

Proyecto Integrado (EA2). Presentación del proyecto y dashboard descriptivo

Integrante:

Juan Diego Urrego Gutiérrez

Proyecto Integrado V - Línea de Énfasis - PREICA2502B020074

Docente: Andrés Felipe Callejas Jaramillo

Institución Universitaria Digital de Antioquia

Medellín - Colombia

7 de diciembre del 2025

Contenido

Introducción	3
Resumen (Abstract).....	4
Descripción	5
Objetivo general	5
Objetivos específicos	5
Variables relevantes	6
Metodología	6
<i>Preparación e Ingesta de Datos</i>	<i>6</i>
<i>Diseño de la Arquitectura del Pipeline</i>	<i>6</i>
<i>Limpieza de Datos</i>	<i>7</i>
<i>Enriquecimiento del Dataset.....</i>	<i>7</i>
<i>Análisis Descriptivo y Exploratorio (EDA)</i>	<i>8</i>
<i>Construcción del Dashboard Descriptivo en Power BI.....</i>	<i>8</i>
Desarrollo y Análisis de Datos	9
Planeación (Diagrama de Gantt)	16
Presentación de Resultados	16
Repositorio	18
https://github.com/JuanUrrego/EA_proyecto_integrado_V_20251-2	18
Conclusiones	19
Referencias	20

Introducción

El presente proyecto desarrolla la formulación y el análisis de una necesidad de ingeniería de datos enfocada en comprender cómo los estilos de vida influyen en los riesgos de salud dentro de una población adulta. Esta iniciativa se enmarca en el contexto de la ciudad de Medellín, donde las instituciones de salud enfrentan el reto de anticiparse a la aparición de enfermedades crónicas no transmisibles como la hipertensión y la diabetes, patologías fuertemente relacionadas con hábitos de alimentación, actividad física y descanso.

Para este propósito se seleccionó un conjunto de datos público disponible en la plataforma Kaggle, el cual contiene información sobre variables asociadas a estilos de vida, tales como actividad física, horas de sueño, ingesta de azúcar, IMC y edad. A este dataset se le realizaron procesos de enriquecimiento temporal, generando fechas comprendidas entre los años 2022 y 2024, con el fin de habilitar análisis comparativos y gráficos de tendencias.

Para analizar los datos se procedió a depurar, enriquecer y analizar el dataset mediante técnicas de limpieza, normalización y visualización, con el objetivo de transformar los datos en información valiosa que facilite la identificación de patrones y potenciales factores de riesgo para la salud. Los resultados obtenidos nos confirman que la estructura del dataset y las relaciones identificadas permiten avanzar hacia técnicas predictivas que puedan estimar probabilidad de desarrollar condiciones relacionadas con estilo de vida (obesidad, riesgo metabólico, mala calidad del sueño, inactividad física).

Resumen (Abstract)

Este documento presenta el análisis de un conjunto de datos sobre estilos de vida y riesgos de salud, con el objetivo de generar información útil que apoye a una IPS pública de Medellín en la toma de decisiones preventivas. A partir de un dataset de Kaggle, se desarrolló un proceso integral de limpieza, normalización, enriquecimiento temporal y análisis descriptivo, centrado en cinco variables clave: edad, IMC, actividad física, horas de sueño e ingesta de azúcar.

Para este proyecto, se eliminaron duplicados y valores inconsistentes, se estandarizaron los tipos de datos, se generaron nuevas columnas temporales (año, mes, día) y se construyeron variables categóricas que permiten una interpretación más clara del comportamiento poblacional. Posteriormente, se aplicaron técnicas de análisis exploratorio (EDA), incluyendo estadísticas descriptivas, histogramas, gráficos de barras, mapas de correlación y tendencias temporales.

Los resultados obtenidos evidencian patrones relacionados con sobrepeso, niveles bajos de ejercicio, hábitos de sueño moderados y consumo de azúcar medio-bajo, lo cual es relevante para evaluar riesgos metabólicos y de enfermedades crónicas. Este análisis sienta las bases para estudios posteriores enfocados en modelamiento predictivo y generación de indicadores que contribuyan a fortalecer la salud pública en la ciudad haciendo uso de la visualización mediante un dashboard que resumirá gráficamente el comportamiento de las cinco variables clave.

Palabras clave: estilos de vida, riesgo de salud, IMC, ejercicio, sueño, Kaggle, ingeniería de datos, Medellín, EDA.

Descripción

En Medellín, las entidades prestadoras de salud requieren herramientas que permitan analizar de forma anticipada los factores asociados al desarrollo de enfermedades crónicas. Este proyecto busca aportar a ese propósito mediante la preparación, depuración, análisis y visualización de un dataset de estilos de vida, con el fin de identificar patrones relevantes que contribuyan al entendimiento de los riesgos de salud en población adulta.

El proyecto integra elementos de ingeniería de datos y análisis exploratorio para organizar la información, mejorar su calidad y generar visualizaciones interpretativas que permitan comprender la dinámica de variables clave como el IMC, la actividad física, el sueño y la alimentación.

Objetivo general

Implementar un proceso de transformación y análisis descriptivo de datos mediante un dashboard interactivo basado en un dataset público de estilos de vida, con el fin de comprender patrones asociados a riesgos de salud en población adulta de Medellín.

