Estudio de las Medidas Antropométricas para determinar el riesgo de obesidad y ENT

> Autora: Diana Chacón Ocariz

Contexto:

Los datos antropométricos se refieren a las medidas y características físicas de los seres humanos. Estos datos se obtienen mediante la medición de diferentes dimensiones corporales, como la estatura, el peso, la longitud de extremidades, el diámetro de huesos, el perímetro de la cabeza, etc.

Estos datos pueden ser utilizados para diversos fines, como el estudio de la población, el diseño y la ergonomía, la moda y la confección de prendas de vestir, la evaluación de la salud, el rendimiento físico, etc.

Este proyecto, se limita al estudio de un subconjunto de medidas para calcular varios indicadores que pueden ayudar a determinar el riesgo de padecer obesidad o enfermedades no transmisibles (ENT) como hipertensión, enfermedades cardiovasculares, diabetes, etc.

Más adelante, espero poder llevar a cabo otros estudios utilizando el mismo conjunto de datos, con el objetivo de demostrar que es posible crear procesos genéricos de ingeniería de datos sobre un mismo conjunto de datos, y que éstos pueden ser utilizados en la resolución de diversos problemas, ya sean de análisis o de ciencia de datos.

Estudio del sobrepeso y la obesidad y su incidencia en las ENT:

La obesidad y el sobrepeso son condiciones de salud que han alcanzado proporciones epidémicas a nivel mundial, siendo un factor determinante en el desarrollo de diversas enfermedades no transmisibles (ENT). Estas condiciones se caracterizan por un exceso de grasa corporal, lo que puede tener consecuencias significativas para la salud a largo plazo.

El sobrepeso y la obesidad aumentan sustancialmente el riesgo de padecer enfermedades crónicas, como la diabetes tipo 2, la hipertensión y las enfermedades cardiovasculares. Estas afecciones pueden tener un impacto negativo en la calidad de vida, aumentando la morbilidad y la mortalidad en las poblaciones afectadas.

Para evaluar y cuantificar el grado de sobrepeso y obesidad, se utilizan medidas antropométricas clave:

1. **Índice de Masa Corporal (IMC):** Relaciona el peso y la altura de una persona. Sin embargo, es importante señalar que el IMC no distingue entre masa muscular y grasa, por lo que puede

haber limitaciones en su interpretación, especialmente en atletas o personas con una distribución de grasa atípica.

2. **Relación cintura-cadera (RCC) y la relación cintura-talla (RCT):** Estos indicadores ofrecen una perspectiva más completa sobre la distribución de la grasa corporal. Una acumulación excesiva de grasa en la región abdominal, medida mediante el **contorno de cintura**, está particularmente asociada con un mayor riesgo de enfermedades metabólicas y cardiovasculares. Un contorno de cintura elevado puede indicar la presencia de grasa visceral, que se asocia con inflamación y otros procesos patológicos.

3. **La edad y el sexo:** A medida que las personas envejecen, es común experimentar cambios en el metabolismo y en la composición corporal, lo que puede influir en la tendencia a ganar peso. Además, existe evidencia de que las tasas de obesidad varían según el género, con diferencias en la distribución de grasa y en las respuestas hormonales.

Medidas como el IMC, RCC y RCT no se encuentran en los datos utilizados para el estudio. Sin embargo, pueden ser calculados a partir del peso, la estatura, el contorno de cintura y el de cadera con las siguientes fórmulas:

$$IMC = Peso(Kg)/Altura(m)^2$$

$$RCC = Cintura(cm)/Cadera(cm)$$

$$RCT = Cintura(cm)/Altura(cm)$$

Luego, existen tablas que permiten clasificar el grado de sobrepeso y obesidad y el riesgo de tener exceso de grasa abdominal, en función de los valores obtenidos:

1. Clasificación de obesidad según el BMI:

Riesgo	BMI
Peso bajo	< 18.5
Peso normal	18.5 - 25
Sobrepeso	25 - 30
Obesidad	> 30

1. Clasificación de riesgo de grasa abdominal según el contorno de cintura (CC)

Riesgo	Femenino	Masculino
Bajo o nulo	< 80 cm	< 94 cm
ALto	> 80 cm	> 94 cm

1. Clasificación de riesgo de de grasa abdominal según el racio circunferencia de cintura y circunferencia de cadera (RCC)

Riesgo	Femenino	Masculino
Bajo	< 0.8	< 0.95

Riesgo	Femenino	Masculino
Medio	0.81 - 0.85	0.96 - 1
Alto	> 0.86	> 1

1. Clasificación de riesgo sobrepeso y riesgo de enfermedades no transmisibles según racio circunferencia cintura y talla (ICT):

Riesgo	Femenino	Masculino
Delgado	< 0.41	< 0.42
Sano	0.41 - 0.48	0.42 - 0.52
Sobrepeso	0.48 - 0.57	0.52 - 0.62
Obesidad	> 0.57	> 0.62

Sin embargo, no existe una fórmula que combine todos estos datos y nos permita saber de manera objetiva el riesgo que tiene una persona de padecer obesidad y ENT.

