UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la Información: Desarrollo y Gestión de Software

Reporte de limpieza de datos

IDGS91N - Kevin Iván Aguirre Silva (1123250020)
Extracción de Conocimiento en Bases de Datos - Ing. Luis Enrique Mascote Cano

Chihuahua, Chih., 04 de octubre de 2025

ÍNDICE

INTRODUCCIÓN

La limpieza de datos constituye una etapa fundamental dentro del proceso de análisis y ciencia de datos, ya que garantiza la calidad, precisión y confiabilidad de la información que se utiliza para la toma de decisiones, el desarrollo de modelos predictivos y la generación de conocimiento. Este proceso permite identificar y corregir errores, inconsistencias, duplicados, valores faltantes y formatos incorrectos que, de no ser atendidos, podrían introducir sesgos o invalidar los resultados obtenidos.

Sin una adecuada limpieza, los análisis pueden verse distorsionados y los modelos de aprendizaje automático producir predicciones erróneas o sesgadas, lo cual impacta negativamente en la calidad de las decisiones y puede acarrear consecuencias significativas para las organizaciones o proyectos que dependen de los datos. Además, la limpieza contribuye a agilizar los procesos operativos y mejorar la productividad, al asegurar que los datos estén consistentes, completos y listos para ser utilizados de forma eficiente. En este sentido, se considera la base sobre la cual se construyen todos los análisis y decisiones confiables en el ámbito de la ciencia y análisis de datos (Daniel, 2023).

En este reporte se aborda el proceso de análisis y limpieza de un conjunto de datos proveniente de la plataforma Kaggle, titulado Meal Plan & Exercise Schedule (Gender, Goal, BMI). Dicho dataset contiene 80,000 registros y 5 variables, y fue diseñado para sistemas de recomendación en el ámbito de la salud y el fitness. Este conjunto de datos mapea las entradas básicas del usuario Gender, Goal y BMI Category con recomendaciones personalizadas de Exercise Schedule (rutina de ejercicio) y Meal Plan (plan alimenticio).

El objetivo del presente documento es analizar la procedencia, los tipos y las fuentes de los datos, así como aplicar y documentar técnicas de limpieza empleando la biblioteca Pandas de Python, a fin de garantizar la integridad del conjunto y dejarlo preparado para su uso en modelos de aprendizaje automático o sistemas de recomendación.

1. PROCEDENCIA DE LOS DATOS

1.1. Plan de Comidas y Horario de Ejercicios (Género, Objetivo, IMC)

Este es un conjunto de datos limpio y estructurado (80,000 filas, 5 columnas) que mapea entradas básicas del usuario como por ejemplo; Género, Objetivo y Categoría de IMC. Incluye un horario de ejercicios y un plan de comidas recomendados. Ideal para sistemas de recomendación, clasificación multi-salida y líneas de base basadas en reglas en aplicaciones de salud y fitness.

Este conjunto de datos proporciona recomendaciones de fitness alineadas basadas en atributos simples del usuario. Dados el Género, el Objetivo (quemar grasa o ganar músculo) y la categoría de IMC, genera un horario de ejercicios y un plan de comidas sugeridos. Está diseñado para predicción multi-salida, prototipos de sistemas de recomendación y proyectos educativos de aprendizaje automático (Meal Plan & Exercise Schedule (Gender ,Goal ,BMI), 2025).

2. TIPOS Y FUENTES DE DATOS

2.1. Estructura del conjunto de datos

- El archivo es un CSV estructurado, con 80,000 registros y 5 columnas.
- Todas las variables son categóricas o textuales, no hay variables numéricas directas.
- Procede de un sistema curado de recomendaciones (conjunto de datos creado y publicado en Kaggle para proyectos de fitness).

2.2. Clasificación de las variables

1. ***Gender (Male, Female)***

- **Tipo:** cualitativa nominal (son categorías sin orden jerárquico).
- **Fuente:** datos generados por humanos (atributo de identificación personal en encuestas o registros).

2. ***Goal (fat_burn, muscle_gain)***

- **Tipo:** cualitativa nominal (dos opciones posibles, sin jerarquía).
- **Fuente:** datos generados por humanos (objetivo declarado por el usuario).

3. ***BMI Category (Underweight, Normal weight, Overweight, Obesity)***

- **Tipo:** cualitativa ordinal (categorías con un orden lógico relacionado con el peso corporal).
- **Fuente:** puede considerarse mixta:
    - Derivada de datos biométricos (índice de masa corporal).
    - Registrada a través de interacción humana o médica.

4. ***Exercise Schedule (texto con rutinas de ejercicio)***

- **Tipo:** cualitativa no estructurada (texto descriptivo que puede variar en redacción).
- **Fuente:** datos generados por humanos y curados para recomendación.

5. ***Meal Plan (texto con dietas recomendadas)***

- **Tipo:** cualitativa no estructurada (texto libre con listas de alimentos).
- **Fuente:** datos generados por humanos y compilados por el creador del dataset.

2.3. Síntesis

- **Naturaleza de los datos:** principalmente cualitativos (nominales, ordinales y textuales).
- **Estructura:** datos estructurados (tabla) pero con columnas que contienen texto no estructurado (ej. rutinas y planes de dieta).
- **Procedencia/fuentes:**
    - Datos generados por humanos (género, objetivos declarados, rutinas, dietas).
    - Datos biométricos derivados (categoría del IMC).