Objetivos específicos

- Preparar y depurar los datos descargados desde Kaggle, registrando su origen, estandarizando nombres y tipos de datos y corrigiendo duplicados o inconsistencias para asegurar su calidad.
- Enriquecer la información mediante la creación de variables temporales (año, mes, día) y categorías analíticas como grupos de edad, niveles de IMC, actividad física, sueño e ingesta de azúcar.
- Realizar un análisis exploratorio que incluya estadísticas descriptivas y visualizaciones para identificar patrones, distribuciones y relaciones entre las variables clave.
- Construir un dashboard descriptivo en Power BI que permita sintetizar visualmente el comportamiento de las variables clave, integrar filtros interactivos

y facilitar la interpretación de los patrones identificados.

- Documentar todo el proceso en el repositorio GitHub, garantizando trazabilidad de los cambios y estructurando el informe académico según Normas APA.

Variables relevantes

- age: Edad del individuo (factor de riesgo base).
- bmi: Índice de masa corporal, proxy nutricional.
- sleep: Horas de sueño por día.
- sugar_intake: Ingesta estimada de azúcar.
- exercise: Nivel de ejercicio

Metodología

El proyecto se desarrolla bajo un enfoque ágil, empleando la metodología Scrum. La metodología aplicada incluye:

Preparación e Ingesta de Datos

Se realizó la descarga del dataset desde la plataforma Kaggle, asegurando el registro adecuado de la licencia, la fuente original y la ruta de acceso utilizada. El conjunto de datos fue trasladado a un entorno virtual alojado en GitHub para facilitar la reproducibilidad y el control del proyecto. Posteriormente, los datos fueron cargados en una base de datos SQLite como mecanismo centralizado de almacenamiento, y se generó un archivo CSV de verificación con el fin de validar la correcta ingestión y estructuración del contenido.

Diseño de la Arquitectura del Pipeline

Se definió una arquitectura mínima de procesamiento basada en el flujo Dataset → SQLite → CSV, lo que permite asegurar un manejo ordenado y replicable de los datos. Este diseño se integró con un sistema de control de versiones en GitHub, garantizando la trazabilidad completa de los cambios realizados, así como la documentación de cada una de las etapas del proceso, desde la ingestión hasta la generación de artefactos finales.

Limpieza de Datos

Se desarrolló un proceso de depuración orientado a garantizar la calidad y consistencia del dataset antes del análisis. Este incluyó la identificación y eliminación de duplicados, la corrección de valores nulos e inconsistencias, y la normalización de tipos de datos mediante reglas estandarizadas. Asimismo, se homogenizó la estructura del dataset aplicando una nomenclatura uniforme a las columnas. Todas las transformaciones realizadas fueron documentadas dentro del notebook para asegurar trazabilidad metodológica.

```
# 2.1 Duplicados
dup = df.duplicated().sum()
print("Duplicados:", dup)
df = df.drop_duplicates()

# 2.2 Normalizar nombres
df.columns = [c.strip().lower() for c in df.columns]

df.head()
```

4] .. Duplicados: 0

/>

	age	weight	height	exercise	sleep	sugar_intake	smoking	alcohol	married	profession	bmi	health_risk
0	56	67	195	low	6.1	medium	yes	yes	yes	office_worker	17.6	high
1	69	76	170	high	6.9	high	no	no	no	teacher	26.3	high
2	46	106	153	high	6.6	low	yes	no	no	artist	45.3	high
3	32	54	186	medium	8.5	medium	no	no	no	artist	15.6	low
4	60	98	195	high	8.0	low	no	no	yes	teacher	25.8	high

Enriquecimiento del Dataset

Una vez limpio el dataset, se generaron nuevas variables para ampliar su valor analítico. Dado que el conjunto de datos original no incluía fechas, se creó una columna temporal con valores aleatorios en el rango 2022–2024, a partir de la cual se derivaron las variables de año, mes y día. De forma complementaria, se construyeron variables categóricas comparativas —como rangos de edad, categorías de IMC, niveles de sueño, actividad física e ingesta de azúcar— que facilitaron la interpretación de patrones. El dataset enriquecido fue exportado en formato CSV y se dejó constancia del proceso en el notebook.

6. Enriquecimiento temporal (fecha, año, mes, día)

Fechas simuladas entre 2022 y 2024 para análisis temporal.

```
fecha_inicio = pd.to_datetime("2022-01-01")
fecha_fin = pd.to_datetime("2024-12-31")

df_clave["fecha"] = fecha_inicio + (fecha_fin - fecha_inicio) * np.random.rand(len(df_clave))
df_clave["fecha"] = pd.to_datetime(df_clave["fecha"])

df_clave["anio"] = df_clave["fecha"].dt.year
df_clave["mes"] = df_clave["fecha"].dt.month
df_clave["dia"] = df_clave["fecha"].dt.day

df_clave.head()
```

	age	bmi	exercise	sleep	sugar_intake	age_group	bmi_category	sleep_level	exercise_level	sugar_level	fecha	anio	mes	dia
0	56	17.6	1.0	6.1	2.0	45_59	bajo_peso	medio	bajo	medio	2024-01-11 17:12:14.746842752	2024	1	11
1	69	26.3	3.0	6.9	3.0	60_mas	sobrepeso	medio	alto	alto	2023-07-30 05:58:08.930922456	2023	7	30
2	46	45.3	3.0	6.6	1.0	45_59	obesidad_III	medio	alto	bajo	2022-09-01 16:34:03.235812652	2022	9	1
3	32	15.6	2.0	8.5	2.0	30_44	bajo_peso	alto	moderado	medio	2024-06-18 12:47:57.319705232	2024	6	18
4	60	25.8	3.0	8.0	1.0	60_mas	sobrepeso	alto	alto	bajo	2023-12-08 16:46:34.252867064	2023	12	8

