Natalia Parrado
Edwin Piñeros
Jose Suarez

Universidad de los Andes

Estimación de Niveles de Obesidad en América Latina: Un Enfoque de Aprendizaje No Supervisado

1. Resumen

El objetivo de este proyecto es desarrollar un modelo de aprendizaje no supervisado para estimar los niveles de obesidad en personas de México, Perú y Colombia, utilizando datos de hábitos alimenticios y condición física. El desafío consiste en identificar patrones y tendencias para clasificar a las personas en categorías de obesidad definidas por la OMS.
Este modelo se podría aplicar en la creación de herramientas de software que detecten tempranamente niveles de obesidad y sirvan como un insumo para generar sistemas de recomendación personalizados para mejorar la salud de las personas.
Para el desarrollo se utilizará un conjunto de datos que combina datos reales y datos sintéticos, se utilizan 77% datos generados y 23% de los datos recolectados directamente de usuarios a través de una encuesta en una plataforma web.
Se implementaron tres algoritmos de clusterización, se encuentra que DBScan es el método con mejores resultados por calidad de los clusters y coherencia con la marcación de niveles de obesidad que contiene el dataset. Sin embargo, como veremos en el desarrollo es necesario profundizar en la recolección de datos adicionales y el perfeccionamiento de los métodos para encontrar mejores resultados que apoyen el desarrollo de sistemas de recomendación personalizados.

2. Introducción

La obesidad es una de las principales preocupaciones de salud pública a nivel global, con graves problemas para la salud individual y para los sistemas de salud social. Según la Organización Mundial de la Salud (OMS, 2023)[1], la obesidad ha llegado a ser considerada como una enfermedad epidémica, que ha afectado a más de 650 millones de adultos a nivel mundial. Este tipo de condición no solo aumenta el riesgo de enfermedades graves como la diabetes tipo 2 y las enfermedades cardiovasculares, sino que también tiene un impacto significativo en la calidad de vida y en los costos de atención médica tanto para el individuo como la carga económica al sistema de salud social.

Por ello se considera que analizar este problema y tratar de encontrar un método de solución o al menos reducir los índices de obesidad a nivel global, es fundamental el poder clasificar e identificar a las personas en diferentes niveles de obesidad como lo son Peso bajo (o desnutrición), peso normal, Sobrepeso nivel 1, Sobrepeso nivel 2, Obesidad tipo 1, Obesidad tipo 2 y Obesidad tipo 3. Cada uno de estos niveles tiene características distintas y requiere planes de tratamiento y prevención específicos. La identificación precisa de estos niveles permite a los profesionales de la salud proporcionar tratamientos personalizados y desarrollar estrategias de prevención efectivas al encontrar los patrones dentro de los datos de los pacientes.

Entidades como la Organización Panamericana de la Salud (OPS, 2022)[2], también manifiestan la importancia de una clasificación adecuada de pacientes e identificación de las posibles causas de manera adecuada para diseñar políticas de salud pública que puedan enfrentar la obesidad de manera más efectiva. Por ello, la implementación de algoritmos de aprendizaje no supervisado podría permitir la identificación de patrones de los pacientes que

contienen cierto niveles de obesidad tratando de identificar las posibles causas o las condiciones de las personas que puedan llegar ser más propensas a tener dichas condiciones. Una solución como esta, permitirá a organizaciones de la salud plantear campañas de prevención y también establecer el nivel de obesidad de la persona con la finalidad de prestar los mejores tratamientos para evitar que el problema de salud siga o se evite complicaciones.

Nuestro proyecto se posiciona como una extensión de estos estudios al utilizar un enfoque de aprendizaje no supervisado, aplicando técnicas de clusterización para estimar los niveles de obesidad sin necesidad de una clasificación previa.

