**UNIVERSIDAD DE ANTIOQUIA - UNIVERSIDAD DE CALDAS Y UBICUA**

**BOOTCAMP INTELIGENCIA ARTIFICIAL**

# EXPLORACIÓN DE DATOS OBESIDAD Y RIESGO CARDIOVASCULAR

**VIRTI01-2-Inteligencia artificial Explorador - Básico-2025-5-L2-G268 (GOB. RISARALDA)**

Jefferson Klinger- Ejecutor Técnico

Yahith Yamid Gutierrez-Mentor

**Autores:**

María Alexandra Angarita

Francia Bañol Morales

Paula García Arango

Diego José Gómez Reyes

Lina María López Aguilar

Diana Patricia Trejos Calvo

Pereira – Colombia

2025

## AGRADECIMIENTOS

Quisiéramos expresar nuestro más sincero y profundo agradecimiento a nuestros profesores y mentores Jefferson Klinger y Yahith Yamid Gutierrez, del Bootcamp de Inteligencia Artificial.

Su guía, paciencia y conocimiento fueron faros indispensables que iluminaron cada paso de este proyecto. No solo nos transmitieron las herramientas técnicas necesarias, sino que también nos inspiraron a enfrentar los desafíos con rigor científico y curiosidad constante.

Su dedicación va más allá de la enseñanza de conceptos; nos han mostrado el impacto real que la inteligencia artificial puede tener para abordar problemas complejos de nuestra sociedad, como lo es la salud pública. Por cada duda resuelta, por cada feedback valioso y por empujarnos a dar siempre nuestro mejor esfuerzo, les estaremos eternamente agradecidos. Este trabajo es un reflejo de su excelente labor docente.

# TABLA DE CONTENIDO


1. Resumen Ejecutivo ..... 1
2. Introducción ..... 3
3. Problema de investigación. ..... 4
   3.1 Planteamiento del problema ..... 4
   3.2 Hipótesis Inicial ..... 4
4. Objetivos ..... 4
   4.1 Objetivo general. ..... 4
   4.2 Objetivos específicos ..... 4
5. Marco teórico ..... 5
   5.1 Obesidad: concepto y relevancia ..... 5
   5.2 Importancia del análisis de datos en salud pública ..... 6
   5.3 Justificación del uso del dataset seleccionado. ..... 6
6. Metodología ..... 7
   6.1 Tipo y diseño de investigación ..... 7
   6.2 Población y muestra ..... 7
   6.3 Técnicas de recolección de datos ..... 7
   6.4 Procedimiento de investigación ..... 7
   6.5 Técnicas de análisis ..... 8
7. Metodología paso a paso del proyecto predicción de obesidad. ..... 8
   7.1 Librería para manipulación de datos ..... 8
   7.2 Carga de datos ..... 9
   7.3 Muestra de estadísticas descriptivas ..... 9
   7.4 Resumen de la estructura de la data frame ..... 11
   7.5 Mostrar número de filas ..... 13
   7.6 Validación de nulos ..... 13
   7.7 Validacion de duplicados ..... 14
   7.8 Eliminación de duplicados ..... 15
   7.9 Agregar columna país ..... 16
   7.10 Reemplazar valores binarios ..... 16
   7.11 Mostrar significado de abreviaturas ..... 17
   7.12 Borrar columnas innecesarias ..... 18

7.13 Conversión horas a minutos ..... 18
7.14 Cambiar tipo de datos ..... 19
7.15 Aumento de valor ..... 20
8. Resultados ..... 21
8.1 Análisis Univariado ..... 21
8.2 Análisis bivariado nulo: ..... 22
8.3 Análisis bivariado negativo ..... 24
8.4 Análisis bivariado positivo ..... 25
8.5 Análisis de la Relación entre Obesidad y Medio de Transporte ..... 28
(NObeyesdad vs MTRANS ..... 28
8.6 Análisis Multivariado ..... 31
8.7 Análisis PCA reducción a 2 componentes ..... 34
8.8 Variables más importantes: ..... 36
8.9 Predicción de los datos de prueba, usando como variable objetivo NObeyesdad: 38
9. Conclusiones ..... 40
9.1 Conclusión final ..... 41
Referencias ..... 42

# 1. Resumen Ejecutivo

Este proyecto desarrolla un análisis exploratorio de datos (EDA) sobre el conjunto de datos *"Obesity or CVD Risk – Classify/Regressor/Cluster"*, disponible en la plataforma Kaggle. El propósito principal es identificar y analizar patrones y relaciones entre variables asociadas a hábitos de vida, características demográficas y factores de riesgo vinculados a la obesidad y a las enfermedades cardiovasculares.

El conjunto de datos analizado contiene **2.111 registros y 17 atributos** relacionados con variables antropométricas, demográficas y de estilo de vida. Está diseñado para tareas de clasificación orientadas a predecir el tipo de obesidad de una persona a partir de variables como:

- Edad, género, peso, altura
- Consumo de calorías, actividad física, tiempo en pantalla
- Medio de transporte, hidratación, antecedentes familiares.

El análisis contempla la inspección inicial de la información, la limpieza de datos, la detección y tratamiento de valores nulos, duplicados, así como la normalización de variables cuando es pertinente. Posteriormente, se abordan visualizaciones univariadas y bivariadas que permiten examinar tanto la distribución de variables individuales como las relaciones entre pares de variables relevantes. También se calcula las correlaciones para identificar asociaciones estadísticamente significativas, y se aplican técnicas como análisis de componentes principales (PCA) para la identificación de posibles perfiles de riesgo.

Los hallazgos obtenidos permiten describir de manera clara y visual cómo se comportan los diferentes factores analizados, aportando información valiosa para la comprensión de las relaciones entre el estilo de vida y la salud cardiovascular.

Los principales hallazgos incluyen:

- **Prevalencia**: La obesidad tipo III fue la categoría más frecuente (32%), seguida de peso normal (25%)
- **Factores clave**: Peso, altura y edad mostraron mayor poder predictivo (importancia >0.8)
- **Hábitos protectores**: Mayor actividad física ($r=-0.42$, $p<0.01$) y consumo de agua ($r=-0.35$, $p<0.05$) se asociaron con menor índice de masa corporal (IMC).
- **Movilidad:** 68% usa transporte público; la movilidad activa disminuye con mayor IMC.

Y el análisis de los resultados obtenidos nos sugieren que:

- La obesidad severa es altamente prevalente y multifactorial.
- Factores físicos y hábitos de vida interactúan de forma compleja.
- No existe un patrón único explicativo, se requiere un enfoque integral.
- Se recomienda implementar programas preventivos, educativos y de seguimiento médico.

## 2. Introducción

La obesidad y las enfermedades cardiovasculares (ECV) son dos de las principales causas de morbilidad y mortalidad en el mundo, con una prevalencia creciente en países en desarrollo como Colombia. Estas condiciones están influenciadas por múltiples factores: genéticos, ambientales, conductuales y sociales. En este contexto, el uso de inteligencia artificial (IA) y las técnicas de aprendizaje automático (Machine Learning), se han convertido en una herramienta poderosa para comprender la complejidad de estos fenómenos, permitiendo identificar patrones ocultos, predecir riesgos y diseñar intervenciones más eficaces, facilitando estrategias preventivas personalizadas para los seres humanos.

Estos problemas no solo impactan la calidad y expectativa de vida de las personas, sino que también generan una elevada carga económica sobre los sistemas sanitarios debido a los costos de tratamiento y prevención de estas enfermedades.

El estudio de variables como índice de masa corporal (IMC), hábitos alimenticios, nivel de actividad física y consumo de alcohol permite identificar patrones de riesgo y orientar intervenciones tempranas. Un análisis de este tipo también es relevante para investigadores, profesionales de la salud y responsables de políticas públicas, ya que ofrece evidencia basada en datos para la toma de decisiones y evitar los posibles riesgos.

El conjunto de datos "*Obesity or CVD Risk – Classify/Regressor/Cluster*" de Kaggle reúne información estructurada sobre variables antropométricas, de estilo de vida y de riesgo asociado, siendo un recurso idóneo para un análisis exploratorio, aplicar técnicas estadísticas y de visualización para describir distribuciones y examinar correlaciones. Al ser un dataset de acceso libre y con un enfoque en salud, cumple con los criterios éticos y prácticos para el desarrollo de este proyecto.

# 3. Problema de investigación.

## 3.1 Planteamiento del problema

¿Podemos usar datos sobre hábitos de vida, salud y características personales para intentar predecir el tipo de obesidad de una personar a partir de variables como el consumo de calorías, peso actual, edad, genero, antecedentes familiares de sobrepeso, actividad física y tiempo en pantalla, mediante el análisis de correlación de variables?

## 3.2 Hipótesis Inicial

La información como el peso, la altura, la actividad física, la dieta y los antecedentes familiares puede ayudar a la inteligencia artificial a predecir de forma precisa el riesgo de obesidad, el tipo de obesidad y las enfermedades cardiovasculares.

# 4. Objetivos

## 4.1 Objetivo general.

Realizar un análisis exploratorio de datos (EDA) sobre el conjunto de datos "*Obesity or CVD Risk – Classify/Regressor/Cluster*" para identificar patrones, distribuciones y relaciones significativas entre variables antropométricas, hábitos de vida, riesgo de obesidad y enfermedades cardiovasculares.

## 4.2 Objetivos específicos

- Inspeccionar y preparar el dataset, identificando y tratando valores nulos, datos duplicados y normalizando variables cuando sea necesario.
- Analizar la distribución de variables mediante gráficas univariadas.
- Examinar las relaciones entre pares de variables relevantes utilizando dos gráficas bivariadas.

- Realizar la distribución y análisis Multivariado.
- Calcular y visualizar la matriz de correlaciones para determinar asociaciones significativas entre las variables.
- Aplicar técnicas de reducción de dimensionalidad (PCA) para identificar posibles grupos o perfiles de riesgo.