3. TÉCNICAS DE LIMPIEZA DE DATOS

Para el desarrollo de las diferentes técnicas de limpieza de datos se utiliza el lenguaje de programación Python en conjunto de la librería Pandas principalmente debido a que Pandas es una biblioteca potente, flexible y fácil de usar diseñada para manipular y analizar datos de manera eficiente. Pandas ofrece estructuras de datos como DataFrame que funcionan como tablas, facilitando la carga, limpieza, manipulación, segmentación y análisis de datos, incluso en grandes volúmenes (Pandas - Python Data Analysis Library, s. f.).

3.1. Desarrollo

Valores nulos

```
7    print("Valores nulos por columna:")
8    print(df.isnull().sum())
```

```
Valores nulos por columna:
Gender              0
Goal                0
BMI Category        0
Exercise Schedule   0
Meal Plan           0
```

No hay valores nulos en ninguna de las 5 columnas por lo tanto no es necesario imputar ni eliminar registros por este motivo.

Duplicados

```
11   df_clean = df.drop_duplicates()
12   print(f"Registros antes: {len(df)}, después de eliminar duplicados: {len(df_clean)}")
```

```
Registros antes: 80000, después de eliminar duplicados: 16
```

Se encontraron 79,984 registros duplicados de un total de 80,000. Esto significa que el conjunto de datos es altamente redundante por lo tanto es necesario eliminar duplicados, esto se logró con la función **drop_duplicates()**.

Errores de formato

```
15   df_clean["Exercise Schedule"] = df_clean["Exercise Schedule"].str.replace(r",\s*", ", ", regex=True)
16   df_clean["Meal Plan"] = df_clean["Meal Plan"].str.replace(r",\s*", ", ", regex=True)
```

1. Las variables categóricas (Gender, Goal, BMI Category) tienen valores consistentes (Male/Female, fat_burn/muscle_gain, categorías de IMC).
2. No hay variaciones de mayúsculas/minúsculas ni errores de ortografía.

3. En las columnas de texto (Exercise Schedule, Meal Plan), el formato es descriptivo y coherente, aunque se detecta falta de espacios después de comas en algunos casos (ej. "carrot sticks,grilled chicken breast").

Como acción se limpiaron espacios y estandarizó el formato de texto.

Valores atípicos

```
print("Categorías en Gender:", df_clean["Gender"].unique())
print("Categorías en Goal:", df_clean["Goal"].unique())
print("Categorías en BMI Category:", df_clean["BMI Category"].unique())
```

```
Categorías en Gender: ['Female' 'Male']
Categorías en Goal: ['muscle_gain' 'fat_burn']
Categorías en BMI Category: ['Normal weight' 'Underweight' 'Overweight' 'Obesity']
```

Como todas las columnas son categóricas o textuales, no hay valores numéricos extremos. Los únicos posibles “atípicos” serían categorías no reconocidas, pero aquí todo es consistente.

3.2. Resumen

En el análisis de calidad de datos se identificó que el conjunto no presenta valores nulos en ninguna de sus variables, por lo que no fue necesario aplicar técnicas de imputación o eliminación. Sin embargo, se detectó un número considerable de registros duplicados (79,984 de 80,000), lo cual representa un 99.9% de redundancia. Para garantizar la validez del análisis, dichos registros fueron eliminados mediante la función drop_duplicates() de Pandas.

En cuanto a los errores de formato, las variables categóricas se encuentran estandarizadas (ejemplo: Gender contiene únicamente los valores “Male” y “Female”). No obstante, en las variables textuales se identificaron inconsistencias menores, como la ausencia de espacios tras algunas comas en descripciones de comidas. Estos errores fueron corregidos mediante expresiones regulares para asegurar uniformidad en el texto.

Finalmente, no se identificaron valores atípicos numéricos debido a que las variables del dataset son mayormente categóricas o textuales. Las categorías presentes en cada columna corresponden al dominio esperado, por lo que no fue necesario aplicar técnicas de detección y corrección de outliers.

4. CONCLUSIÓN

El proceso de limpieza de datos constituye un pilar esencial dentro del análisis de información, ya que garantiza que los resultados obtenidos sean válidos, consistentes y confiables. A lo largo de este caso de estudio, se evidenció que incluso un conjunto de datos aparentemente limpio como el Meal Plan & Exercise Schedule (Gender, Goal, BMI) de Kaggle puede contener problemas de redundancia y formato que deben ser tratados antes de su uso en análisis o modelos predictivos.

La aplicación de la biblioteca Pandas en Python permitió detectar y corregir eficazmente los principales problemas de calidad presentes en el conjunto de datos: la eliminación de registros duplicados y la estandarización de formato en los textos descriptivos. Estas acciones no solo mejoraron la integridad del conjunto, sino que también aseguraron que las variables categóricas y textuales mantuvieran una estructura coherente y apta para posteriores procesos de modelado o recomendación.

Asimismo, el análisis permitió reforzar el entendimiento de los tipos y fuentes de datos, distinguiendo entre variables cualitativas nominales, ordinales y textuales, así como entre datos generados por humanos y biométricos derivados. Este ejercicio demuestra que la limpieza no se limita a la corrección técnica, sino que implica una comprensión profunda del origen, naturaleza y propósito de los datos.

REFERENCIAS

Meal plan & Exercise Schedule (Gender ,Goal ,BMI). (14 de agosto de 2025). Kaggle. https://www.kaggle.com/datasets/kavindavimukthi/meal-plan-and-exercise-schedule-gender-goal-bmi?resource=download

pandas - Python Data Analysis Library. (s. f.). https://pandas.pydata.org/

Daniel. (30 de octubre de 2023). *Datacleaning Limpieza de datos: definición, técnicas, importancia en Data Science*. DataScientest. https://datascientest.com/es/datacleaning-limpieza-de-datos-definicion-tecnicas-importancia-en-data-science