Definición del problema y objetivo del estudio:

Determinar, a partir de ciertas antropométricas, el grado de riesgo que corre una persona de padecer obesidad y/o ENT.

Crear una aplicación que pueda ser utilizada por cualquier persona y que mediante la introducción de datos sencillos de obtener, pueda darle una predicción de su grado de riesgo de padecer sobrepeso/obesidad y ENT.

Se trata de determinar la variable **obesity** que indica 3 grados de riesgo:

1. Riesgo bajo o nulo (0)
2. Riesgo medio (1)
3. Riesgo alto (2)

a partir de la edad, el género, el peso, la estatura, el contorno de cintura y el contorno de cadera de una persona.

La variable objetivo no se encuentra dentro del conjunto de datos.

Con los datos existentes, es posible calcular los indicadores definidos más arriba y que permiten determinar el grado de riesgo. Sin embargo, no existe una fórmula o algún criterio objetivo que permita calcular el riesgo tomando en cuenta todas estas medidas.

El objetivo es usar modelos de ML para:

1. Identificar grupos de personas con características similares (clasificación no supervisada) y poderlas etiquetar (asignación manual de la variable objetivo).
2. A partir de los datos etiquetados, entrenar un modelo de clasificación supervisada que permita predecir el grado de riesgo.

Fuente de los datos:

Fuente principal:

Existen muchos datasets de datos antropométricos. Sin embargo, este me pareció el más pertinente por la cantidad y variedad de la información:

https://www.kaggle.com/datasets/thedevastator/3-d-anthropometry-measurements-of-human-body-sur?select=caesar.csv

Créditos: Andy R. Terrel: https://data.world/andy

Subconjunto de datos utilizados en el estudio:

- **age:** Edad
- **age_range:** Rango de edad. Variable categórica
- **gender:** Sexo (male, female). Variable categórica
- **height:** Altura (en pulgadas)
- **hip_circum:** Contorno de cadera (en pulgadas)
- **weight:** Peso (en libras)
- **waist_circum_preferred:** Contorno de cintura (en pulgadas)

Aquí una descripción completa del dataset: Metadatos

Enfoque técnico para la realización del estudio:

Pasos generales que seguí para llevar a cabo el estudio:

- Buscar y seleccionar un subconjunto de datos o medidas antropométricas pertinentes para el estudio.
- Calcular los indicadores que permiten determinar el riesgo de padecer obecidad o ENT.
- Utilizar modelos de clasificación no supervisada para encontrar grupos y así etiquetar los datos.
- Seleccionar el mejor modelo de ML para predecir el riesgo de que una persona pueda sufrir de obesidad o ENT - Implementar un prototipo de aplicación que permita a cualquier persona, conocer el grado de riesgo de padecer obesidad o ENT.

Resumen del Estudio y Notebooks:

Repostitorio GitHub de los Notebooks.

1. NB1: Lectura, limpieza, transformación y cálculo de nuevas variables:

- Carga, limpieza y transformación de los datos.
- Cálculo de nuevas variables necesarias al estudio.
- Almacenamiento de los datos en archivo .parquet para el resto del estudio.

Resultado: Función **load_clean_transform** con 3 workflows:

- **Extracción:** Cambia según la fuente de los datos
  - Lectura del archivo
  - Limpieza o eliminación de datos incorrectos
  - Selección del subconjunto de datos
- **Transformación:**
  - Transformación de los datos del sistema inglés al sistema métrico: Función **custom_transformations** Puede ser aplicada a todo el dataset o a un subconjunto. Se aplica sólo si es necesario
  - Cálculo de nuevas variables: Función **custom_calculations**. Indispensable para continuar el estudio
- **Salvaguarda de los datos (load):** En un archivo parquet. Más adelante debería ser en una base de datos.

1. NB2: EDA y Visualización de los Datos: Exploración y análisis de los datos:

- Estudio de la forma de los datos y la correlación entre ellos.
- Identificación de patrones en los datos.

1. NB3: Clasificación no supervisada y etiquetado de los datos: Clasificación y etiquetado de los datos. Análisis de los clusters. Análisis de la etiqueta.

La variable objetivo no existe dentro del conjunto de datos. Tampoco existe una fórmula que me permita calcularla.

Una solución es utilizar un algoritmo de clasificación no supervisada para identificar los grupos de datos similares y así poder etiquetarlos más fácilmente.