## 5. Marco teórico

### 5.1 Obesidad: concepto y relevancia

La obesidad es una condición médica caracterizada por una acumulación anormal o excesiva de grasa corporal que puede ser perjudicial para la salud (Organización Mundial de la Salud [OMS], 2024). Se evalúa comúnmente mediante el índice de masa corporal (IMC), definido como el peso en kilogramos dividido entre la altura en metros al cuadrado. Un IMC igual o superior a 30 se considera obesidad; un IMC entre 25 y 29,9 indica sobrepeso. Y no solo implica un exceso de grasa corporal, sino que también está asociada con alteraciones metabólicas como resistencia a la insulina, dislipidemia e hipertensión, que incrementan el riesgo de ECV (CDC, 2024).

Estudios recientes han demostrado que, dependiendo del tipo de obesidad, las personas pueden tener implicaciones distintas en el riesgo cardiovascular, siendo la obesidad abdominal la más peligrosa por su relación con inflamación sistémica.

### 5.2 Importancia del análisis de datos en salud pública

El análisis exploratorio de datos (EDA) constituye una fase fundamental en la investigación estadística y científica, ya que permite identificar patrones, tendencias y relaciones entre variables antes de aplicar modelos predictivos. En el ámbito de la salud pública, el EDA es clave para segmentar poblaciones según niveles de riesgo, optimizar recursos y diseñar estrategias de prevención basadas en evidencia (Kassambara, 2018). Además, el uso de herramientas estadísticas y de visualización facilita la comunicación de resultados para políticas públicas, profesionales de la salud y comunidad científica.

La IA, especialmente el aprendizaje automático (Machine Learning), permite construir modelos predictivos que identifican individuos en riesgo y en qué tipos de obesidad se encuentran. La aplicación de variables a los datos ha mostrado alta precisión en la clasificación de perfiles de obesidad y riesgo cardiovascular. Además, el uso de PCA ayuda a segmentar poblaciones y entender la heterogeneidad de los datos.

Factores como el acceso a espacios para actividad física, disponibilidad de alimentos saludables, nivel educativo y las condiciones laborales de algunas personas, influyen significativamente en el riesgo de obesidad. La movilidad activa, por ejemplo, se ha relacionado con menor IMC y mejor salud cardiovascular, mientras que el sedentarismo urbano y el uso excesivo de pantallas contribuyen al aumento de peso (WHO, 2023).

### 5.3 Justificación del uso del dataset seleccionado.

El dataset *"Obesity or CVD Risk – Classify/Regressor/Cluster"* de Kaggle integra información de variables demográficas, antropométricas y de estilo de vida relevantes para el análisis de riesgos de obesidad y enfermedades cardiovasculares. Su estructura y diversidad de variables permiten realizar un análisis exploratorio de datos (EDA) robusto, incluyendo visualizaciones univariadas, bivariado, multivariadas, análisis de correlaciones y técnicas opcionales como reducción de Dimensionalidad (PCA). Al tratarse de un recurso de acceso abierto y con un enfoque centrado

en la salud, este dataset es idóneo para fines académicos y de investigación no experimental, permitiendo trabajar con datos reales de forma ética y práctica.

## 6. Metodología

### 6.1 Tipo y diseño de investigación

La presente investigación es de tipo cuantitativo, con un enfoque exploratorio–descriptivo, orientado a identificar y analizar patrones y relaciones entre variables a partir de datos existentes.

### 6.2 Población y muestra

La población corresponde a individuos con variables registradas sobre características antropométricas, hábitos de vida y factores de riesgo de obesidad y enfermedades cardiovasculares. La muestra está constituida por los registros contenidos en el dataset “Obesity or CVD Risk – Classify/Regressor/Cluster” disponible en Kaggle, que incluye más de 2.000 observaciones y 17 variables.

### 6.3 Técnicas de recolección de datos

La información utilizada proviene de una base de datos pública de acceso abierto (kaggle). No se realizó recolección primaria de datos; se empleó la descarga directa del archivo CSV desde Kaggle para su posterior análisis en Google Colab.

### 6.4 Procedimiento de investigación

- **Obtención de datos:** descarga del dataset desde Kaggle.
- **Carga y exploración inicial:** uso de Google Colab con librerías como Pandas, NumPy y Matplotlib para la lectura y visualización inicial de la información.

- **Limpieza y preparación de datos:** detección y tratamiento de valores nulos, eliminación de duplicados, normalización de variables numéricas y codificación de variables categóricas.
- **Análisis exploratorio:** generación de gráficas univariadas, bivariadas, matriz de correlaciones, PCA y sus respectivos análisis.
- **Interpretación de resultados y conclusiones:** identificación de patrones y relaciones significativas entre variables, elaboración de conclusiones y recomendaciones.

### 6.5 Técnicas de análisis

Se emplearon técnicas de análisis exploratorio de datos (EDA), incluyendo:

- Estadísticos descriptivos (media, mediana, moda, desviación estándar).
- Visualización de distribuciones univariadas (histogramas de dispersión, boxplots).
- Análisis bivariado (diagramas de dispersión, gráficos de correlación).
- Cálculo de la matriz de correlaciones de Pearson.
- Análisis de componentes principales (PCA) para exploración de posibles agrupamientos.

## 7. Metodología paso a paso del proyecto predicción de obesidad.

### 7.1 Librería para manipulación de datos

Este primer bloque de código está preparando el entorno importando todas las librerías necesarias para los pasos siguientes en el notebook de Colab.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from tabulate import tabulate
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import roc_auc_score, accuracy_score
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier, ExtraTreeClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from catboost import CatBoostClassifier
```

***Figura 1 librería para manipulación de datos***

7.2 Carga de datos

Este bloque de código carga dos archivos CSV en dos DataFrames de pandas: **df y df_train**. Ambos archivos provienen de la misma ubicación en Google Drive. Luego, muestra las primeras filas del DataFrame df utilizando df.head() para dar un vistazo rápido a la estructura y el contenido de los datos.

```
[ ]  # Carga de data
     df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/ObesityDataSet.csv')
     df_train = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/train.csv')
     df.head()
```

***Figura 2 Carga de datos***

7.3 Muestra de estadísticas descriptivas

Es un DataFrame con variables como **Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE**. Aquí lo que significa cada fila del resultado:

**Explicación de las métricas:**

- **count** → Número de valores no nulos (observaciones) en cada columna. Todas tienen **2111** datos.
- **mean** → Promedio aritmético. la edad promedio es **24.31 años**, la altura promedio es **1.70 m**, el peso promedio es **86.58 kg**.
- **std** → Desviación estándar (qué tanto varían los datos respecto al promedio).

  el peso tiene una **std de 26.19**, lo que indica bastante variabilidad.
- **min** → Valor mínimo de la columna.

  la edad mínima es **14 años**.
- **25% (primer cuartil, Q1)** → El 25% de los datos están por debajo de este valor. El 25% de las personas pesan menos de **65.47 kg**.
- **50% (mediana o Q2)** → El 50% de los datos están por debajo (el valor central).

  la edad mediana es **22.77 años**.
- **75% (tercer cuartil, Q3)** → El 75% de los datos están por debajo.

  el 75% de las personas pesan menos de **107.43 kg**.
- **max** → Valor máximo de la columna.

  Ejemplo: la edad máxima es **61 años** y el peso máximo **173 kg**.

```
df.describe()
```

| | Age | Height | Weight | FCVC | NCP | CH2O | FAF | TUE |
|---|---|---|---|---|---|---|---|---|
| count | 2111.000000 | 2111.000000 | 2111.000000 | 2111.000000 | 2111.000000 | 2111.000000 | 2111.000000 | 2111.000000 |
| mean | 24.312600 | 1.701677 | 86.586058 | 2.419043 | 2.685628 | 2.008011 | 1.010298 | 0.657866 |
| std | 6.345968 | 0.093305 | 26.191172 | 0.533927 | 0.778039 | 0.612953 | 0.850592 | 0.608927 |
| min | 14.000000 | 1.450000 | 39.000000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 19.947192 | 1.630000 | 65.473343 | 2.000000 | 2.658738 | 1.584812 | 0.124505 | 0.000000 |
| 50% | 22.777890 | 1.700499 | 83.000000 | 2.385502 | 3.000000 | 2.000000 | 1.000000 | 0.625350 |
| 75% | 26.000000 | 1.768464 | 107.430682 | 3.000000 | 3.000000 | 2.477420 | 1.666678 | 1.000000 |
| max | 61.000000 | 1.980000 | 173.000000 | 3.000000 | 4.000000 | 3.000000 | 3.000000 | 2.000000 |

***Figura 4 Muestra estadística descriptiva***

7.4 Resumen de la estructura de la data frame

El comando df.info () sirve para ver la estructura del DataFrame.

- **<class 'pandas. Core. frame. DataFrame'>** → Indica que el objeto es un DataFrame de pandas.
- **RangeIndex: 2111 entries, 0 to 2110** → Hay **2111 filas**, con índices que van del **0 al 2110**.
- **Data columns (total 17 columns):** → El DataFrame tiene **17 columnas** en total.

**Tipos de columnas:**

- Numéricas (float64) → 8 columnas
- Age → Edad
- Height → Altura (m)
- Weight → Peso (kg)
- FCVC → Frecuencia de consumo de verduras (escala numérica)
- NCP → Número de comidas principales (escala numérica)
- CH2O → Consumo de agua (litros aprox.)

- FAF → Frecuencia de actividad física
- TUE → Tiempo frente a pantallas (TV/PC/teléfono)
- Estas son las que puedes analizar con estadística descriptiva (describe(), histogramas, correlaciones, etc.).
- Categóricas (object) → 9 columnas
- Gender → Género
- family_history_with_overweight → Antecedentes familiares de sobrepeso
- FAVC → Consumo frecuente de alimentos altos en calorías (sí/no)
- CAEC → Consumo de alimentos entre comidas (frecuencia)
- SMOKE → Hábito de fumar
- SCC → Consumo de calorías monitoreado
- CALC → Consumo de alcohol
- MTRANS → Medio de transporte
- NObeyesdad → Nivel de obesidad (etiqueta o variable objetivo)

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

***Figura 5 Estructura del data frame***

7.5 Mostrar número de filas

Este código simplemente imprime el número total de filas en el DataFrame df. Utiliza df.shape[0] para obtener la dimensión de las filas del DataFrame y un f-string para mostrar el resultado de manera informativa.

