

Informe sobre la Selección de las Tres Bases de Datos

Sergio Mínguez Cruces

18 de noviembre de 2025

Índice

1. Introducción	3
2. Dataset 1: FitLife - Health and Fitness Tracking Dataset	3
2.1. Descripción de la Base	3
2.2. Fuente de Descarga	4
2.3. Selección y Justificación de la Variable Objetivo	4
3. Dataset 2: Air Quality Monitoring in European Cities	4
3.1. Descripción de la Base	4
3.2. Fuente de Descarga	5
3.3. Selección y Justificación de la Variable Objetivo	5
4. Dataset 3: 90 Day Habit Tracker for Personal Growth	6
4.1. Descripción de la Base	6
4.2. Fuente de Descarga	6
4.3. Selección y Justificación de la Variable Objetivo	6
5. Conclusión	6

1. Introducción

Este informe analiza tres bases de datos relacionadas con salud y medio, con el propósito de explicar su estructura, origen y relevancia para análisis de datos. Los *datasets* provienen de Kaggle, un repositorio abierto. Para cada una, se describirá su contenido, se proporcionará la URL de descarga y se justificará la selección de la variable objetivo basada en los siguientes criterios: relevancia y aplicabilidad práctica.

2. Dataset 1: FitLife - Health and Fitness Tracking Dataset

2.1. Descripción de la Base

FitLife - Health and Fitness Tracking Dataset es una base de datos sintética diseñada en la que se recopilan métricas de *salud* y *fitness*, incluyendo actividad física, hábitos nutricionales y niveles de estrés. Su propósito es realizar análisis predictivos para el bienestar personal. Contiene aproximadamente 690,000 registros totales, con 22 columnas que cubren distintos aspectos de la salud.

Cuadro 1: Variables principales del Health and Fitness Dataset

Variable	Tipo	Descripción
date	Categorico	Fecha (YYYY-MM-DD)
participant_id	Numérico	Identificador único de cada participante
age	Numérico	Edad del participante
gender	Categorico	Género (M/F/Otro)
height_cm	Numérico	Altura en centímetros
weight_kg	Numérico	Peso en kilogramos
bmi	Numérico	Índice de Masa Corporal (calculado)
activity_type	Categorico	Tipo de ejercicio (Correr, Nadar, Ciclismo, etc.)
duration_minutes	Numérico	Duración de la sesión de actividad (minutos)
intensity	Categorico	Intensidad del ejercicio (Baja/Media/Alta)
calories_burned	Numérico	Calorías quemadas estimadas durante la actividad
daily_steps	Numérico	Conteo de pasos diarios
avg_heart_rate	Numérico	Ritmo cardíaco promedio durante la actividad
resting_heart_rate	Numérico	Ritmo cardíaco en reposo
blood_pressure_systolic	Numérico	Presión arterial sistólica
blood_pressure_diastolic	Numérico	Presión arterial diastólica
health_condition	Categorico	Presencia de condiciones de salud

Variable	Tipo	Descripción
smoking_status	Categorico	Historial de tabaquismo (Nunca/Exfumador/Actual)
hours_sleep	Numérico	Horas de sueño por noche
stress_level	Numérico	Nivel de estrés diario (1–10)
hydration_level	Numérico	Consumo diario de agua en litros
fitness_level	Numérico	Puntuación de condición física basada en actividad acumulada

2.2. Fuente de Descarga

- **URL:** <https://www.kaggle.com/datasets/jijagallery/fitlife-health-and-fitness-tracking-dataset>

2.3. Selección y Justificación de la Variable Objetivo

- **Variable Objetivo:** *Stress_Level* (Numérico rango (1–10)).
- **Relevancia:** Como multiples estudios han confirmado el nivel de estrés impacta significativamente en la salud del paciente.
- **Aplicabilidad:** Desarrollo de aplicaciones de bienestar, como rutinas de reducción del nivel de estrés o implementaciones en planes de estudio para mejorar la salud a los estudiantes.

3. Dataset 2: Air Quality Monitoring in European Cities

3.1. Descripción de la Base

Air Quality Monitoring in European Cities recopila mediciones de calidad del aire de estaciones en varias ciudades europeas, pero solo se estudiará *ancona_data* que recoplia en la ciudad de Ancona, Italia, cubriendo contaminantes clave y condiciones meteorológicas. Su propósito es apoyar análisis ambientales y modelado de impactos en la salud pública. Incluye datos horarios o diarios de 2021–2023, con alrededor de 420.000 observaciones y 19 columnas.