Consideré el algoritmo **KModes** que se basa en encontrar grupos o clusters en un conjunto de datos utilizando la distancia entre las observaciones y los modos de las categorías.

Utilicé este algoritmo con las variables categóricas siguientes: age_range, obesity_bmi, obesity_cc, obesity_rcc, obesity_ict y risk_factors

Como lo que busco es ayudarme a etiquetar los datos, escogí el algoritmo de clasificación que me generara menor cantidad de clusters y datos más homogéneos dentro de cada grupo.

Por esta razón, escogí **KModes** para definir las etiquetas (con KMeans obtuve resultados menos óptimos).

Estudé cluster por cluster para definir las etiquetas de cada uno.

Este proceso es de lejos, el que más me tomó tiempo durante el estudio.

Resultado: Función **label_data**** con 3 workflows:

- Lectura de los datos
- Clasificación y generación de los clusters
- Etiquetado de los datos
- Salvagurada de los datos etiquetados

Nota importante: Este proceso de etiquetado debería ser hecho o estar supervisado por personal médico o científico experto en el tema.

1. NB4: Entrenamiento del modelo: Escogencia del modelo. Ajuste del modelo. Salvaguarda del modelo entrenado. Predicciones.

Variable objetivo: Obesity:

- **0** -> Riesgo nulo o bajo
- **1** -> Riesgo medio
- **2** -> Riesgo alto

Se trata de un problema de clasificación supervisada.

1. Primero evalué los modelos: regresión logística, árboles de decisión, vector support machines y ramdon forest; utilizando los datos originales del dataset (no los calculados).
2. Escogí los dos mejores modelos: árboles de decisión y ramdon forest.
3. Los comparé con diferentes subconjuntos de datos.
4. Seleccioné el modelo y el subconjunto de datos definitivos.
5. Construí 2 workflows:

- **Workflow para entrenar el modelo con todos los datos y guardarlo:** Función **save_model**
- **Workflow para hacer la predicción para una persona:** Función **make_obesity_prediction**

Tareas:

1. Cargar los datos
2. Selección y separación del dataset
3. Transformación de los datos (si necesario)
4. Entrenar y evaluar diferentes modelos de clasificación
5. Escoger el(los) mejor(es) modelo(s)
6. Ajustar y mejorar el modelo escogido
7. Contruir los workflows

Repostitorio GitHub

Implementación del prototipo de la aplicación:

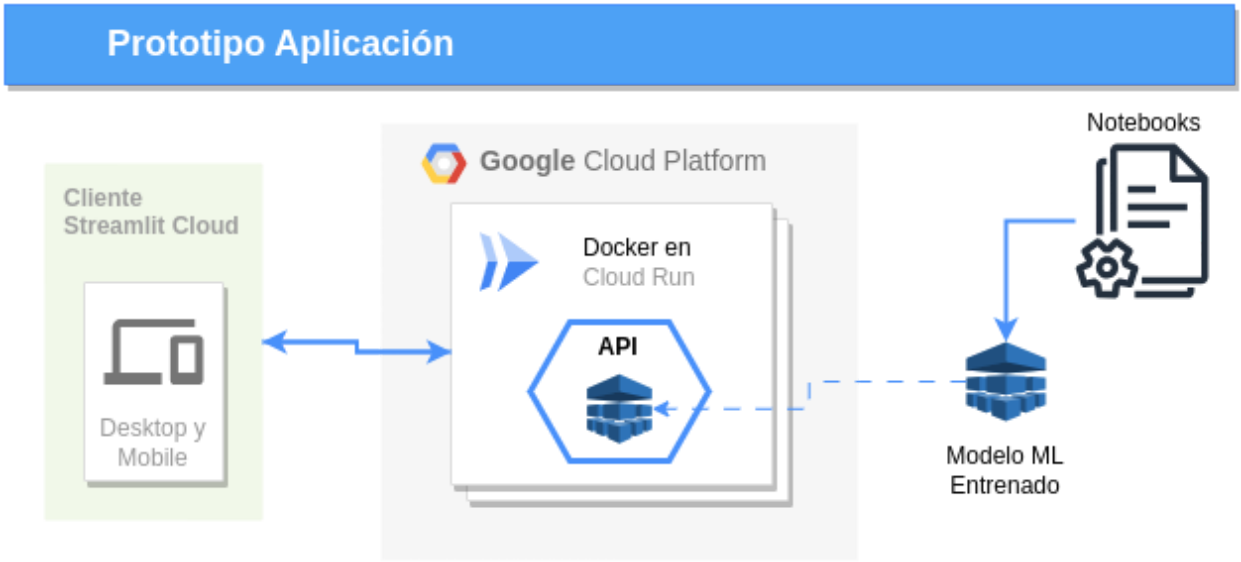
A partir de los datos que el usuario introduce, la aplicación **dIAna** (hecha utilizando Streamlit) consulta una API (desarrollada con FastAPI) que carga el modelo de ML entrenado, y devuelve:

1. Indicadores y sus índices de riesgo: Índice de masa corporal, contorno de cintura, racio entre cintura y cadera, racio entre cintura y estatura.
2. Índice de riesgo predicho por el modelo: Riesgo bajo o nulo, riesgo medio, riesgo alto.