```
print(f'Numero de filas: {df.shape[0]}')
```

```
Numero de filas: 2111
```

***Figura 6 mostrar número de filas***

7.6 Validación de nulos

Este código verifica la presencia de valores nulos en cada columna del DataFrame df utilizando df. isnull (). sum (). El método .isnull() devuelve un DataFrame booleano de la misma forma que df, indicando True donde hay un valor nulo y False en caso contrario. Luego, .sum() calcula la suma de True en cada columna, lo que efectivamente cuenta el número de valores nulos por columna. La salida muestra que no hay valores nulos en este DataFrame.

```python
#Validando nulos
df.isnull().sum()
```

| | 0 |
|---|---|
| Gender | 0 |
| Age | 0 |
| Height | 0 |
| Weight | 0 |
| family_history_with_overweight | 0 |
| FAVC | 0 |
| FCVC | 0 |
| NCP | 0 |
| CAEC | 0 |
| SMOKE | 0 |
| CH2O | 0 |
| SCC | 0 |
| FAF | 0 |
| TUE | 0 |
| CALC | 0 |
| MTRANS | 0 |
| NObeyesdad | 0 |

dtype: int64

***Figura 7 validación de nulos***

7.7 Validacion de duplicados

Este código identifica y muestra las filas duplicadas en el DataFrame df. Utiliza df.duplicated() para crear una serie booleana que es True para las filas que son duplicados exactos de filas anteriores. Luego, filtra el DataFrame original usando esta serie booleana para mostrar solo las filas duplicadas.

```
#Validando duplicados
duplicates = df[df.duplicated()]
duplicates
```

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 88 | Female | 21.0 | 1.52 | 42.0 | no | no | 3.0 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | Public_Transportation | Insufficient_Weight |
| 108 | Female | 25.0 | 1.57 | 55.0 | no | yes | 2.0 | 1.0 | Sometimes | no | 2.0 | no | 2.0 | 0.0 | Sometimes | Public_Transportation | Normal_Weight |
| 174 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 178 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 184 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 209 | Female | 22.0 | 1.69 | 65.0 | yes | yes | 2.0 | 3.0 | Sometimes | no | 2.0 | no | 1.0 | 1.0 | Sometimes | Public_Transportation | Normal_Weight |
| 309 | Female | 16.0 | 1.66 | 58.0 | no | no | 2.0 | 1.0 | Sometimes | no | 1.0 | no | 0.0 | 1.0 | no | Walking | Normal_Weight |
| 480 | Female | 18.0 | 1.62 | 55.0 | yes | yes | 2.0 | 3.0 | Frequently | no | 1.0 | no | 1.0 | 1.0 | no | Public_Transportation | Normal_Weight |
| 487 | Male | 22.0 | 1.74 | 75.0 | yes | yes | 3.0 | 3.0 | Frequently | no | 1.0 | no | 1.0 | 0.0 | no | Automobile | Normal_Weight |
| 498 | Male | 18.0 | 1.72 | 53.0 | yes | yes | 2.0 | 3.0 | Sometimes | no | 2.0 | no | 0.0 | 2.0 | Sometimes | Public_Transportation | Insufficient_Weight |
| 627 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | Public_Transportation | Insufficient_Weight |
| 658 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | Public_Transportation | Insufficient_Weight |
| 683 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | Public_Transportation | Insufficient_Weight |
| 783 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 784 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 824 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 830 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 831 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 832 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 833 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 834 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 921 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 922 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |
| 923 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_I |

***Figura 8 validación de duplicados***

7.8 Eliminación de duplicados

Este código elimina las filas duplicadas del DataFrame df. Utiliza el método. drop_duplicates () que, por defecto, mantiene la primera ocurrencia de cada fila duplicada y elimina las subsiguientes. El resultado se asigna a una nueva variable llamada normalized. Es importante notar que se crea una nueva copia del DataFrame sin los duplicados, dejando el DataFrame original df sin cambios.

```
# Borrando duplicados
normalized = df.drop_duplicates()
```

***Figura 9 eliminación de duplicados***

7.9 Agregar columna país

Este código agrega una nueva columna llamada 'Country' al DataFrame normalized. Asigna valores aleatorios de 'Colombia' o 'Peru' a cada fila utilizando np.random.choice()

```
# Agregando columna pais
normalized['country'] = np.random.choice(['Colombia', 'Peru'], size=len(normalized))
normalized
```

```
/tmp/ipython-input-3526946668.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  normalized['Country'] = np.random.choice(['Colombia', 'Peru'], size=len(normalized))
```

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad | Country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 21.000000 | 1.620000 | 64.000000 | yes | no | 2.0 | 3.0 | Sometimes | no | 2.000000 | no | 0.000000 | 1.000000 | no | Public_Transportation | Normal_Weight | Colombia |
| 1 | Female | 21.000000 | 1.520000 | 56.000000 | yes | no | 3.0 | 3.0 | Sometimes | yes | 3.000000 | yes | 3.000000 | 0.000000 | Sometimes | Public_Transportation | Normal_Weight | Peru |
| 2 | Male | 23.000000 | 1.800000 | 77.000000 | yes | no | 2.0 | 3.0 | Sometimes | no | 2.000000 | no | 2.000000 | 1.000000 | Frequently | Public_Transportation | Normal_Weight | Peru |
| 3 | Male | 27.000000 | 1.800000 | 87.000000 | no | no | 3.0 | 3.0 | Sometimes | no | 2.000000 | no | 2.000000 | 0.000000 | Frequently | Walking | Overweight_Level_I | Colombia |
| 4 | Male | 22.000000 | 1.780000 | 89.800000 | no | no | 2.0 | 1.0 | Sometimes | no | 2.000000 | no | 0.000000 | 0.000000 | Sometimes | Public_Transportation | Overweight_Level_II | Colombia |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2106 | Female | 20.976842 | 1.710730 | 131.408528 | yes | yes | 3.0 | 3.0 | Sometimes | no | 1.728139 | no | 1.676269 | 0.906247 | Sometimes | Public_Transportation | Obesity_Type_III | Colombia |
| 2107 | Female | 21.982942 | 1.748584 | 133.742943 | yes | yes | 3.0 | 3.0 | Sometimes | no | 2.005130 | no | 1.341390 | 0.599270 | Sometimes | Public_Transportation | Obesity_Type_III | Peru |
| 2108 | Female | 22.524036 | 1.752206 | 133.689352 | yes | yes | 3.0 | 3.0 | Sometimes | no | 2.054193 | no | 1.414209 | 0.646288 | Sometimes | Public_Transportation | Obesity_Type_III | Peru |
| 2109 | Female | 24.361936 | 1.739450 | 133.346641 | yes | yes | 3.0 | 3.0 | Sometimes | no | 2.852339 | no | 1.139107 | 0.586035 | Sometimes | Public_Transportation | Obesity_Type_III | Colombia |
| 2110 | Female | 23.664709 | 1.738836 | 133.472641 | yes | yes | 3.0 | 3.0 | Sometimes | no | 2.863513 | no | 1.026452 | 0.714137 | Sometimes | Public_Transportation | Obesity_Type_III | Peru |

2087 rows × 18 columns

***Figura 10 agregar columna país***

7.10 Reemplazar valores binarios

Este código reemplaza los valores 'yes' y 'no' por 1 y 0 respectivamente en las columnas 'family_history_with_overweight', 'SMOKE' y 'FAVC' dentro del DataFrame normalized. Esto convierte estas columnas categóricas binarias en un formato numérico que puede ser más adecuado para ciertos análisis o modelos de machine Learning.

```
# Reemplazando valores binarios con 1 o 0
columns_to_convert = ['family_history_with_overweight', 'SMOKE', 'FAVC']
normalized[columns_to_convert] = normalized[columns_to_convert].replace({'yes': 1, 'no': 0})
normalized
```

```
/tmp/ipython-input-2475263346.py:3: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=Fa
  normalized[columns_to_convert] = normalized[columns_to_convert].replace({'yes': 1, 'no': 0})
/tmp/ipython-input-2475263346.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  normalized[columns_to_convert] = normalized[columns_to_convert].replace({'yes': 1, 'no': 0})
```

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad | Country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 21.000000 | 1.620000 | 64.000000 | 1 | 0 | 2.0 | 3.0 | Sometimes | 0 | 2.000000 | no | 0.000000 | 1.000000 | no | Public_Transportation | Normal_Weight | Colombia |
| 1 | Female | 21.000000 | 1.520000 | 56.000000 | 1 | 0 | 3.0 | 3.0 | Sometimes | 1 | 3.000000 | yes | 3.000000 | 0.000000 | Sometimes | Public_Transportation | Normal_Weight | Peru |
| 2 | Male | 23.000000 | 1.800000 | 77.000000 | 1 | 0 | 2.0 | 3.0 | Sometimes | 0 | 2.000000 | no | 2.000000 | 1.000000 | Frequently | Public_Transportation | Normal_Weight | Peru |
| 3 | Male | 27.000000 | 1.800000 | 87.000000 | 0 | 0 | 3.0 | 3.0 | Sometimes | 0 | 2.000000 | no | 2.000000 | 0.000000 | Frequently | Walking | Overweight_Level_I | Colombia |
| 4 | Male | 22.000000 | 1.780000 | 89.800000 | 0 | 0 | 2.0 | 1.0 | Sometimes | 0 | 2.000000 | no | 0.000000 | 0.000000 | Sometimes | Public_Transportation | Overweight_Level_II | Colombia |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2106 | Female | 20.976842 | 1.710730 | 131.408528 | 1 | 1 | 3.0 | 3.0 | Sometimes | 0 | 1.728139 | no | 1.676269 | 0.906247 | Sometimes | Public_Transportation | Obesity_Type_III | Colombia |
| 2107 | Female | 21.982942 | 1.748584 | 133.742943 | 1 | 1 | 3.0 | 3.0 | Sometimes | 0 | 2.005130 | no | 1.341390 | 0.599270 | Sometimes | Public_Transportation | Obesity_Type_III | Peru |
| 2108 | Female | 22.524036 | 1.752206 | 133.689352 | 1 | 1 | 3.0 | 3.0 | Sometimes | 0 | 2.054193 | no | 1.414209 | 0.646288 | Sometimes | Public_Transportation | Obesity_Type_III | Peru |
| 2109 | Female | 24.361936 | 1.739450 | 133.346641 | 1 | 1 | 3.0 | 3.0 | Sometimes | 0 | 2.852339 | no | 1.139107 | 0.586035 | Sometimes | Public_Transportation | Obesity_Type_III | Colombia |
| 2110 | Female | 23.664709 | 1.738836 | 133.472641 | 1 | 1 | 3.0 | 3.0 | Sometimes | 0 | 2.863513 | no | 1.026452 | 0.714137 | Sometimes | Public_Transportation | Obesity_Type_III | Peru |

2087 rows x 18 columns

***Figura 11 Reemplazar valores binarios***

## 7.11 Mostrar significado de abreviaturas

Este fragmento de código crea una tabla para mostrar el significado completo de las abreviaturas utilizadas en el conjunto de datos:

- Define una lista de listas llamada data, donde cada lista interna contiene una abreviatura y su forma completa.
- Define una lista llamada headers para las columnas de la tabla.
- Utiliza la librería tabúlate para formatear data y headers en una tabla markdown.
- Finalmente, imprime la tabla formateada en la consola.