Cuadro 2: Variables principales del Air Quality Monitoring Dataset

Variable	Tipo	Descripción
Date	Categorico	Fecha y hora de la observación (YYYY-MM-DD HH:MM:SS)
NO2	Numérico	Concentración de dióxido de nitrógeno ($\mu\text{g}/\text{m}^3$)
O3	Numérico	Concentración de ozono, con efectos aditivos o sinérgicos ($\mu\text{g}/\text{m}^3$)

Variable	Tipo	Descripción
PM10	Numérico	Concentración de aerosoles con diámetro $\leq 10 \mu\text{m}$ ($\mu\text{g}/\text{m}^3$)
PM2.5	Numérico	Concentración de aerosoles con diámetro $\leq 2.5 \mu\text{m}$ ($\mu\text{g}/\text{m}^3$)
Latitude	Numérico	Coordenada geográfica (grados)
Longitude	Numérico	Coordenada geográfica (grados)
station_name	Categorico	Nombre de la estación de monitoreo
Wind-Speed (U)	Numérico	Componente de viento en dirección longitudinal (m/s)
Wind-Speed (V)	Numérico	Componente de viento en dirección latitudinal (m/s)
Dewpoint Temp	Numérico	Temperatura de punto de rocío ($^{\circ}\text{C}$)
Temp	Numérico	Temperatura del aire ($^{\circ}\text{C}$)
Vegetation (High)	Numérico	Cobertura vegetal de alto nivel
Vegetation (Low)	Numérico	Cobertura vegetal de bajo nivel
Soil Temp	Numérico	Temperatura promedio del suelo ($^{\circ}\text{C}$)
Total Percipitation	Numérico	Flujo de agua equivalente (lluvia o nieve) (mm)
Relative Humidity	Numérico	Humedad relativa del aire (%)
code	Categorico	Código de la estación de monitoreo
id	Categorico	ID de la estación de monitoreo

3.2. Fuente de Descarga

- **URL:** <https://www.kaggle.com/datasets/yekenot/air-quality-monitoring-in-european-cities>

3.3. Selección y Justificación de la Variable Objetivo

- **Variable Objetivo:** *PM2.5* (numérico, escala 0–200).
- **Relevancia:** Las partículas PM2.5 son de los contaminantes más crítico para la salud (causan problemas respiratorios y cardiovasculares, según la OMS). Sirve como indicador para calidad general del aire y es el foco principal de regulaciones europeas.
- **Aplicabilidad:** Alertas urbanas en tiempo real integrar predicciones de PM2.5 para recomendar rutas con menor exposición a contaminación a la población.

4. Dataset 3: 90 Day Habit Tracker for Personal Growth

4.1. Descripción de la Base

El *90 Day Habit Tracker for Personal Growth* es un *dataset* que recopila hábitos diarios durante 90 días para fomentar el desarrollo personal, incluyendo métricas de productividad, ejercicio y estado emocional. Su propósito es analizar patrones de hábitos para mejorar en coaching y auto-mejora. Contiene 90 registros, con 8 columnas.

Cuadro 3: Variables principales del Habit Tracker Dataset

Variable	Tipo	Descripción
Date	Númérico	Día del desafío (1–90)
Workout_Duration_Min	Númérico	Minutos de ejercicio
Reading_Min	Númérico	Minutos dedicados a la lectura
Sleep_Hours	Númérico	Horas de sueño
Journaling	Catégorico	Escribir un diario (sí/no)
Screen_Time_Hours	Númérico	Horas enfrente de una pantalla
Daily_Expense (RM)	Númérico	Gasto diario (Malaysian Ringgit)
Mood_Score	Númérico	Puntuación del estado de ánimo autoevaluada (1–10)

4.2. Fuente de Descarga

- **URL:** <https://www.kaggle.com/datasets/uthaya1995/90-day-habit-tracker-for-personal-growth>

4.3. Selección y Justificación de la Variable Objetivo

- **Variable Objetivo:** *Mood_Score* (numérica, escala 1–10).
- **Relevancia:** *Mood_Score* permite medir cómo los hábitos diarios influye en el bienestar emocional.
- **Aplicabilidad:** Permite personalizar programas de crecimiento personal, además de correlacionar aspectos del día a día como ejercicio, horas de sueño, tiempo de lectura y uso de pantallas con el estado de ánimo.

5. Conclusión

Los tres *datasets* (*FitLife*, *Air Quality*, *Habit Tracker*) se han seleccionado por su enfoque en salud y medio ambiente, sumado a que hay variación en el tiempo de recogida de datos: diaria (*FitLife* y *Habit Tracker*) y horaria (*Air Quality*). Las variables objetivo (*Stress_Level*, *PM2.5*, *Mood_Score*) se seleccionaron por su relevancia y posibles aplicaciones ya explicadas. Para ello se tendrán que realizar modelos predictivo o clasificatorios ya que pese a que las tres variables son numéricas, *Stress_Level* y *Mood_Score* son discretas.