Análisis Descriptivo y Exploratorio (EDA)

Con la base de datos lista, se aplicaron técnicas de análisis descriptivo para comprender la distribución, comportamiento y relaciones entre las variables seleccionadas. Se generaron estadísticas básicas (media, desviación estándar, mínimos, máximos y cuartiles) y se visualizaron las distribuciones mediante histogramas y gráficos de barras. Se elaboró una matriz de correlación para identificar relaciones lineales y un análisis temporal apoyado en la columna de fecha simulada. Cada visualización fue acompañada de una interpretación breve que permitió contextualizar los hallazgos.

Construcción del Dashboard Descriptivo en Power BI

Con el dataset enriquecido final, se desarrolló un dashboard interactivo en Power BI con el propósito de sintetizar gráficamente la información más relevante.

El tablero incluyó:

- Visualizaciones para las cinco variables clave (edad, IMC, sueño, ingesta de azúcar y ejercicio).
- Gráficos de columnas y barras comparativas para analizar distribuciones y patrones por categorías.
- Filtros interactivos (slicers) por año, grupo de edad y categoría de IMC permitiendo explorar distintos segmentos.
- Descripciones y títulos que facilitan la interpretación de los hallazgos.

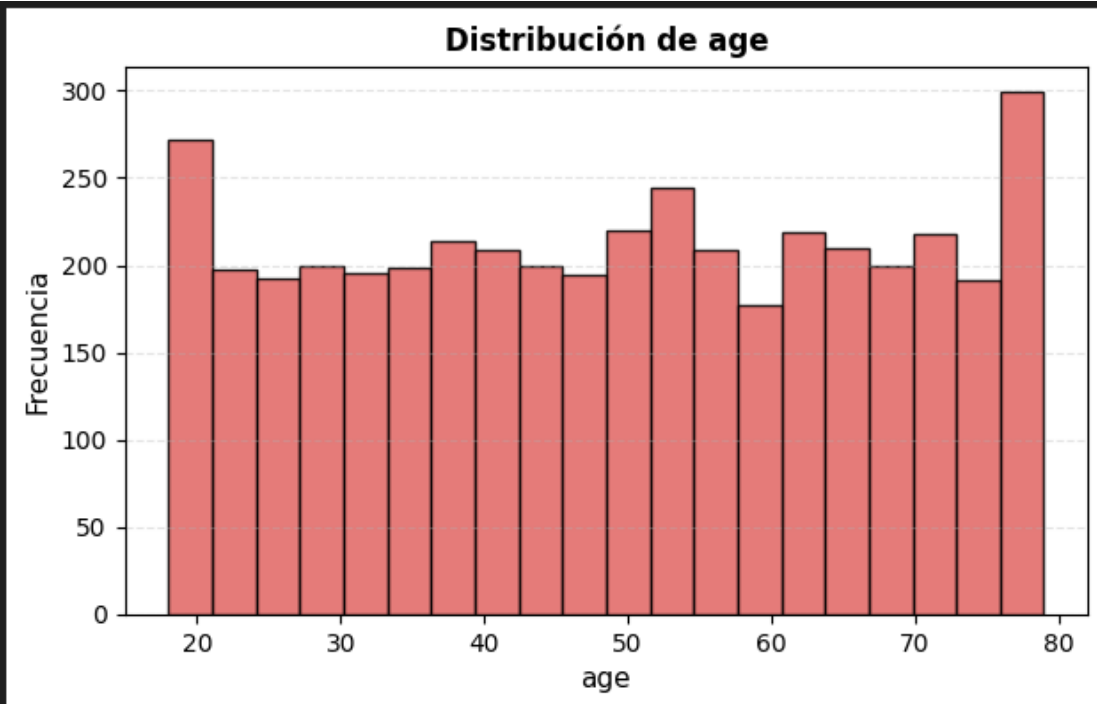
La construcción del dashboard complementó el análisis exploratorio al permitir una lectura visual integrada del comportamiento de la población.

Desarrollo y Análisis de Datos

El análisis descriptivo y exploratorio realizado sobre el dataset enriquecido brindó una comprensión clara del comportamiento de las cinco variables clave del estudio: edad, índice de masa corporal, ejercicio, sueño e ingesta de azúcar. A partir de 4.257 registros, se identificó que la población es mayoritariamente adulta, con una edad media cercana a los 49 años y un BMI promedio de 26.8, indicador de sobrepeso. Las tendencias muestran niveles bajos y moderados de actividad física, patrones de sueño relativamente estables alrededor de 7 horas y una ingesta de azúcar mayoritariamente baja o media.

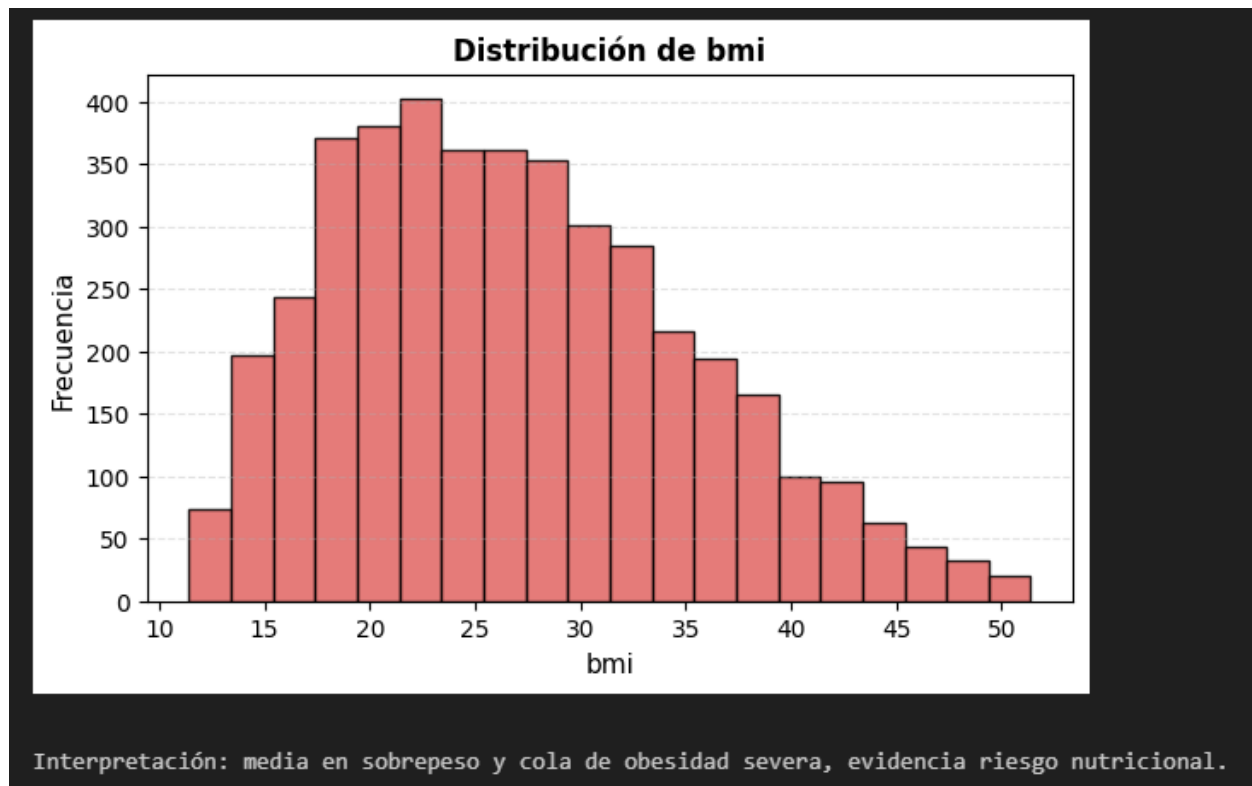
Las visualizaciones —como histogramas, gráficos de barras y una matriz de correlación— permitieron evidenciar concentraciones específicas, asimetrías en la distribución de BMI y ejercicio, así como correlaciones lineales débiles entre las variables, lo que refleja una independencia relativa entre los hábitos analizados. Además, el análisis temporal con fechas generadas mostró una estabilidad general en los promedios mensuales sin estacionalidades relevantes.

En conjunto, estos hallazgos aportan una visión integral del estado y hábitos de la población estudiada, destacan posibles riesgos asociados y confirman la calidad del dataset para avanzar hacia etapas posteriores de análisis e interpretación en el marco del proyecto.



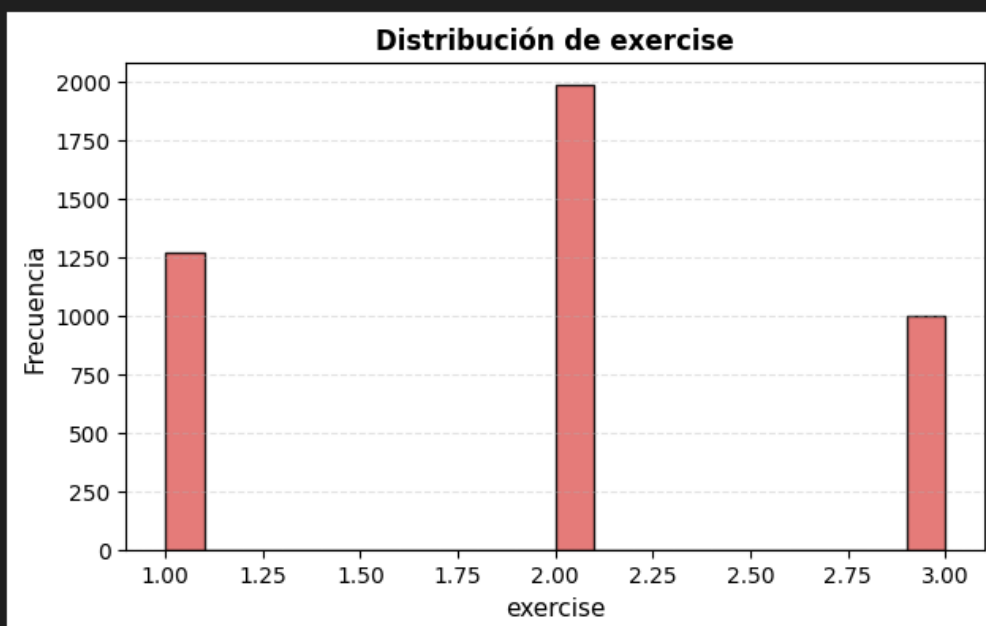
Interpretación: predominan edades adultas (30-65), útil para análisis cardiometabólicos.