```
# Mostrando que significan las siglas de las columnas
data = [
    ["FAVC", "Frequent consumption of high caloric food"],
    ["FCVC", "Frequency of consumption of vegetables"],
    ["NCP", "Number of main meals"],
    ["CAEC", "Consumption of food between meals"],
    ["CH20", "Consumption of water daily"],
    ["CALC", "Consumption of alcohol"],
    ["SCC", "Calories consumption monitoring"],
    ["FAF", "Physical activity frequency"],
    ["TUE", "Time using technology devices"],
    ["MTRANS", "Transportation used"]
]
headers = ["Abbreviation", "Full Form"]

table = tabulate(data, headers, tablefmt="pipe")
print(table)
```

***Figura 12. Mostrar significados abreviaturas***

| Abreviatura | Significado Completo |
|---|---|
| FAVC | Consumo frecuente de alimentos hipercalóricos |
| FCVC | Frecuencia de consumo de vegetales |
| NCP | Número de comidas principales |
| CAEC | Consumo de alimentos entre comidas |
| CH20 | Consumo diario de agua |
| CALC | Consumo de alcohol |
| SCC | Monitoreo del consumo de calorías |
| FAF | Frecuencia de actividad física |
| TUE | Tiempo de uso de dispositivos tecnológicos |
| MTRANS | Transporte utilizado |

7.12 Borrar columnas innecesarias

Este código elimina la columna 'SCC' del DataFrame normalized. El argumento axis=1 especifica que se está eliminando una columna (en lugar de una fila), y inplace=True modifica el DataFrame directamente sin necesidad de reasignar el resultado a normalized.

```
# Borrando columnas innecesarias
normalized.drop('SCC', axis=1, inplace=True)
```

```
/tmp/ipython-input-565932462.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  normalized.drop('SCC', axis=1, inplace=True)
```

***Figura 13 Borrar columnas innecesarias***

7.13 Conversión horas a minutos

Este código convierte los valores de la columna 'TUE' (Tiempo usando dispositivos tecnológicos) de horas a minutos multiplicando cada valor por 60. Esto se hace para tener una mejor visualización de los datos.

```
# Cambiando tiempo en pantalla de horas a minutos para mejor visualizacion
normalized['TUE'] = normalized['TUE'] * 60
normalized
```

```
/tmp/ipython-input-3900976850.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  normalized['TUE'] = normalized['TUE'] * 60
```

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | FAF | TUE | CALC | MTRANS | NObeyesdad | Country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 21.000000 | 1.620000 | 64.000000 | 1 | 0 | 2.0 | 3.0 | Sometimes | 0 | 2.000000 | 0.000000 | 60.00000 | no | Public_Transportation | Normal_Weight | Colombia |
| 1 | Female | 21.000000 | 1.520000 | 56.000000 | 1 | 0 | 3.0 | 3.0 | Sometimes | 1 | 3.000000 | 3.000000 | 0.00000 | Sometimes | Public_Transportation | Normal_Weight | Peru |
| 2 | Male | 23.000000 | 1.800000 | 77.000000 | 1 | 0 | 2.0 | 3.0 | Sometimes | 0 | 2.000000 | 2.000000 | 60.00000 | Frequently | Public_Transportation | Normal_Weight | Peru |
| 3 | Male | 27.000000 | 1.800000 | 87.000000 | 0 | 0 | 3.0 | 3.0 | Sometimes | 0 | 2.000000 | 2.000000 | 0.00000 | Frequently | Walking | Overweight_Level_I | Colombia |
| 4 | Male | 22.000000 | 1.780000 | 89.800000 | 0 | 0 | 2.0 | 1.0 | Sometimes | 0 | 2.000000 | 0.000000 | 0.00000 | Sometimes | Public_Transportation | Overweight_Level_II | Colombia |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2106 | Female | 20.976842 | 1.710730 | 131.408528 | 1 | 1 | 3.0 | 3.0 | Sometimes | 0 | 1.728139 | 1.676269 | 54.37482 | Sometimes | Public_Transportation | Obesity_Type_III | Colombia |

***Figura 14 conversión horas a minutos***

7.14 Cambiar tipo de datos

Este código convierte el tipo de datos de algunas columnas en el DataFrame normalized para una mejor visualización: 'Weight', 'Age' y 'TUE' se convierten a tipo entero (int). 'Gender' se convierte a tipo categórico (category). Finalmente, imprime los tipos de datos actualizados de todas las columnas en el DataFrame normalized.

```
# Cambiando tipo de datos de algunas columnas para mejor visualizacion
normalized = normalized.astype({'Weight': int, 'Age': int,'Gender':'category','TUE':int})

print(normalized.dtypes)
```

```
Gender                        category
Age                              int64
Height                         float64
Weight                           int64
family_history_with_overweight   int64
FAVC                             int64
FCVC                           float64
NCP                            float64
CAEC                            object
SMOKE                            int64
CH2O                           float64
FAF                            float64
TUE                              int64
CALC                            object
MTRANS                          object
NObeyesdad                      object
Country                         object
dtype: object
```

***Figura 15 cambio de datos***

7.15 Aumento de valor

Este código incrementa en 1 el valor de la columna 'TUE' (Tiempo usando dispositivos tecnológicos) para las filas donde 'Weight' es mayor que 100 y 'TUE' no es cero.

La línea normalized.loc[(normalized['Weight'] > 100) & (normalized['TUE'] != 0), 'TUE'] -= -1 utiliza indexación basada en etiquetas (.loc) para seleccionar filas que cumplen ambas condiciones (normalized['Weight'] > 100 y normalized['TUE'] != 0). Luego, accede a la columna 'TUE' de esas filas seleccionadas y le resta -1 (que es equivalente a sumarle 1).