3. Materiales y Métodos

Los datos utilizados en este análisis provienen del dataset titulado "Estimation of Obesity Levels Based on Eating Habits and Physical Condition", disponible en el UCI Machine Learning Repository. Este conjunto de datos contiene información sobre los niveles de obesidad de individuos en México, Perú y Colombia..
El dataset consta de 2111 registros y 17 variables, de las cuales 16 son características del individuo que lo clasifican en un nivel de obesidad registrado en la variable "*NObeyesdad*" A continuación se describen las variables:

Variables Categóricas:

- Gender: Género del individuo (Masculino/Femenino).
- family_history_with_overweight: Indica si el individuo tiene antecedentes familiares de sobrepeso (Sí/No).
- FAVC: Indica si el individuo consume alimentos altos en calorías frecuentemente (Sí/No).
- CAEC: Indica si el individuo consume alimentos entre comidas (No/Sí).
- SMOKE: Indica si el individuo fuma (Sí/No).
- SCC: Indica si el individuo monitorea las calorías que consume diariamente (Sí/No).
- CALC: Frecuencia con la que el individuo consume alcohol (Nunca/Rara vez/Frecuentemente/Siempre).
- MTRANS: Medio de transporte que el individuo utiliza normalmente (Automóvil/Moto/Bicicleta/A pie/Transporte público).
- NObeyesdad: Nivel de obesidad, categorizado en 'Insufficient_Weight', 'Normal_Weight', 'Overweight_Level_I', 'Overweight_Level_II', 'Obesity_Type_I', 'Obesity_Type_II' y 'Obesity_Type_III'.

Variables Continuas:

- Age: Edad del individuo (en años).
- Height: Altura del individuo (en metros).
- Weight: Peso del individuo (en kilogramos).
- NCP: Número de comidas principales que el individuo consume diariamente.
- CH2O: Cantidad de agua que el individuo consume diariamente (en litros).
- FAF: Frecuencia de actividad física del individuo (en días por semana).
- TUE: Tiempo de uso de dispositivos tecnológicos por día (en horas).

Los estadísticos descriptivos de las variables continuas se presentan en la Tabla 1 :

Tabla 1. Estadísticos descriptivos del conjunto de datos

Variable	Media	Desviación Estándar	Mínimo	25% Percentil	50% Percentil	75% Percentil	Máximo
Age	24,31	6,35	14,00	19,95	22,78	26,00	61,00
Height	1,70	0,09	1,45	1,63	1,70	1,77	1,98
Weight	86,59	26,19	39,00	65,47	83,00	107,43	173,00
FCVC	2,42	0,53	1,00	2,00	2,39	3,00	3,00
NCP	2,69	0,78	1,00	2,66	3,00	3,00	4,00
CH2O	2,01	0,61	1,00	1,58	2,00	2,48	3,00
FAF	1,01	0,85	0,00	0,12	1,00	1,67	3,00
TUE	0,66	0,61	0,00	0,00	0,63	1,00	2,00

A través de histogramas y diagramas de caja, se puede observar la distribución de las variables continuas, estas herramientas permiten identificar la presencia de valores atípicos y entender la dispersión de los datos. A Continuación se presentan los resultados obtenidos.

Figura 1. Distribución peso

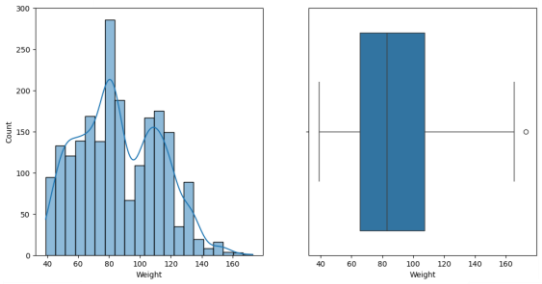


Figura 2. Distribución altura

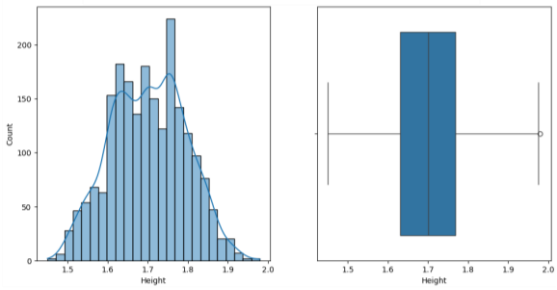


Figura 3. Distribución edad

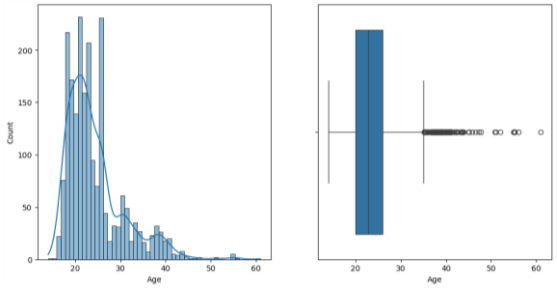


Figura 4. Distribución FCVC

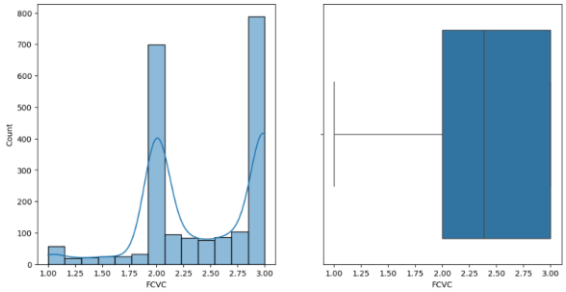


Figura 5. Distribución NCP

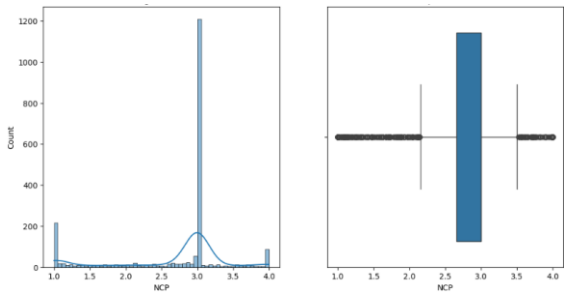


Figura 6. Distribución CH20

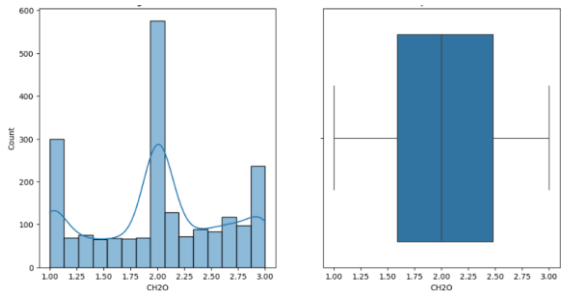


Figura 7. Distribución FAF

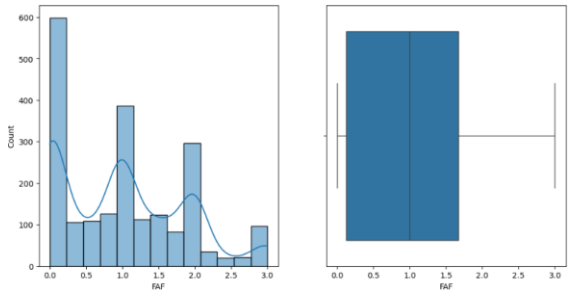
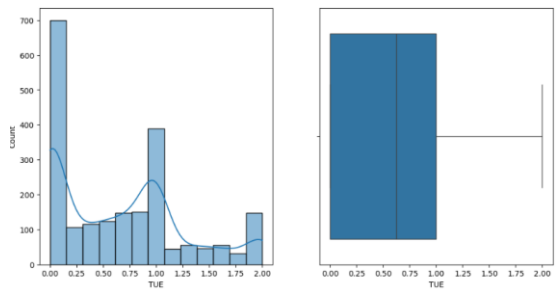


Figura 8. Distribución TUE



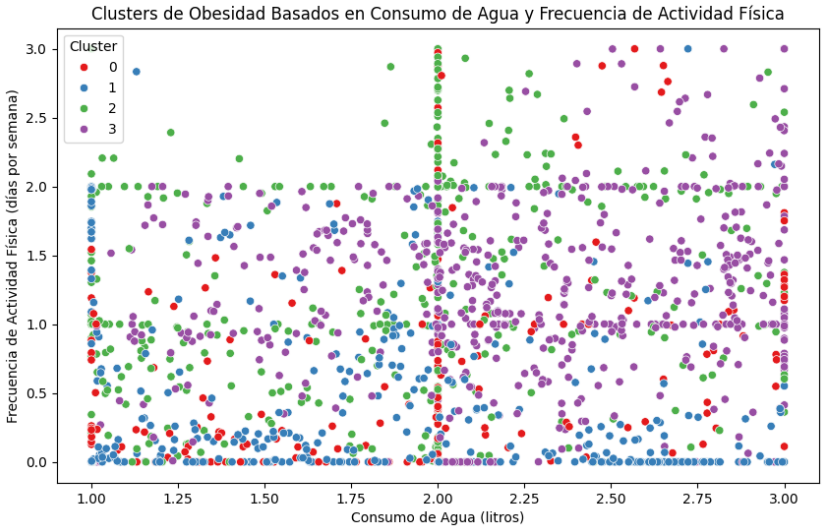
Dado que el enfoque del análisis es realizar un **aprendizaje no supervisado** para clasificar los clústeres según el nivel de obesidad, se realizaron las siguientes acciones para el procesamiento de los datos:

- Selección de Variables
- Conversión de variables categóricas con One-Hot-Encoding.
- Estandarización de las variables numéricas

Durante el proceso de clustering, las variables numéricas ayudarán a definir la similitud entre los individuos, mientras que las variables categóricas proporcionarán información adicional para interpretar los clústeres resultantes. La interpretación de los clústeres se basará en la comparación de las características dentro de cada clúster, buscando correlaciones implícitas con los niveles de obesidad previamente definidos.
Finalmente, aunque NObeyesdad no se utilizará directamente en el proceso de clustering, se evaluará la coherencia de los clústeres formados comparándolos con esta variable para entender cómo se alinean los clústeres con los niveles de obesidad establecidos. Este análisis permitirá validar si los patrones descubiertos son coherentes con el conocimiento previo sobre la obesidad o si revelan nuevas perspectivas.

4. Resultados y Discusión

Se implementó k-means con bajos resultados en la calidad de los clusters al mostrar superposición de varios de ellos, se atribuye a que las variables de hábitos utilizadas se distribuyen de manera similar en todos los grupos poblacionales.



Los resultados obtenidos muestran que el algoritmo DBSCAN proporcionó los mejores resultados en cuanto a la calidad de los clusters y su coherencia con los niveles de obesidad establecidos por la OMS. Este enfoque fue capaz de identificar agrupaciones significativas basadas en los hábitos alimenticios y la actividad física de los individuos, permitiendo una clasificación más precisa en los distintos niveles de obesidad. Sin embargo, uno de los principales desafíos identificados es la limitada cantidad de datos reales disponibles, ya que un 77% de los datos fueron generados sintéticamente mediante SMOTE, lo que podría haber introducido ciertos sesgos en los resultados. La falta de más datos recolectados directamente

de los usuarios puede haber limitado la capacidad del modelo para capturar todas las variaciones en los patrones de comportamiento asociados con la obesidad

5. Conclusiones

El aplicar los tres algoritmos nos dio la ventaja de poder observar y contrastar diferentes enfoques para producir las agrupaciones e implícitamente intentar identificar patrones sobre estos mismos. Cabe destacar que los resultados no fueron los mejores, pero se entiende que la mayor razón de estos es por la naturaleza que tiene los datos recopilados ya que más del 505 de estos datos son sintéticos y producidos de manera sub-muestreada lo cual podría implicar que se solapen ciertas propiedades o características de cada agrupación.

Como resultado de este trabajo es la motivación de continuar indagando sobre la generación de proyectos que permitan extraer conocimiento y más aún enfocado en la parte de salud, no solo con la finalidad de identificar enfermedades sino con la finalidad de generar un prevención. Finalmente, es importante tener en cuenta que proyectos como estos deben estar acompañados de especialistas en el área que permitan corroborar los resultados.

6. Bibliografía

1. **Estimation of Obesity Levels Based On Eating Habits and Physical Condition .** (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.
2. **Organización Mundial de la Salud.** (2023). *Obesidad y sobrepeso*. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight
3. **Organización Panamericana de la Salud.** (2022). *Obesidad: una preocupación de salud pública*. Prevención de la obesidad - OPS/OMS | Organización Panamericana de la Salud (paho.org)
4. **Gómez, M. A., & Henao, H.** (2020). Aplicación de técnicas de clustering para la detección de obesidad en datos de salud en Colombia. *Revista Colombiana de Estadística, 43*(2), 209-225. Las variables más influyentes en la obesidad: un análisis desde la minería de datos (scielo.cl)
5. **Agamez Julio, Wilmer Jesús.** (2022). Predicción de riesgos en salud, para personas con obesidad empleando técnicas de aprendizaje de máquinas. Predicción de riesgos en salud, para personas con obesidad empleando técnicas de aprendizaje de maquinas (unal.edu.co)