La distribución de **age** muestra una población predominantemente adulta, con concentración entre los 30 y 65 años. Hay pocos valores en los extremos (jóvenes <25 y mayores >70), lo que indica que el dataset se enfoca en población adulta, ideal para estudios de salud y estilo de vida.



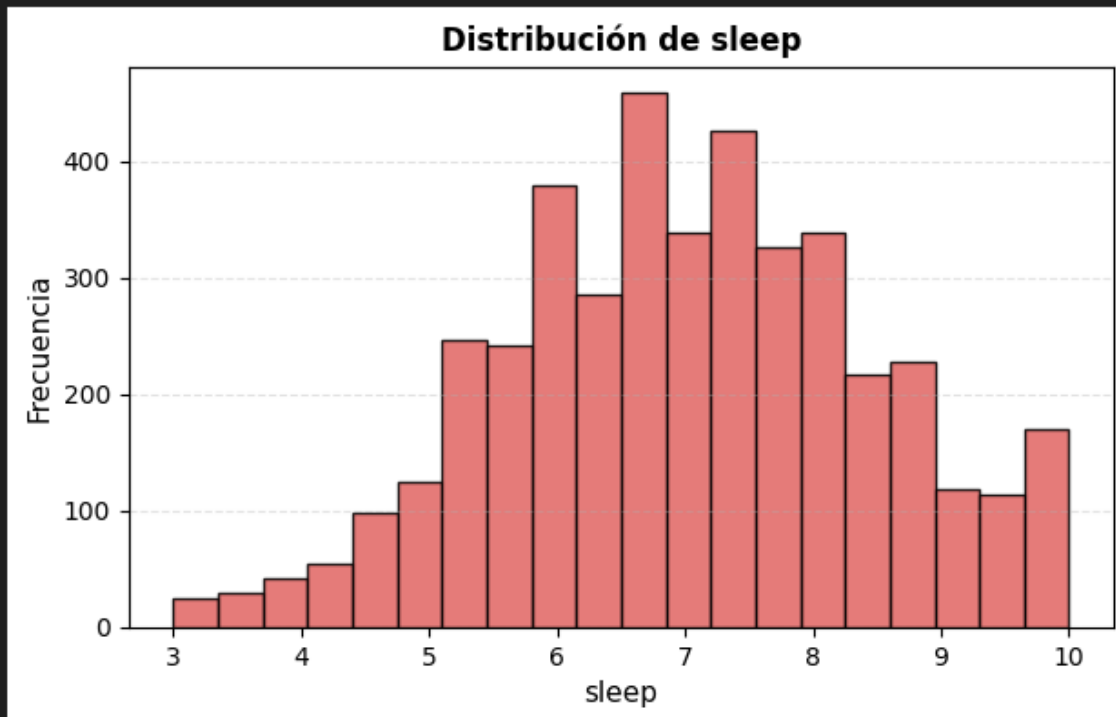
El IMC o BMI se distribuye de forma amplia, con un promedio alrededor de 26–27, lo cual corresponde a **sobrepeso** según clasificación OMS. Se observan valores extremos que representan casos de obesidad severa ($\text{IMC} > 40$), lo que aporta variabilidad útil para análisis de riesgo.

Interpretación: media en sobrepeso y cola de obesidad severa, evidencia riesgo nutricional.



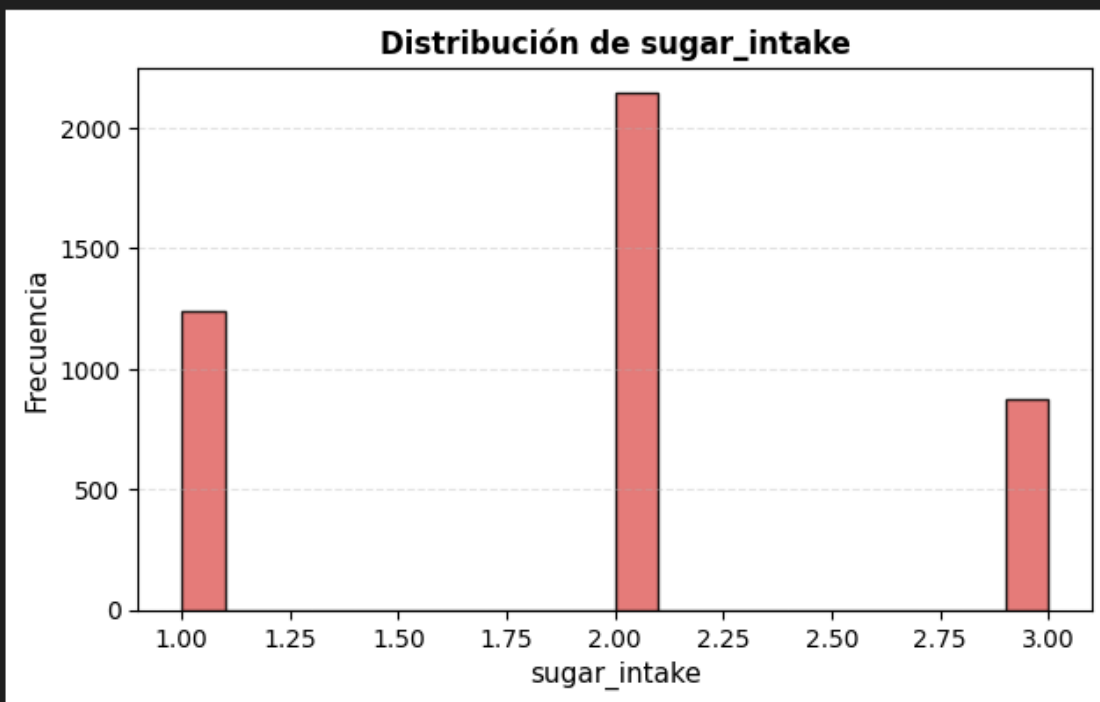
Interpretación: concentración en niveles bajos/medios, sugiere tendencia al sedentarismo.

La mayor concentración está en niveles 1 y 2 (actividad baja y moderada). Los niveles altos (3) son minoritarios. Esto sugiere que la población tiene una **tendencia al sedentarismo**, lo cual puede relacionarse con IMC elevados o hábitos no saludables.



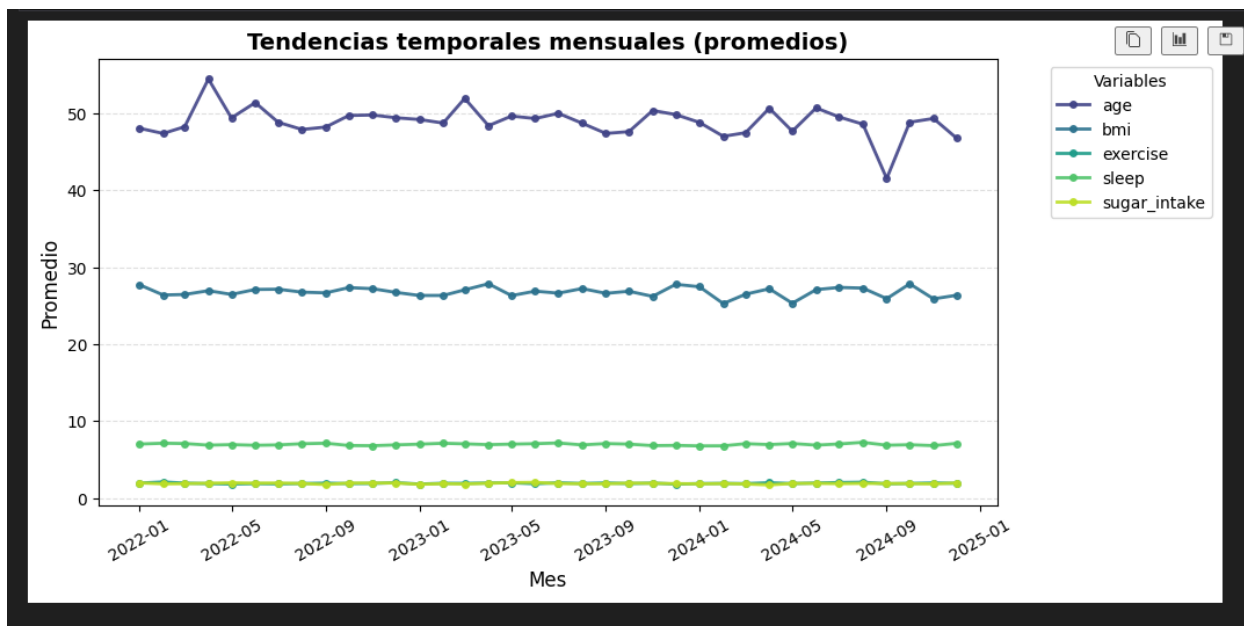
Interpretación: centrado en 6-8h, con grupo minoritario en déficit (<6h).

La distribución del **sueño** está centrada entre 6 y 8 horas, indicando hábitos de descanso relativamente adecuados. Sin embargo, existe una cola en valores bajos (3–5 horas) que representa posibles casos de privación de sueño.



Interpretación: mayoría baja-media, pocos casos altos relevantes para riesgo metabólico.

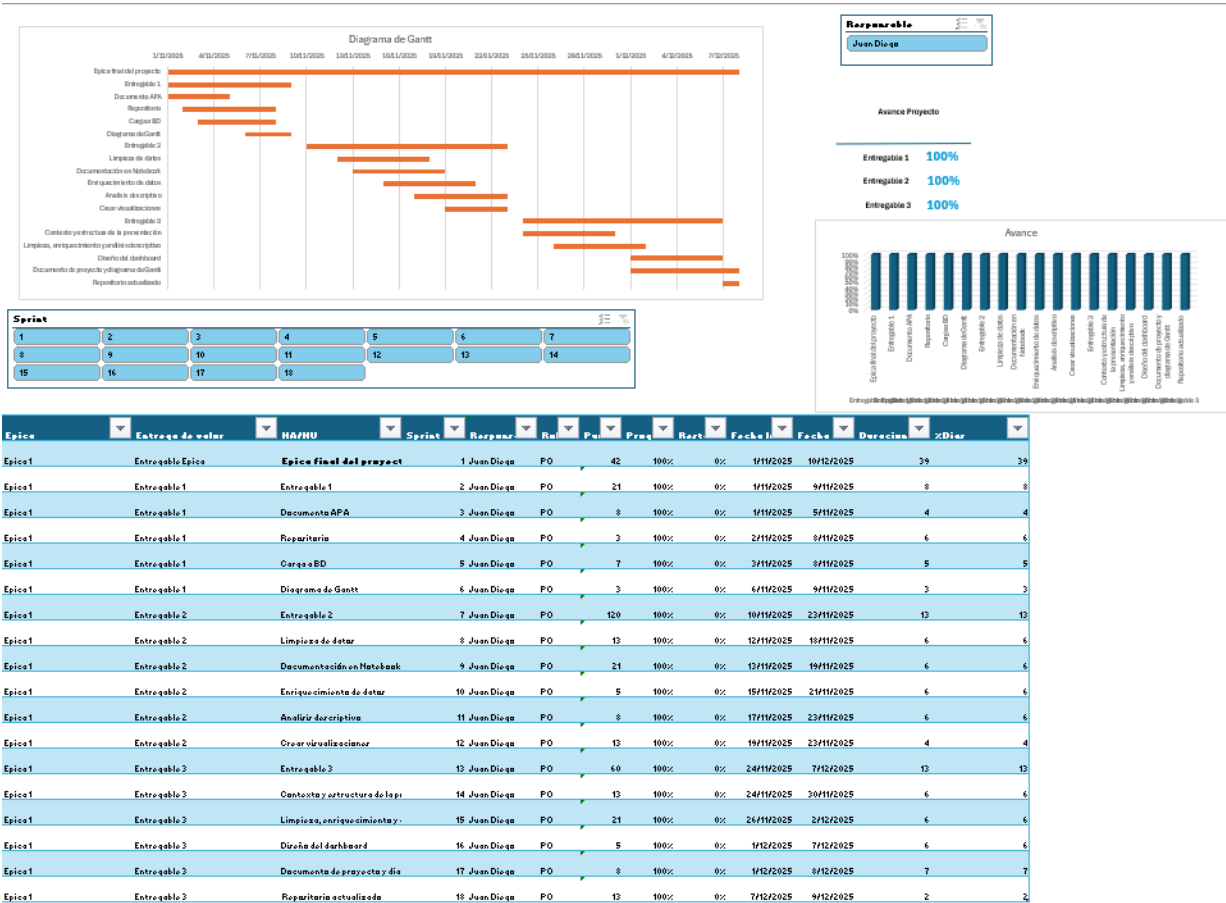
La ingesta de azúcar se concentra en niveles 1 y 2 (baja y media). Nivel 3 (alta) aparece con poca frecuencia. Esto demuestra que la mayoría de la población mantiene una **ingesta moderada**, aunque el pequeño grupo con consumo alto puede ser relevante para análisis de salud metabólica.



El análisis de tendencias temporales mostró que los promedios mensuales de las variables estudiadas —edad, IMC, ejercicio, sueño e ingesta de azúcar— se mantuvieron estables entre 2022 y 2024, sin fluctuaciones relevantes. Esta ausencia de variación es coherente con el uso de fechas generadas de manera aleatoria para fines metodológicos.

Aun así, la revisión temporal aporta valor al proceso exploratorio: permite confirmar que no existen picos artificiales ni comportamientos irregulares, valida la homogeneidad del dataset enriquecido y contribuye a la calidad del análisis exploratorio.

Planeación (Diagrama de Gantt)



El proyecto se desarrolló con metodología scrum, siguiendo la secuencia planteada en el diagrama de Gantt. Primero se dedicó tiempo a la ingesta, limpieza y preparación del dataset; luego se avanzó con el enriquecimiento de variables y el análisis exploratorio. Más adelante, se destinó un espacio específico para la construcción del dashboard en Power BI y, finalmente, se realizaron los ajustes y la documentación del informe. Esta organización permitió avanzar de manera ordenada y coherente entre cada etapa del proceso.

Presentación de Resultados

Características generales de la población

La población analizada se encuentra mayoritariamente entre los 45 y 59 años, seguida de cerca por adultos mayores de 60 años. Este predominio de personas en etapas avanzadas de la vida influye en varios de los comportamientos observados posteriormente, especialmente en lo relacionado con el IMC, el sueño y la actividad física.

Estado nutricional según IMC

Los resultados muestran una situación que merece atención:

- Apenas cerca del **28,8 %** de los individuos se encuentran dentro de un IMC saludable.
- Más de **la mitad de la población** presenta **sobrepeso u obesidad** en alguno de sus grados.

Este comportamiento refleja un riesgo metabólico significativo y sugiere que buena parte de las personas podrían beneficiarse de intervenciones relacionadas con nutrición, actividad física y hábitos de salud.

Relaciones entre variables

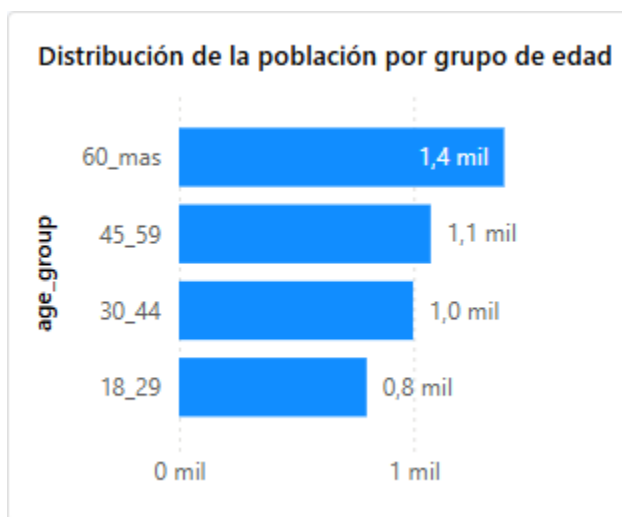
El análisis conjunto arroja patrones relevantes:

- **IMC y edad:** los grupos de mayor edad concentran las tasas más altas de sobrepeso y obesidad.
- **Ejercicio y azúcar:** quienes realizan menos ejercicio tienden también a consumir más azúcar, un hallazgo coherente con perfiles de riesgo metabólico.
- **Sueño y edad:** el sueño insuficiente se hace más visible en los grupos mayores, sugiriendo un deterioro progresivo del descanso con la edad.

Distribución de la Población por Grupo de Edad (Gráfico de Barras)

Este visual muestra el total de personas agrupado por rangos etarios (18–29, 30–44, 45–59, 60+).

Permite identificar cuál grupo representa la mayor proporción en el análisis y cómo se relaciona esto con otros factores de riesgo.



Tarjetas de Resumen de Indicadores Clave

Incluyen los promedios de edad, IMC, horas de sueño, ingesta de azúcar y nivel de ejercicio.

Permiten una comprensión rápida del estado general de la población, destacando un IMC promedio en rango de sobrepeso y niveles moderados de sueño y ejercicio.



Indicadores Clave: promedios de edad, IMC, horas de sueño, ingesta de azúcar y nivel de ejercicio

48,86

Prom Edad

26,80

Prom BMI

6,99

Prom Sueño (h)

1,91

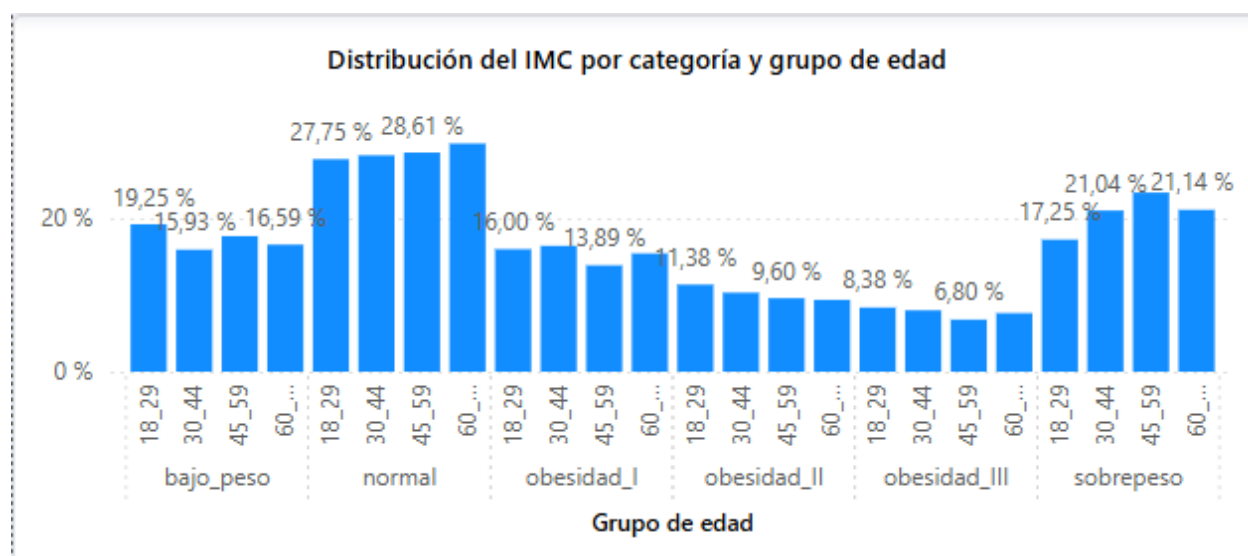
Prom Azúcar (nivel)

1,94

Prom Ejercicio (nivel)

Distribución del IMC por Categoría y Grupo de Edad (%)

El gráfico permite observar cómo se distribuyen las diferentes categorías de índice de masa corporal (IMC) dentro de cada grupo de edad. Al combinar ambas variables, se obtiene una visión más detallada del estado nutricional de la población, lo que facilita identificar patrones y segmentos de mayor riesgo.



Repositorio

https://github.com/JuanUrrego/EA_proyecto_integrado_V_20251-2

Conclusiones

El trabajo realizado permitió consolidar un conjunto de datos limpio, estructurado y enriquecido, adecuado para el análisis de estilos de vida y su relación con los riesgos de salud en población adulta de Medellín. A través de procesos rigurosos de depuración, normalización, categorización y generación de variables temporales, se logró transformar un dataset crudo en una fuente de información confiable y coherente. El análisis descriptivo y las visualizaciones desarrolladas aportaron una comprensión clara de los patrones presentes en la población, destacando tendencias relevantes en IMC, niveles de ejercicio, hábitos de sueño e ingesta de azúcar. Este resultado constituye un insumo analítico sólido que evidencia la calidad del tratamiento de los datos y la capacidad del proceso para revelar comportamientos significativos relacionados con la salud.

La construcción del dashboard en Power BI permitió visualizar de forma clara los patrones identificados en el análisis, facilitando la exploración de variables como IMC, ejercicio, sueño e ingesta de azúcar mediante tarjetas, gráficos y filtros interactivos. Esto hizo evidente la alta prevalencia de sobrepeso, la baja actividad física y las diferencias entre grupos de edad. En conjunto, el dashboard reforzó y amplió los hallazgos del análisis descriptivo, convirtiéndose en una herramienta útil para interpretar la información y apoyar la toma de decisiones sobre el bienestar de la población.

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).