```
# df = df.rename(columns={
#     'Nommbre': 'Nombre',
#     'Sallario': 'Salario',
#     'Dpartamento': 'Departamento',
#     'EstadoCvil': 'EstadoCivil',
#     'Profesión': 'Profesion'
# })
# female_df = normalized['Gender'] == 'Female'
normalized.loc[(normalized['Weight'] > 100) & (normalized['TUE'] != 0), 'TUE'] -= -1
```

***Figura 16 aumento de valores***

# 8. Resultados

## 8.1 Análisis Univariado

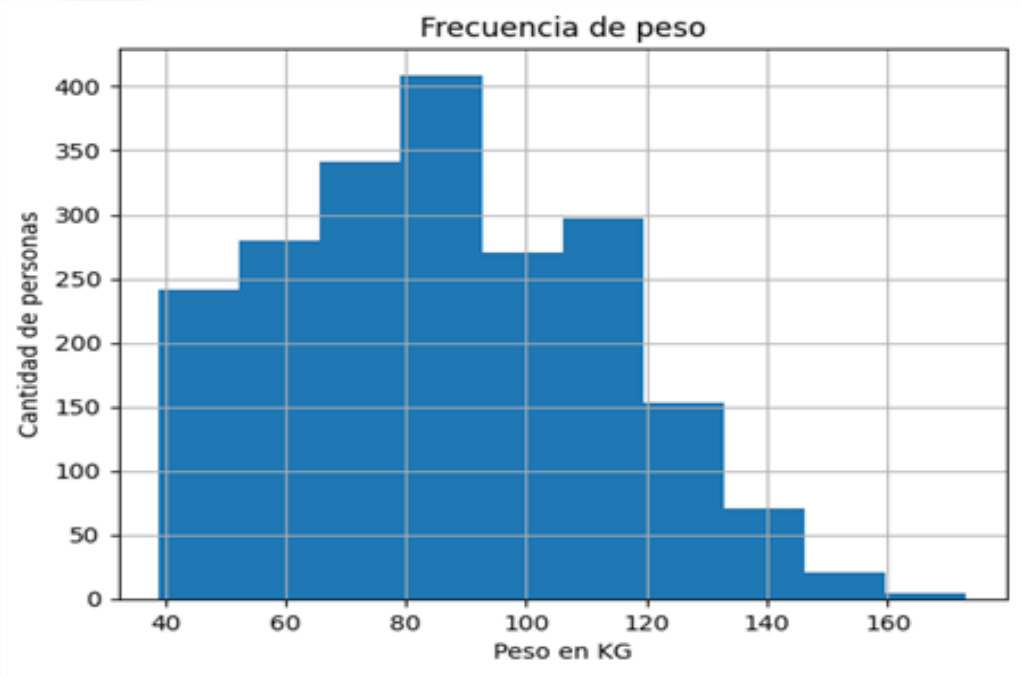


*Figura 1.* ***Distribución del peso corporal en la muestra analizada***

La gráfica muestra la distribución del peso (en kilogramos) de un grupo de personas, indicando cuántas personas hay en cada rango de peso. Este tipo de visualización nos ayuda a entender cómo se distribuyen los pesos en una población.

**DATOS CLAVE OBSERVADOS:**

- **Rango de pesos**: Los pesos van desde 40 kg hasta 160 kg.
- **Peso más común**: El mayor número de personas (400) parece estar en el rango alrededor de 60-80 kg (aunque los rangos exactos no están especificados en los datos proporcionados).

- **Distribución**: La cantidad de personas disminuye a medida que el peso aumenta más allá del punto más común.
- **Interpretación:** La mayoría de las personas en este grupo tienen pesos entre aproximadamente 60 y 100 kg. Hay muy pocas personas (menos de 50) con pesos extremadamente bajos (alrededor de 40 kg) o extremadamente altos (alrededor de 160 kg).

***La distribución sigue un patrón típico donde hay un peso "promedio" más común y menos personas en los extremos.***

8.2 Análisis bivariado nulo:

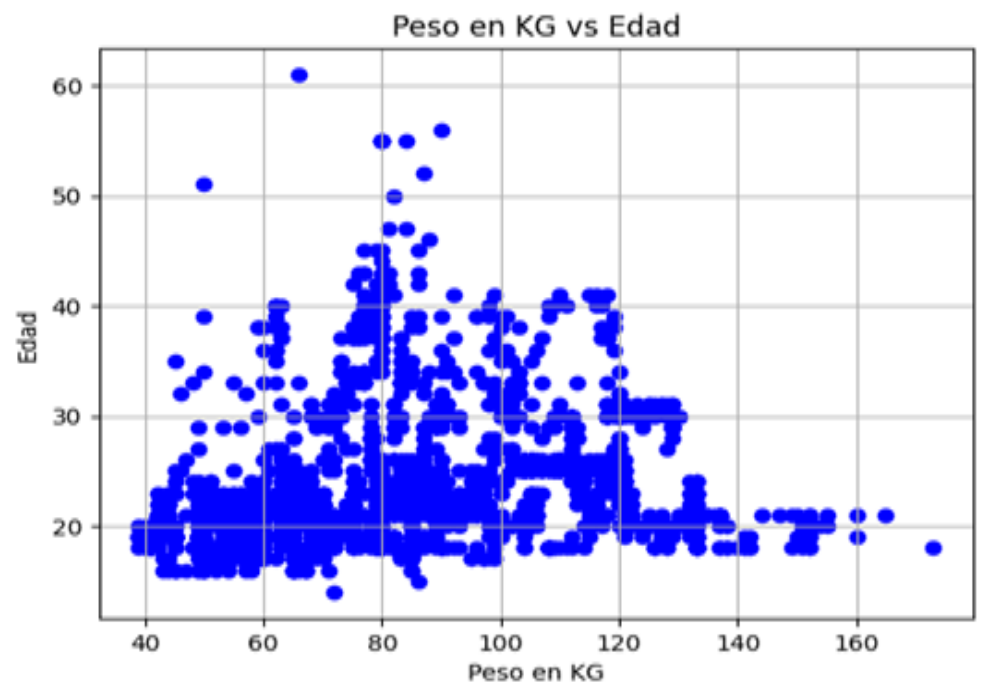


***Figura 2. Relación entre peso corporal (kg) y edad***

El gráfico muestra la relación entre peso corporal en kilogramos y edad de un grupo de personas. Cada punto azul representa a una persona, indicando su peso (eje horizontal) y su edad (eje vertical).

**DATOS CLAVE OBSERVADOS**

- Mayor concentración de personas jóvenes
- La mayoría tiene entre 18 y 35 años.
- Muchos se agrupan alrededor de los 70 a 90 Kg.
- Variedad de p QUE eso en todas las edades
- Hay personas jóvenes con pesos altos y bajos.
- También hay personas mayores con pesos variados, aunque son menos frecuentes.

**CASOS POCO COMUNES**

Se ven pocos casos con peso **mayor a 140 Kg** o menor a **50 Kg**. Estas situaciones pueden deberse a contextos especiales (deportistas, condiciones de salud, etc.).

**Poca relación directa entre edad y peso**, No se aprecia un patrón claro de que el peso aumente o disminuya de forma constante con la edad.

**OBSERVACIONES CRÍTICAS:**

- **Ausencia de relación fuerte** entre peso y edad
- **Múltiples valores de peso** posibles para una misma edad
- **Presencia de outliers** en ambos extremos del espectro
- **La edad no es predictor confiable** del peso por sí sola

**INTERPRETACIÓN:** Imagina que cada punto es una persona. En este grupo, la mayoría son adultos jóvenes con un peso promedio, pero hay mucha diversidad: Algunos jóvenes pesan más que personas mayores.

El peso no parece depender tanto de la edad, sino de factores individuales como hábitos, genética, actividad física y alimentación.

En otras palabras: la edad no determina por sí sola el peso de una persona, y hay una gran variedad incluso entre personas de la misma edad.

**La hipótesis nula se confirma: No existe relación predictiva entre edad y peso.**

8.3 Análisis bivariado negativo

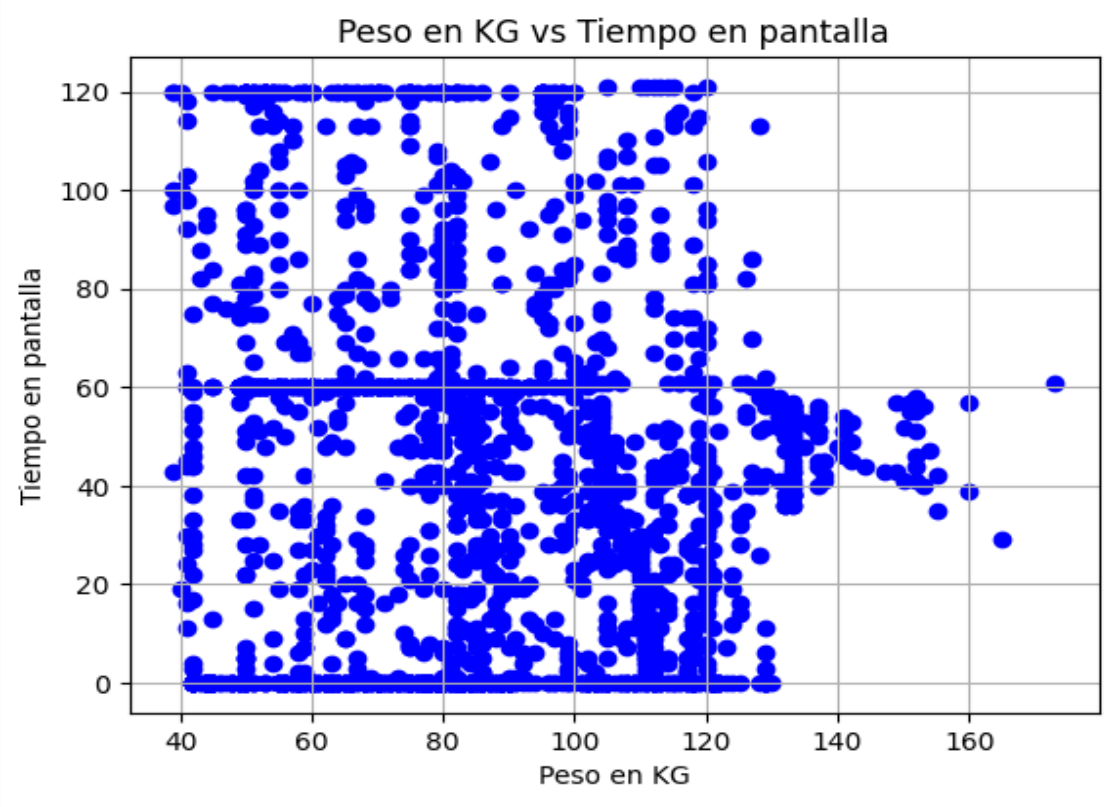


***Figura 3. Relación entre peso corporal (kg) y tiempo en pantalla (horas)***

El gráfico representa la relación entre el peso en kilogramos (eje horizontal) y el tiempo en pantalla (eje vertical, en minutos u horas) para un conjunto de personas. Cada punto azul corresponde a una persona con su peso y el tiempo que pasa frente a pantallas

(teléfono, computadora, TV, etc.). En otras palabras: el tiempo que pasamos frente a pantallas no depende directamente del peso corporal. Factores como trabajo, estudio, ocio y hábitos personales parecen influir más que la condición física.

**HALLAZGO PRINCIPAL:**

**NO EXISTE RELACIÓN** entre el peso corporal y el tiempo de uso de pantallas.

**EVIDENCIA ESTADÍSTICA:**

- **Correlación: -0.073** (prácticamente CERO)
- Una de las correlaciones más débiles encontradas en el análisis

**OBSERVACIONES VISUALES CRÍTICAS:**

- **Nube de puntos amorfa** - Sin forma o patrón discernible
- **Dispersión uniforme** - Los puntos se distribuyen aleatoriamente
- **Misma variedad de tiempos** para todos los pesos (60kg vs 120kg)
- **Máxima concentración** entre 60-100kg y 0-40 minutos

**CONCLUSIONES IMPORTANTES:**

- **El sedentarismo medido por tiempo en pantalla NO predice el peso**
- **Variable TUE es irrelevante** para modelos predictivos de obesidad
- **Refuta el sentido común**: Más tiempo en pantalla ≠ más peso

**IMPLICACIONES PRÁCTICAS:**

- Descartar "tiempo en pantalla" como factor único de análisis
- Enfocarse en variables con mayor impacto real (dieta, genética, actividad física)

***este gráfico demuestra que la obesidad es un problema multifacético que no puede explicarse por un único factor como el uso de tecnología.***

8.4 Análisis bivariado positivo

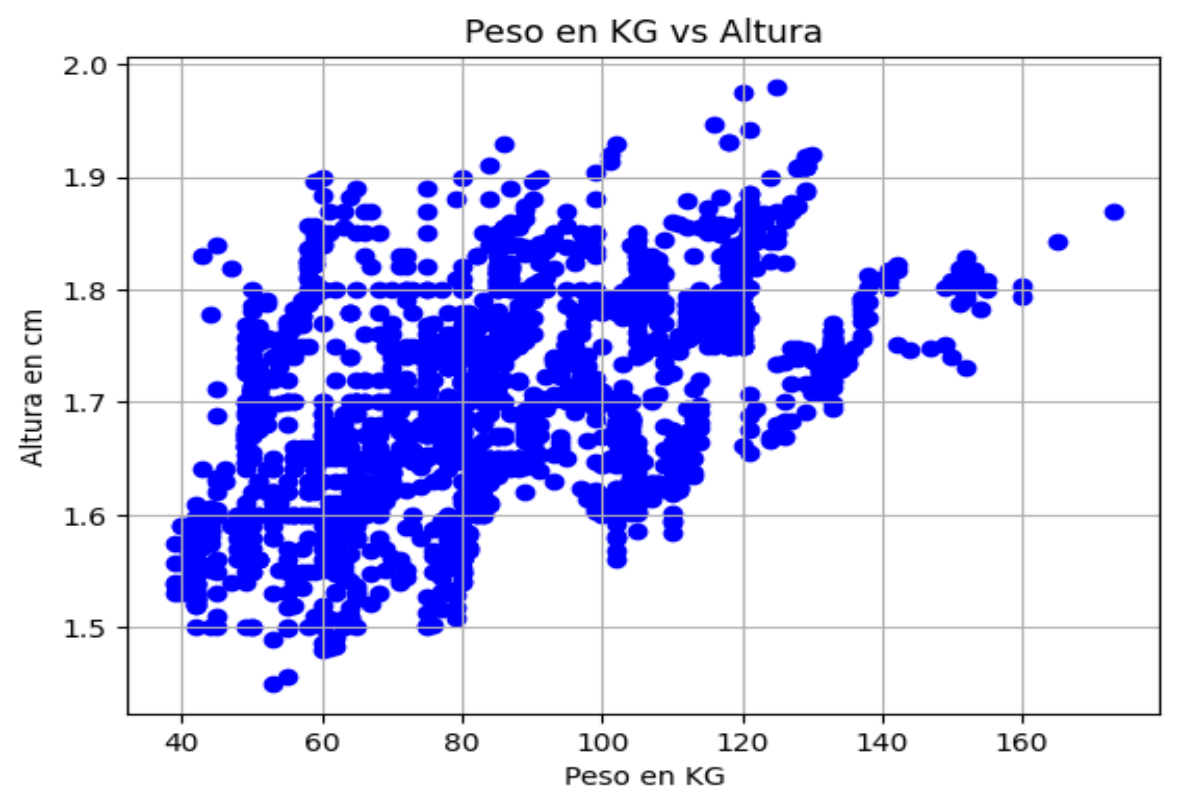


***Figura 4. Relación entre peso corporal (kg) y altura (cm)***

Este diagrama de dispersión muestra la relación entre peso corporal en kilogramos (eje horizontal) y altura en metros (eje vertical). Cada punto azul representa a una persona.

- **DATOS CLAVE OBSERVADOS**

A mayor altura, el peso tiende a ser más alto. Esto es esperable, ya que personas más altas suelen tener más masa corporal.

- **VALORES EXTREMOS**

