', 'se', 'sw').

Valores Atípicos

- IMC con valores imposibles (negativos, cero o extremadamente altos).
- Edades fuera del rango razonable (negativas o superiores a 100 años).
- Costos médicos negativos.
- Número de hijos con valores negativos.

Registros Duplicados

- Se identificaron 20 registros duplicados.

3. Proceso de Limpieza Aplicado

Eliminación de Duplicados

Se eliminaron todos los registros duplicados, reduciendo el tamaño del dataset.

Manejo de Valores Nulos

- **Variables numéricas**:
  - `age`, `bmi`, `children`: se imputó la mediana.
  - `charges`: se imputó la media.
- **Variables categóricas**:
  - `sex`, `smoker`, `region`: se utilizó la moda.

Estandarización de Categorías

- `sex`: conversión a minúsculas y mapeo de abreviaturas a 'male' y 'female'.
- `smoker`: conversión a minúsculas y mapeo a 'yes' y 'no'.
- `region`: mapeo de abreviaturas a nombres completos.

Validación con Expresiones Regulares

- Validación de `sex` con regex; valores incorrectos fueron reemplazados por la moda.

Manejo de Valores Atípicos

- **Edad**: entre 18 y 100 años.
- **IMC**: entre 15 y 50.
- **Hijos**: entre 0 y 10, convertido a entero.
- **Costos médicos**: los valores negativos se reemplazaron por la mediana.

Transformaciones Adicionales (Opcionales)

- **Normalización**: se aplicó `StandardScaler` a variables numéricas (`age`, `bmi`, `charges`).
- **One-Hot Encoding**: aplicado a variables categóricas (`sex`, `smoker`, `region`).

4. Resultados de la Limpieza

Comparación Cuantitativa

- Registros duplicados eliminados: 20
- Valores nulos tratados: 100 %
- Registros con sexo inválido corregidos: 100 %
- Valores atípicos tratados: 100 %

Mejoras en la Calidad de los Datos

- **Consistencia**: formatos estandarizados
- **Completitud**: no hay valores nulos
- **Validez**: rangos razonables
- **Precisión**: errores corregidos

5. Archivos Generados

- `insurance_clean.csv`: Dataset limpio con variables originales
- `insurance_normalized.csv`: Variables numéricas normalizadas
- `insurance_encoded.csv`: Variables categóricas codificadas

6. Conclusiones y Recomendaciones

El proceso de limpieza ha mejorado significativamente la calidad de los datos, eliminando inconsistencias y preparando el dataset para su uso en análisis estadístico o modelos de *machine learning*. Se recomienda:

- Usar `insurance_clean.csv` para análisis exploratorio
- Usar `insurance_normalized.csv` o `insurance_encoded.csv` para modelos de *machine learning*

7. Consideraciones Adicionales

- La mediana fue preferida sobre la media debido a la presencia de outliers.
- La normalización facilita la aplicación de modelos sensibles a escalas.
- El proceso es completamente reproducible mediante código Python documentado.