Proyecto Integrado (EA1). Formulación de una necesidad de ingeniería de datos

Integrante:

Juan Diego Urrego Gutiérrez

Institución Universitaria Digital de Antioquia

Proyecto Integrado V - Linea de Énfasis - PREICA2502B020074

Docente: Andrés Felipe Callejas Jaramillo

9 de noviembre del 2025

Contenido

Introducción

El presente documento desarrolla la formulación de una necesidad de ingeniería de datos enfocada en el análisis de estilos de vida y su relación con riesgos de salud en la población. El propósito principal es diseñar un flujo técnico que permita a una IPS pública de Medellín identificar patrones entre los hábitos de las personas y la aparición de enfermedades crónicas no transmisibles, como la hipertensión y la diabetes.

A partir de un conjunto de datos público disponible en Kaggle, se busca transformar la información en conocimiento útil mediante procesos de extracción, modelado, validación y documentación reproducible, fortaleciendo así las capacidades de análisis preventivo en el sector salud.

Resumen (Abstract)

Este documento formula una necesidad de ingeniería de datos basada en un dataset público de Kaggle sobre estilos de vida y riesgos de salud. El caso de uso busca apoyar a una IPS pública en la priorización de acciones preventivas mediante la identificación de patrones entre hábitos (ejercicio, sueño, consumo de azúcar, tabaco y alcohol) y condiciones como hipertensión o diabetes. En la Etapa 1 se delimita el problema, se describen las variables relevantes (edad, peso, altura, IMC, horas de sueño, frecuencia de ejercicio, consumo), y se planifica una metodología ágil con WBS y diagrama de Gantt. Se materializa una base local en SQLite evidenciando el flujo dataset → SQLite → CSV para asegurar trazabilidad y reproducibilidad. Esta entrega sienta las bases para análisis descriptivos y modelado predictivo en etapas posteriores, garantizando gobernanza mínima de datos (licencias, uso académico), control de calidad básico (tipos y valores faltantes) y documentación en README y borrador APA. Los resultados analíticos se desarrollarán en la Etapa 3, mientras que el detalle metodológico se profundizará en la Etapa 2.

Palabras clave: estilos de vida; riesgo de salud; hipertensión; diabetes; IMC; sueño; ejercicio; Kaggle; SQLite; pipeline de datos.

Descripción

En la ciudad de Medellín, las entidades prestadoras de salud enfrentan el desafío de anticiparse a la aparición de enfermedades crónicas que afectan la calidad de vida de la población. Este proyecto propone aprovechar un dataset público sobre estilos de vida y riesgos de salud para construir una base de datos estructurada que sirva como punto de partida para el desarrollo de indicadores y modelos predictivos futuros.

La iniciativa busca sentar las bases de un flujo de ingeniería de datos que respalde procesos de toma de decisiones en salud preventiva, integrando tecnología, análisis y gestión del conocimiento con fines sociales.

Objetivo general

Diseñar e implementar un flujo básico de ingeniería de datos basado en un dataset público de estilos de vida, que permita preparar la información para estimar riesgos de salud en población adulta y aplicarlo en la ciudad de Medellín.

Objetivos específicos

-Ingestar y preparar los datos: Descargar el conjunto de datos desde Kaggle y trasladarlo a un entorno virtual alojado en GitHub, registrando la licencia, la fuente y la ruta de acceso correspondiente. Posteriormente, cargar la información en una base de datos SQLite y generar un archivo CSV de verificación para garantizar la correcta ingestión.

-Diseñar la arquitectura del pipeline: Construir una arquitectura mínima que describa el flujo Dataset → SQLite → CSV, implementando un esquema de control de versiones en el repositorio de GitHub que permita mantener la trazabilidad de cada etapa del proceso.

-Explorar la estructura y calidad de los datos: Realizar un análisis exploratorio inicial para comprender la composición del dataset, identificar valores faltantes, inconsistencias y evaluar la calidad de la información mediante técnicas básicas de

perfilado de datos.

-Documentar el proceso técnico y académico: Elaborar un README descriptivo, incluir evidencias de ejecución en GitHub (scripts, registros y resultados), y desarrollar el documento formal del proyecto siguiendo las Normas APA (7.ª edición).

-Planificar la metodología de desarrollo: Definir la estructura de trabajo (WBS) e implementar una metodología Scrum, organizando las tareas por etapas en un diagrama de Gantt que muestre las dependencias, responsables y entregables.

-Construir y validar la base de datos: Crear una base de datos en SQLite, alojada en el entorno del repositorio, realizando la carga completa del dataset y la exportación de un CSV de verificación que evidencie la correcta ejecución del pipeline.

Metodología

El proyecto se desarrolla bajo un enfoque ágil, empleando la metodología Scrum para organizar las actividades en etapas iterativas con entregables parciales.

En la Etapa 1, se realiza la selección y preparación del dataset, la definición del caso de uso, el modelado preliminar de la información y la creación de la base de datos en SQLite.

La planeación se apoya en un diagrama de Gantt, que describe la estructura del trabajo (WBS), fechas, responsables y entregables.

Herramientas utilizadas:

- Python 3.9.12: Ingestión y procesamiento de datos.
- SQLite: Almacenamiento estructurado de la información.
- Pandas: Manipulación y análisis de datos tabulares.

- GitHub: Control de versiones y documentación técnica.
- Excel: Planificación de tareas y elaboración del diagrama de Gantt.

Definición y Preparación

Se desarrolla en las siguientes etapas.

Desarrollo y Análisis de Datos

Se desarrolla en las siguientes etapas.

Resultados

Se desarrolla en las siguientes etapas.

Variables relevantes

- age: Edad del individuo (factor de riesgo base).
- bmi: Índice de masa corporal, proxy nutricional.
- exercise_days_per_week: Frecuencia semanal de actividad física.
- sleep_hours: Horas promedio de sueño por día.
- smoking_status: Consumo de tabaco (nunca, exfumador, actual).
- alcohol_frequency: Frecuencia de consumo de alcohol.
- sugar_intake_g: Ingesta estimada de azúcar en gramos/día.
- hypertension/diabetes: Etiquetas o flags de condición de salud.

Conclusiones

En esta primera etapa logré construir la base técnica y documental del proyecto, garantizando que la información se maneje con calidad y pueda reproducirse de forma confiable. Se definieron el flujo de datos, la metodología ágil de trabajo y las variables más relevantes para los próximos análisis. Este avance representa un paso importante, ya que deja lista una estructura sólida para continuar en la siguiente fase, donde se abordará la parte analítica y la visualización de los resultados.

Referencias

Kaggle. (2025). Lifestyle and Health Risk Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/zahranusrat/lifestyle-and-health-risk-prediction-dataset

Purdue OWL. (n.d.). General Format – APA 7th Edition. Purdue Online Writing Lab. https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/general_format.html

Python Software Foundation. (n.d.). Python 3.9.12 Documentation. https://www.python.org/

Pandas Development Team. (n.d.). pandas: Python Data Analysis Library. https://pandas.pydata.org/

GitHub · Change is constant. GitHub keeps you ahead. (s/f).