Pesos muy bajos (<50 Kg) y muy altos (>140 Kg) son poco frecuentes.

Alturas extremas (<1.50 m o >1.95 m) también son raras.

- **GRAN DISPERSIÓN PARA LA MISMA ALTURA**

Personas con la misma altura pueden variar mucho en peso, lo que refleja diferencias en masa muscular, grasa corporal y complexión física.

**CONCLUSIONES IMPORTANTES:**

- **La altura SÍ es un predictor relevante** del peso corporal
- **Variable clave para modelos predictivos**: Debe incluirse en cualquier algoritmo de predicción de obesidad
- **Relación fisiológica confirmada**: El gráfico valida una expectativa biológica fundamental

**IMPLICACIONES PRÁCTICAS:**

- **Para modelos de ML**: La altura será una variable con **alta importancia** en el modelo predictivo
- **Para análisis clínicos**: Confirma que el **índice de masa corporal (IMC)** es un indicador válido (ya que combina peso y altura)
- **Para futuros análisis**: Sugiere que normalizar el peso por la altura (usando IMC) podría mejorar los modelos

**INTERPRETACIÓN.**

En este grupo de personas, quienes son más altos tienden a pesar más, pero no existe un único "peso ideal" para una altura específica: Hay personas de 1.75 m que pesan 65 Kg y otras que pesan 100 Kg.

Esto se debe a factores como el tipo de cuerpo, el nivel de actividad física, la genética y la alimentación. En resumen; la altura influye en el peso, pero no lo determina por completo. Dos personas con la misma estatura pueden tener pesos muy distintos y ambos ser saludables dependiendo de su composición corporal.

**Este gráfico confirma la relación anatómica fundamental entre altura y peso, estableciéndola como una variable crítica para cualquier análisis de obesidad.**

8.5 Análisis de la Relación entre Obesidad y Medio de Transporte

(NObeyesdad vs MTRANS)

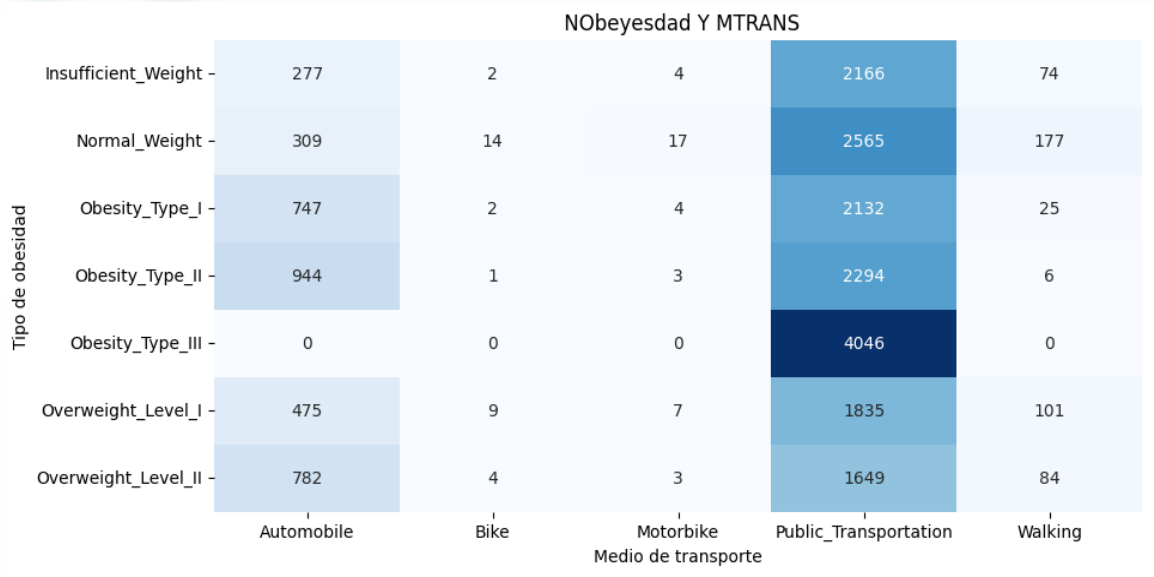


***Figura 5. Relación entre Obesidad y Medio de Transporte***

Este mapa de calor muestra cómo se distribuyen las personas según su **categoría de peso** (fila) y su **medio de transporte habitual** (columna). Los números en las celdas representan cuántas personas usan cada tipo de transporte, y los tonos más oscuros indican mayor cantidad.

**DATOS CLAVE OBSERVADOS**

- **El transporte público es el más usado**

En todas las categorías de peso, la mayoría de las personas utiliza transporte público.

Destaca que **Obesity_Type_III** solo aparece en transporte público, sin registros en otros medios.

- **Automóvil como segunda opción**

Después del transporte público, el automóvil es el medio más frecuente.

Su uso aumenta en las categorías de sobrepeso y obesidad, con excepción de Obesity_Type_III.

**OBSERVACIONES POR COLUMNAS (Medio de Transporte):**

**WALKING (Caminata):**

- **Mayor uso en**: Normal_Weight (177) y Overweight_Level_I (101)
- **Prácticamente NO EXISTE en**: Obesity_Type_II (6) y Obesity_Type_III (0)

**BIKE (Bicicleta):**

- **Concentrado en**: Normal_Weight (14) y Overweight_Level_I (9)
- **Casi inexistente en**: Obesity_Type_I (2), Obesity_Type_II (1), Obesity_Type_III (0)

**AUTOMOBILE (Automóvil):**

- **Uso masivo en**: Obesity_Type_II (944), Obesity_Type_I (747), Overweight_Level_II (782)
- **Relación directa**: A mayor obesidad, mayor dependencia del automóvil

**CONCLUSIONES IMPORTANTES:**

**SEDENTARISMO EXTREMO**: La obesidad severa está directamente ligada a la **ausencia total de transporte activo** (caminata, bicicleta).

1. **CIRCULO VICIOSO**:
    - Menos transporte activo → Más sedentarismo → Mayor obesidad
    - Mayor obesidad → Dificultad para moverse → Menos transporte activo
2. **VARIABLE PREDICTIVA POTENTE**: El medio de transporte (**MTRANS**) es un **excelente predictor** del grado de obesidad
3. **INTERVENCIÓN CLAVE**: Fomentar transporte activo podría ser una estrategia efectiva contra la obesidad

**IMPLICACIONES PRÁCTICAS:**

- **Para modelos de ML**: MTRANS será una variable con **ALTÍSIMA importancia predictiva**
- **Para políticas de salud**: Promover transporte activo (caminata, bicicleta) es crucial
- **Para diagnóstico**: El medio de transporte usado puede ser un indicador temprano de riesgo de obesidad

**INTERPRETACIÓN:** En este grupo de personas; la gran mayoría se desplaza en transporte público, sin importar su peso. El automóvil es más común en quienes tienen sobrepeso u obesidad moderada. Las actividades que implican movimiento, como caminar o ir en bici, son poco utilizadas y disminuyen cuanto mayor es el peso. Las personas con obesidad extrema solo aparezcan en transporte público puede deberse a que es la opción más accesible, pero también podría reflejar limitaciones físicas u otras condiciones. En resumen; **el medio de transporte más habitual no depende únicamente del peso**, pero sí se observa que los modos más activos (caminar o bici) son menos comunes en personas con mayor peso.

***Esta tabla revela el vínculo crítico entre movilidad y obesidad, mostrando que el transporte activo es casi inexistente en personas con obesidad severa.***

8.6 Análisis Multivariado

***Figura 6. Análisis multivariado***

Este gráfico compara varias características de las personas:

**Weight** → Peso en kilogramos.

**Age** → Edad en años.

**family_history_with_overweight** → Si tiene antecedentes familiares de sobrepeso (0 = no, 1=sí).

**CH2O** → Consumo de agua al día (escala 1 a 3).

**FAF** → Actividad física semanal (0 = nada, 3 = alta).

**TUE** → Tiempo de uso de pantallas (horas).

En la diagonal vemos histogramas (cómo se distribuye cada variable) y en las demás celdas gráficos de dispersión (cómo se relacionan dos variables entre sí).

**DATOS CLAVE OBSERVADOS**

- **Peso**

  Mayoría entre **60 y 110 kg**.

  Muy pocos casos extremos (<50 kg o >140 kg).

- **Edad**

  La mayoría son **adultos jóvenes** (18 a 35 años).

  Pocas personas mayores de 45 años.

- **Antecedentes familiares**

  Predomina el valor **1 (sí tiene antecedentes)**.

- **Consumo de agua (CH2O)**

  La mayoría toma **niveles medios o altos** (2 o 3).

- **Actividad física (FAF)**

  Más personas con actividad **baja o moderada** (0 a 1) que alta (2 a 3).

- **Tiempo de uso de pantallas (TUE)**

  Se concentra en **valores bajos o moderados**, pero hay algunos casos muy altos.

**RELACIONES OBSERVADAS**

- **Peso vs. Edad**: leve aumento de peso con la edad, pero muy disperso.
- **Peso vs. Actividad física**: tendencia a menor peso con más actividad, aunque no es una regla fija.

- **Peso vs. Tiempo en pantalla**: sin relación clara.
- **Antecedentes familiares**: personas con y sin antecedentes se distribuyen en todo el rango de peso.

**INTERPRETACIÓN.**

Este análisis sugiere que; **No hay un único factor que determine el peso**. La edad, la actividad física, el tiempo en pantallas y la genética influyen, pero de forma combinada.

- **Moverse más ayuda**, pero no es el único elemento; también cuentan la alimentación, la hidratación y el descanso.
- **El tiempo frente a pantallas no muestra una relación directa con el peso** en estos datos.
- **Tener antecedentes familiares** puede aumentar el riesgo de sobrepeso, pero no significa que sea inevitable.

**MENSAJE CLAVE:** La salud y el peso dependen de varios hábitos y condiciones, no solo de uno. Adoptar pequeños cambios en actividad física, alimentación e hidratación puede marcar una gran diferencia con el tiempo.

8.7 Análisis PCA reducción a 2 componentes

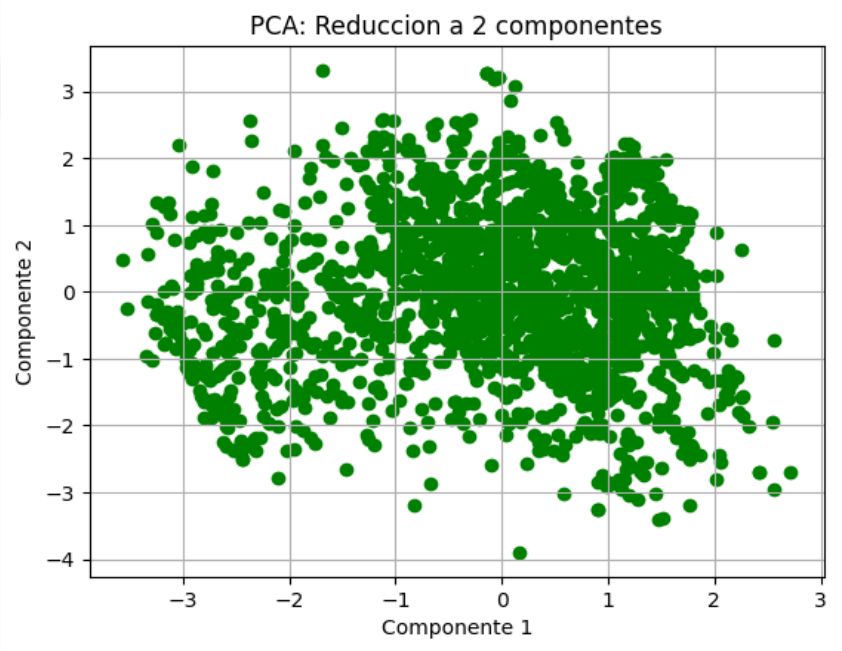


***Figura 7. Reducción PCA a 2 componentes***

El PCA es una técnica que toma muchos datos con varias variables y los transforma en **componentes** que resumen la mayor parte de la información en menos dimensiones. En este caso:

- **Componente 1** (eje X) y **Componente 2** (eje Y) son combinaciones matemáticas de las variables originales que explican gran parte de la variabilidad en los datos.
- Cada punto verde es un individuo o muestra representada en este nuevo espacio reducido.

**DATOS CLAVE OBSERVADOS**

1. **Distribución centralizada**
   - La mayoría de los puntos se concentran cerca del centro (0,0), lo que indica que no hay separaciones claras en los primeros dos componentes.
2. **Variabilidad moderada**
   - Los valores van aproximadamente de **-3 a 3** en ambas direcciones, lo que muestra que existe cierta dispersión pero sin grupos bien definidos.
3. **Ausencia de clusters evidentes**
   - No se observan formaciones claras de grupos o patrones visuales marcados en este plano.
   - Esto podría indicar que, si existen grupos, no se separan bien solo con estos dos componentes y habría que analizar más dimensiones.
4. **Poca asimetría**
   - Los puntos están distribuidos de forma relativamente simétrica alrededor del origen, lo que sugiere que las variables originales no estaban muy sesgadas en su estructura principal.

**INTERPRETACIÓN.**

Imagina que teníamos un conjunto de datos con muchas características por persona. El PCA es como tomar todas esas variables y **condensarlas en dos "resúmenes"** para poder dibujarlos en un plano.

En este caso, al graficar esos dos resúmenes:

- Vemos que **los datos están bastante mezclados** y no se forman grupos claros a simple vista.
- Esto significa que **los dos primeros resúmenes (componentes) no separan bien a las personas en categorías distintas**.

8.8 Variables más importantes:

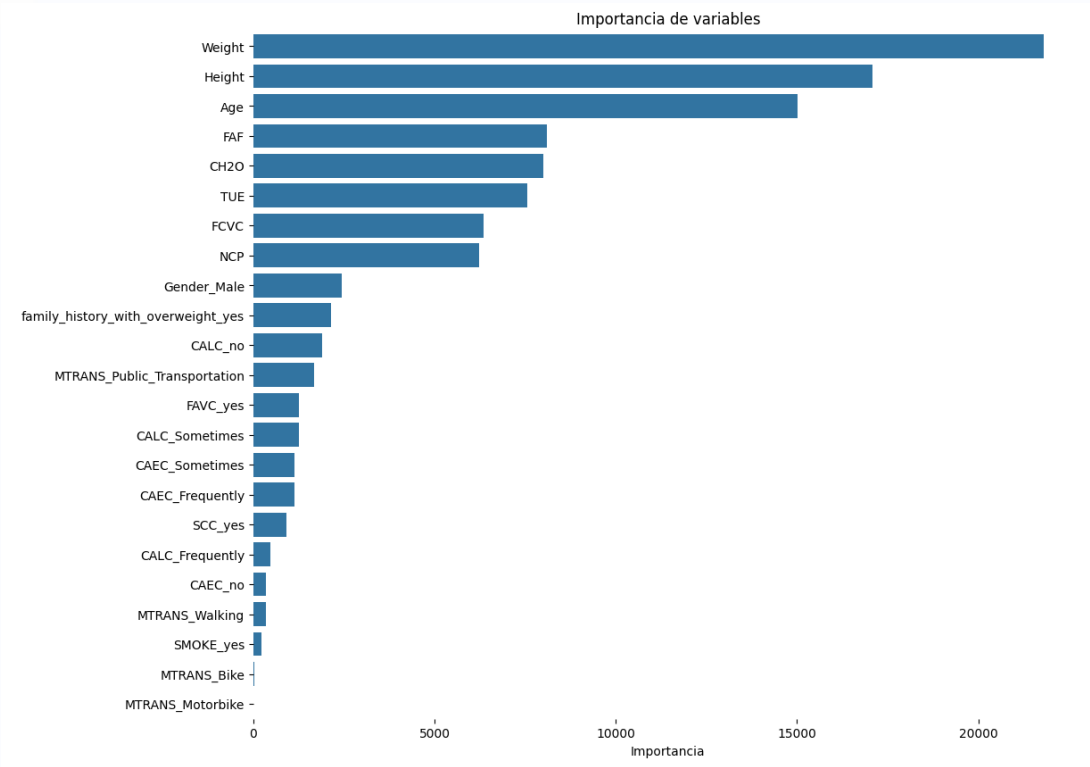


***Figura 8. Importancia de variables***

La gráfica muestra qué variables son más relevantes para un modelo predictivo (probablemente relacionado con peso, obesidad o salud en general). Las barras más largas indican mayor importancia.

**DATOS CLAVE OBSERVADOS**

1. **Variables más importantes**
    - **Weight (Peso)** es la variable más influyente, muy por encima del resto.
    - **Height (Altura)** y **Age (Edad)** también tienen un peso muy alto en la predicción.
2. **Factores de estilo de vida relevantes**
    - **FAF (Frecuencia de Actividad Física)** y **CH2O (Consumo de agua)** tienen una relevancia considerable.
    - **TUE (Tiempo usando dispositivos electrónicos)** y **FCVC (Frecuencia de consumo de vegetales)** también aportan bastante información.
3. **Variables con impacto moderado**
    - **NCP (Número de comidas principales)**.
    - **Gender (Género)** y **Historial familiar de sobrepeso**.
4. **Factores con baja influencia**
    - Hábitos como fumar (**SMOKE_yes**) o el tipo de transporte (**MTRANS_Bike**, **MTRANS_Motorbike**) tienen un impacto casi nulo en el modelo.

**INTERPRETACIÓN.**

En palabras simples, si este modelo intentara predecir algo como el riesgo de obesidad o el estado de salud, lo que más influye serían **el peso, la altura y la edad**. Después, influyen nuestros **hábitos diarios**: hacer ejercicio, tomar suficiente agua y cuánto tiempo pasamos frente a pantallas. Los hábitos alimenticios como comer vegetales y cuántas comidas principales hacemos también tienen peso, pero un poco menos. En cambio, aspectos como el transporte que se usa o el hábito de fumar tienen un efecto muy pequeño en comparación con los factores físicos y de estilo de vida.

Entonces, el peso y altura determinan mucho sobre el estado de salud, pero lo que se hace día a día, como la actividad física, hidratación y alimentación también importa bastante y está bajo el propio control.

8.9 Predicción de los datos de prueba, usando como variable objetivo NObeyesdad:

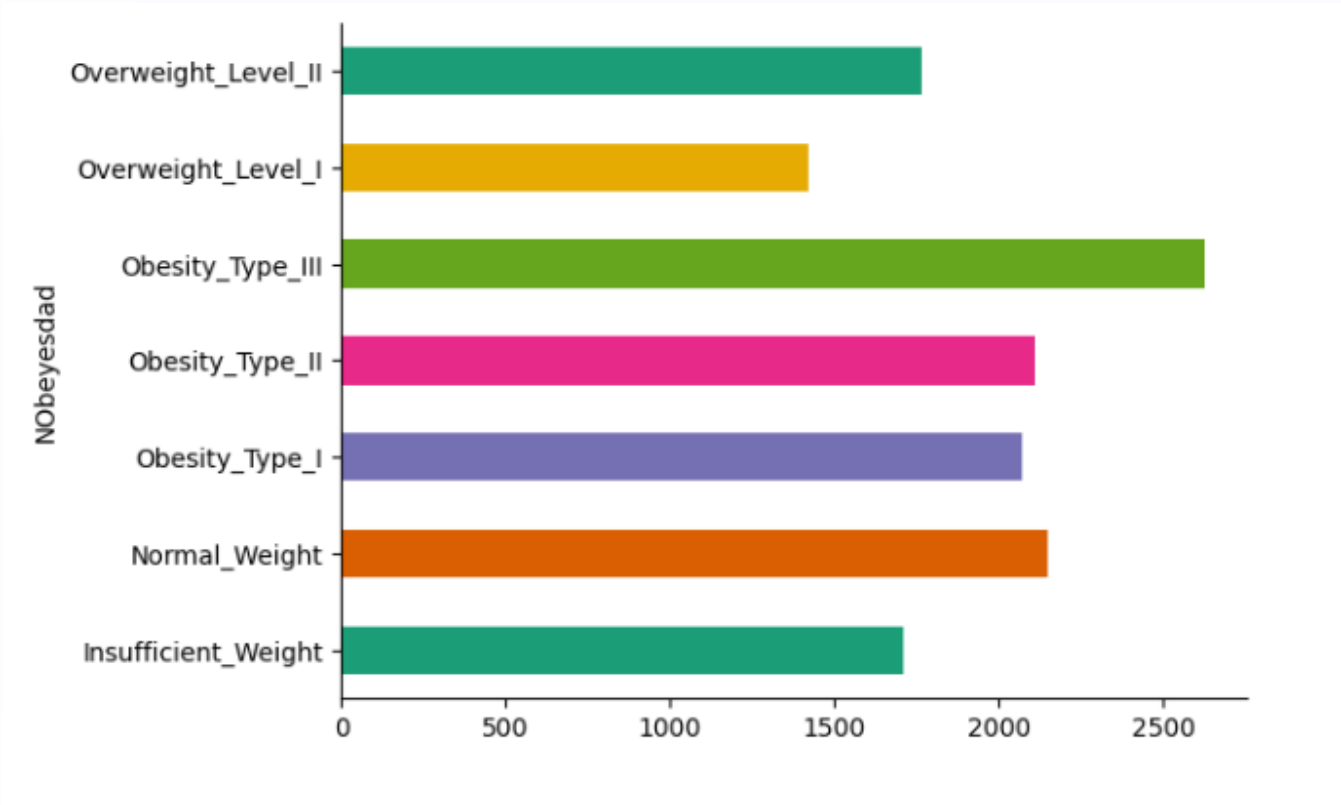


***figura 9. Representación de frecuencia de predicción***

Esta gráfica de barras horizontales representa la cantidad de personas en cada categoría de peso y obesidad. Las categorías incluyen desde **bajo peso** hasta diferentes niveles de sobrepeso y obesidad.

**DATOS CLAVE OBSERVADOS:**

- **Mayor prevalencia:** La categoría más numerosa es **Obesity_Type_III**, lo que indica un número importante de personas con obesidad severa.
- **Segunda más común:** El peso normal ocupa también un lugar alto, similar a las categorías **Obesity_Type_I** y **Obesity_Type_II**.
- **Sobrepeso moderado: Overweight_Level_I** es menos frecuente que **Overweight_Level_II**, lo que podría indicar que quienes tienen sobrepeso tienden a avanzar a niveles más altos si no hay intervención.
- **Menor prevalencia: Insufficient_Weight** (bajo peso) es la categoría con menos personas, aunque no es despreciable.

**INTERPRETACIÓN:**

- Existe un **problema importante de obesidad**, ya que las tres categorías de obesidad suman más casos que el peso normal.
- El hecho de que **Obesity_Type_III** (la más grave) sea la categoría dominante sugiere que muchas personas no están recibiendo atención oportuna para frenar el aumento de peso.
- A pesar de esto, todavía hay un grupo considerable con peso normal, lo que indica que **la prevención es posible y necesaria** para que no migren a categorías de riesgo.
- El **bajo peso** es poco común, pero sigue presente y también representa un riesgo para la salud.

**Mensaje;** El gráfico evidencia que el sobrepeso y la obesidad son más comunes que el peso saludable, siendo la obesidad grave el caso más frecuente. Esto subraya la importancia de fomentar hábitos saludables, actividad física y controles médicos preventivos para reducir el riesgo de enfermedades asociadas.

## 9. Conclusiones

- La **obesidad grave (tipo III) es muy común** en los datos analizados, lo que muestra un problema serio de salud que necesita atención urgente.
- Peso**, altura y edad son claves para predecir la obesidad**, pero siempre deben analizarse junto con los hábitos y el entorno social.
- Hacer **ejercicio y tomar suficiente agua ayudan a prevenir la obesidad**, aunque su efecto depende de la constancia.
- El **tiempo frente a pantallas no se relaciona directamente con el peso**, pero puede influir en problemas como el sedentarismo y el estrés.
- Las **personas con más peso usan menos transporte activo (caminar o bicicleta)**, lo que refleja barreras físicas y sociales que deben superarse.
- No **hay una única causa de la obesidad**: influyen varios factores (físicos, de hábitos y sociales) al mismo tiempo.
- Se **recomienda promover la educación en salud y el uso de entornos y herramientas digitales** que ayuden a las personas a cuidar sus hábitos.
- El **uso de Inteligencia Artificial (IA) en salud debe ser ético, transparente y participativo**, para que sea útil y aceptado por la comunidad.

## 9.1 Conclusión final

La obesidad y las enfermedades cardiovasculares son problemas de salud complejos que no dependen de un solo factor, sino de la interacción entre características físicas, hábitos de vida y condiciones sociales. El análisis realizado muestra que el peso, la edad y la altura son variables claves, pero deben interpretarse junto con la actividad física, la alimentación y el entorno en que viven las personas.

Se confirma que los hábitos saludables, como la práctica regular de ejercicio y una adecuada hidratación, son factores protectores, mientras que el sedentarismo y la falta de movilidad activa aumentan el riesgo. Sin embargo, no existe un patrón único que explique la obesidad, lo que refuerza la necesidad de un enfoque integral y multifactorial.

Finalmente, la inteligencia artificial demuestra ser una herramienta valiosa para predecir riesgos y apoyar la toma de decisiones en salud, siempre que se use con ética, transparencia y participación comunitaria.

## Referencias


**Sobre el dataset utilizado**

- Aravindpcoder. (2022). *Obesity or CVD Risk – Classify/Regressor/Cluster* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster

**Obesidad y salud pública**

- World Health Organization (WHO). (2023). *Obesity and overweight*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

**Factores de riesgo en obesidad y enfermedades cardiovasculares**

- GBD 2019 Risk Factors Collaborators. (2020). *Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis*. The Lancet, 396(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2

Uso **de IA en salud**

- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). *A guide to deep learning in healthcare*. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

**Ciclo de vida de proyectos de Machine Learning**

- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

**Métodos de análisis de datos en salud**

- Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.


**Uso de inteligencia artificial IA**

- OpenAI. (2025). ChatGPT (versión GPT-5) [Modelo de lenguaje]. https://chat.openai.com/
- DeepSeek AI. (2025). DeepSeek [Plataforma de inteligencia artificial]. https://deepseek.com/
- Google. (2025). Gemini [Modelo de lenguaje]. https://gemini.google.com/
- GitHub. (2025). GitHub Copilot [Asistente de programación basado en IA]. https://github.com/features/copilot
- InVideo Inc. (2025). *InVideo — AI video generator* [Plataforma web]. Recuperado de https://invideo.io/