Probar la aplicación dIAna antropometría y obesidad

Repostitorio GitHub

Nota: Los resultados NO deben tomarse como la opinión de un especialista. Esto es un simple ejercicio de ciencia de datos.



Limitaciones y dificultades:

1. **Etiquetado de los datos:** Lo que más se me dificultó fue el proceso de etiquetado de los datos, ya que la variable objetivo no existía dentro del conjunto de datos y tampoco existe una fórmula para calcularla. Para que el estudio pueda tener algún valor científico, este proceso de etiquetado debería ser realizado por algún especialista.

2. **Implementación del prototipo:** También se me dificultó la implementación del prototipo, el uso de docker y la instalación de la API en la nube, por falta de experiencia.

Mejoras a futuro:

Propuesta de solución a futuro:

1. Promover el uso de la aplicación por parte de personal médico para poder integrar nuevos datos y obtener una clasificación más exacta. Los profesionales podrían ayudar a validar las etiquetas asignadas por el modelo de ML, agregar nuevos datos de personas, agregar datos suplementarios, etc.
2. Guardar los datos en una base de datos para incluir datos de diferentes fuentes.

3. Entrenar de nuevo los modelos con todos los datos (actuales y recolectados a través del uso de la aplicación).
4. Usar una herramienta para registrar y monitorear los modelos.
5. Crear pipelines para todos por procesos.
6. Automatizar todo el proceso.
7. Corregir todos los problemas y deficiencias que por falta de tiempo, no pudieron resolverse.

Instrucciones para la ejecución del proyecto:

Ejecutar el procesamiento de los datos:

Ir al Repostitorio GitHub de los Notebooks. Ejecutar los workflows definidos al final de los notebooks en el siguiente orden:

**Para limpiar, transformar, etiquetar los datos y entrenar el modelo:

1. **NB-01:** Función **load_clean_transform**
2. **NB-03:** Función **label_data**
3. **NB-04:** Función **save_model**

Para hacer predicciones:

1. **NB-04:** Función **make_obesity_prediction**

Ejecutar el código de la app:

Ir al Repostitorio GitHub de la aplicación y clonarlo:

1. **Ejecutar la API:** Ir al directorio `api` y ejecutar el comando uvicorn `api:app`. Esto ejecutará la API en local en el puerto 8000
2. **Ejecutar la aplicación Streamlit:** Ejecutar el comando `streamlit run streamlit/Inicio.py`. Esto ejecutará la App en local en el puerto 8501
3. Verificar en el archivo `streamlit/Inicio.py` que se está apuntando a la API en local.

También es posible probar la implementación de la aplicación:

Probar la aplicación dIAna antropometría y obesidad

Conclusiones:

1. Muchos problemas pueden ser resueltos con datos. Sin embargo, es difícil encontrar el conjunto de datos que pueda contestar exactamente las preguntas que nos planteamos.
2. En la mayoría de los casos, debemos completar la información, ya sea calculando nuevas variables a partir de las que tenemos, utilizando herramientas estadísticas o incluso algoritmos

de ML para obtener los datos que nos hacen falta.

3. En estudiar y explorar los datos, se nos puede ir la vida... Siempre encontraremos algo nuevo jejeje

4. Estos preliminares de estudiar los datos y completarlos, es lo que nos lleva más tiempo... ¡Siempre!

5. Si bien es importante tener un conocimiento general de todos los procesos que son necesarios para llevar a cabo un proyecto, desde la recolección de los datos hasta la implementación de una solución que pueda ser utilizada y mantenida fácilmente, es evidente que se necesita un equipo multidisciplinario, para que cada etapa pueda ser llevada a cabo por un especialista.

6. A medida que vamos avanzando en el estudio y conociendo los datos, van surgiendo nuevas preguntas e ideas de lo que podemos hacer con ellos.

Bibliografía

- Índice cintura-altura
- ÍNDICE CINTURA-CADERA
- La medición del cálculo cintura/cadera te ayudarán a saber si tienes sobrepeso
- Metodologías diferentes para medir la Composición Corporal
- OMS: Obesidad y sobrepeso
- Waist circumference action levels in the identification of cardiovascular risk factors: prevalence study in a random sample

In [ ]: