TRABAJO DE GRADO
Opción Seminario-Diplomado.

CARLOS ANDRES MONTEALEGRE MONROY

UNIVERIDAD UNIREMINGTON
Juan Carlos Briñez de León

Machine learning

FACULTAD DE INGENIERÍA
CALI, COLOMBIA
2024

SISTEMA DE PREVENCION Y ESTUDIO CONTRA LA DIABETES, UTILIZANDO ESTRATEGIAS DE MACHINE LEARNING

Corporación Universitaria Remington.
Ingenierías
Ingeniería en Sistemas

Estudiantes:
Carlos Andres Montealegre Monroy.

Tutor: Juan Carlos Briñez de León

Opción de Trabajo de grado Seminario-Diplomado.
2024.

Dedicatoria

A mi familia que me apoyado en estos años de estudios.

Agradecimientos

A Dios por llenarme de salud y Sabiduría.

Tabla de Contenidos

Resumen

Este trabajo de grado se enfoca en el análisis de datos de consumo de clientes del gimnasio MEGA GYM HIT, realizando un seguimiento a sus usuarios con el fin de enfocarse en mejorar los niveles de salud y prevención de enfermedades como la obesidad, propone un sistema de recomendación de prevención de obesidad basado en estrategias de clustering. El análisis de datos inicia con la recolección de información de transacciones, incluyendo características como la edad, genero de los usuarios, el peso y la altura, también se busca conocer los niveles de actividad física que realiza el usuario estos niveles tendrán un rango entre 1 y 4, el índice de masa corporal también es una variable que se va a levantar en el estudio para cada usuario esta medida se tomara por profesionales capacitados. Esta información se somete a un proceso de limpieza y normalización para asegurar la calidad y consistencia de los datos.

A partir de este análisis, se propone utilizar algoritmos de clustering como K-means y DBSCAN con el objetivo de agrupar a los usuarios en segmentos con patrones en edades, niveles de actividad física peso y altura. El proceso de clustering permite identificar grupos homogéneos de usuarios, facilitando la creación de recomendaciones personalizadas de evaluación los niveles de obesidad en el que se encuentra y como generar un plan de trabajo para remediar estas afecciones. Los clusters revelan insights sobre las necesidades de mejorar los hábitos saludables dentro de los usuarios del gimnasio, como alertas para generar de manera oportuna acciones de mejora.

La recomendación basada en clustering optimiza la segmentación y el agrupamiento , mejorando la servicio para usuarios que necesiten un seguimiento adicional para mejorar su salud . El modelo es evaluado mediante métricas como el silhouette score, que valida la cohesión de los clusters, y se ajusta iterativamente para refinar las recomendaciones. Los resultados demuestran que esta metodología ofrece recomendaciones más precisas y relevantes en comparación con enfoques tradicionales, mejorando la experiencia del usuario.

Palabras clave

Sistemas de recomendación, Segmentación de usuarios, Clustering, agrupamientos, obesidad.

1. Marco conceptual y contextual

1.1 Contexto:

1.1.1 **Sistemas de recomendación**.

Si bien los establecimientos deportivos como los gimnasios son espacios de dispersión, frecuentados por un gran grupo de población que busca mejorar de manera estética su físico, moldeando mediante trabajo de pesas y actividad cardio vascular, también este tipo de servicios se enfoca en mejorar la salud de otro gran grupo de personas que no cuentas con los hábitos necesarios para tener una salud optima, es por eso realizar un enfoque donde se pueda aplicar los sistemas de inteligencia artificial para generar modelos de aprendizaje no supervisados para que ayuden a crear alertas de prevención en población que frecuente establecimientos deportivos para ayudar a padecer de enfermedades de la salud.

1.2.2 **Descripción de caso de estudio.**

La obesidad es una enfermedad silenciosa la cual en los últimas décadas se ha evidenciado de manera más amplia en el mundo; esto no significa que se esté previniendo de manera efectiva y oportuna, si bien es un concepto del cual se ha tomado conciencia por los diferentes riesgos que generan para la salud de los seres humanos, se tienen estudios donde se evidencias cifras preocupante por el porcentaje de obesidad por porcentajes de humanos sigue siendo muy alto, 1 de cada 8 personas sufren de obesidad en el mundo, cifra que se a logrado mantener por varios años sin logran ver mejoras significativas, más preocupante es como esta enfermedad también se encuentra presente en menos de edad deteriorando la

calidad de vida de los infantes debido a los grandes riesgos que generan padecer de esta enfermedad.

El gimnasio MEGA HIT GYM busca implementar mediante un modelo de algoritmos de aprendizajes no supervisados, el análisis de información que permita ayudar a prevenir a los usuarios el padecimiento de obesidad o determinar su grado de obesidad.

1.3 **Pregunta problema:**

¿Cómo desarrollar un modelo computacional para identificar y prevenir el padecimiento de obesidad; evaluar los diferentes niveles de obesidad se encuentra mediante factores como el peso, la edad, el género, el índice de masa muscular y otros?

1.4 **Hipótesis:**

El análisis de datos de información tomada de una muestra en un grupo de personas nos permitirá poder evaluar, entender y mitigar las diferentes posibles causas del crecimiento exponencial de casos nuevo de obesidad y sus diferentes niveles, mediante el modelo, e información relacionada a los usuarios del gimnasio poder identificar posibles comportamientos o tendencias hacia la enfermedad.

2. Objetivos

2.1 Objetivo general.

Desarrollar un modelo computacional basado en técnicas de clustering para analizar y prevenir el padecimiento de obesidad en los usuarios del gimnasio MEGA GYM HIT, evaluando los diferentes niveles de obesidad a través de factores como el peso, la edad, el género, el índice de masa corporal y los niveles de actividad física, con el fin de generar recomendaciones personalizadas para la mejora de su salud.

2.2 Objetivos específicos.

- Recolección y limpieza de datos: Recolectar información relevante sobre los usuarios del gimnasio (edad, género, peso, altura, índice de masa corporal e información sobre niveles de actividad física), y aplicar procesos de limpieza y normalización de los datos para garantizar su calidad y consistencia.

- Segmentación de usuarios mediante clustering: Implementar algoritmos de clustering, como K-means y DBSCAN, para segmentar a los usuarios en grupos homogéneos con patrones similares en términos de características físicas y niveles de actividad, con el objetivo de identificar segmentos con mayor riesgo de padecer obesidad.

- Desarrollo de un sistema de recomendación personalizado: Utilizar los resultados del análisis de clustering para diseñar un sistema de recomendación que ofrezca alertas tempranas y recomendaciones personalizadas sobre prevención de obesidad, basadas en los grupos identificados, y sugiera planes de trabajo para mejorar la salud de los usuarios.

- Evaluación y mejora continua del modelo: Evaluar la eficacia del modelo mediante métricas de calidad como el silhouette score, ajustar los algoritmos iterativamente y validar las recomendaciones generadas, con el fin de optimizar la precisión y relevancia del sistema de prevención de obesidad en función de las necesidades de los usuarios del gimnasio.

3. Desarrollo e implementación del aprendizaje

Inicialmente se toma una muestra de información con los usuarios de las diferentes sedes en todo el país de la marca de gimnasio MEGA HIT GYM, no se tuvieron en cuenta factores socioeconómicos para la investigación, el estudio fue acompañado y consolidado con la información de los socios que dieron su consentimiento. Como variables importantes dentro de la investigación se tiene la Edad, el género, peso y estatura, entre otros, esta investigación busca conocer de acuerdo a la información recolectada, como los usuarios estaban viendo afectados por esta enfermedad, conocer que grupo de socios necesitan de un cuidado y un plan de entrenamiento especializado y como prevenir problemas en la salud en las muestras más vulnerables.

3.1 Preparación y análisis de los datos

Daremos un breve resumen de la información que se busca analizar, donde podremos encontrar las variables de entrada y la variable de salida.

3.1.1 Variables de entrada

- **Edad**
- **Genero**
- **Peso**
- **Altura**
- **Índice Masa corporal**
- **Nivel de actividad Física (valor numérico de 1 a 4)**

3.1.2 Variables de salida

- **Escala de obesidad**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Age                    1000 non-null   int64
 1   Gender                 1000 non-null   object
 2   Height                 1000 non-null   float64
 3   Weight                 1000 non-null   float64
 4   BMI                    1000 non-null   float64
 5   PhysicalActivityLevel  1000 non-null   int64
 6   ObesityCategory        1000 non-null   object
dtypes: float64(3), int64(2), object(2)
memory usage: 54.8+ KB
```

De acuerdo a la descripción y análisis realizado a la información suministrada, podemos encontrar información importante y determinante para el estudio de obesidad que se quiere realizar.

	Age	Height	Weight	BMI	PhysicalActivityLevel
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	49.857000	170.052417	71.205769	24.888317	2.534000
std	18.114267	10.309971	15.509849	6.193912	1.116284
min	18.000000	136.115719	26.065730	8.470572	1.000000
25%	35.000000	163.514205	61.129629	20.918068	2.000000
50%	50.000000	169.801665	71.929072	24.698647	3.000000
75%	66.000000	177.353596	81.133746	28.732132	4.000000
max	79.000000	201.419670	118.907366	50.791898	4.000000

- La **Edad** máxima encontrada en el estudio de 79 años, la mínima fue de 18 y el promedio de edad es de 49 años.
- La **Altura** Promedio fue de 170 y encontramos alturas superiores a 2 metros.
- El **BMI** promedio fue de 24 el mínimo de 8 y el máximo de 50.

Tabla de frecuencia

Tabla de Frecuencia:

	Intervalo	Frecuencia Absoluta	Frecuencia Acumulada
0	[66.8, 79.061)	236	236
1	[42.4, 54.6)	214	450
2	[18.0, 30.2)	191	641
3	[30.2, 42.4)	180	821
4	[54.6, 66.8)	179	1000

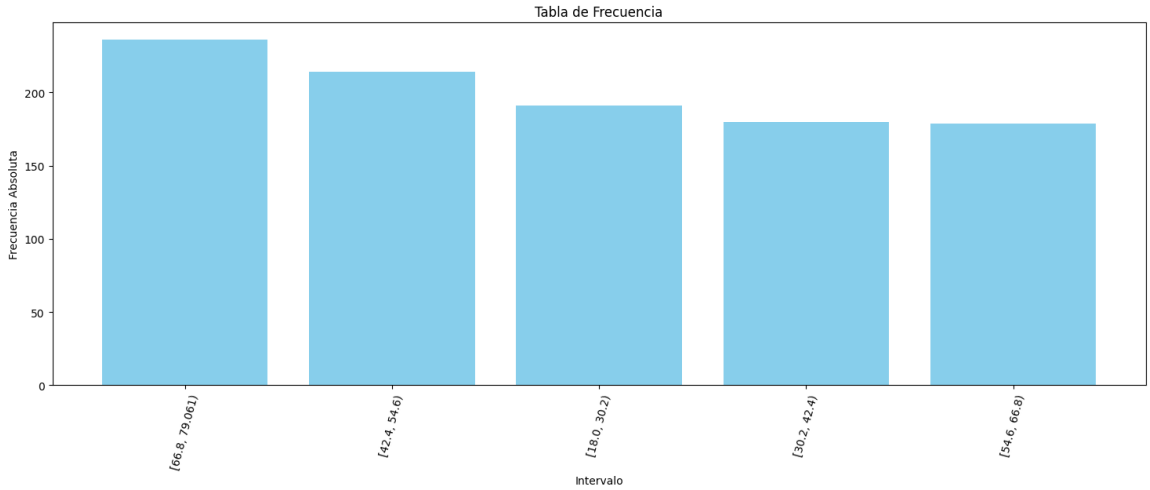

Gráfico 1. Distribución de frecuencia en usuarios según **EDAD**.

De acuerdo al grafico generado podemos encontrar, el rango de edad donde se encuentra el mayor grupo de usuarios es entre los 66 y 79 años, y los 2 grupos de rangos de edad entre 30 y 42, 54 y 66 comparten la misma frecuencia.

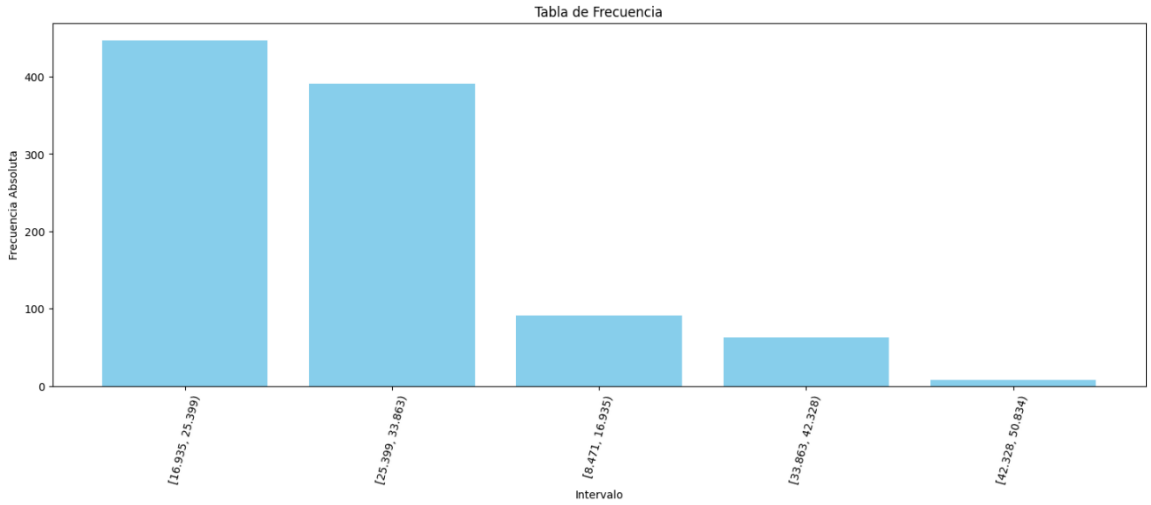

Gráfico 2. Distribución de frecuencia en usuarios según BMI.

De acuerdo al grafico generado podemos encontrar, el rango de índice de masa muscular de los usuarios son entre los 16 y 25 años, el menor grupo es del rango entre 42 y 50.

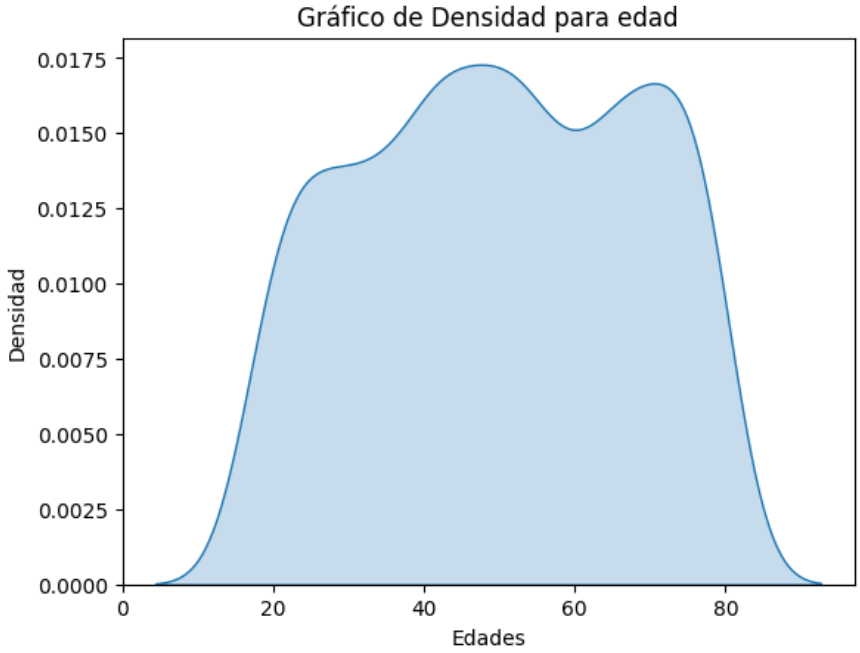

Gráfico 3. Gráfico de densidad por EDAD.

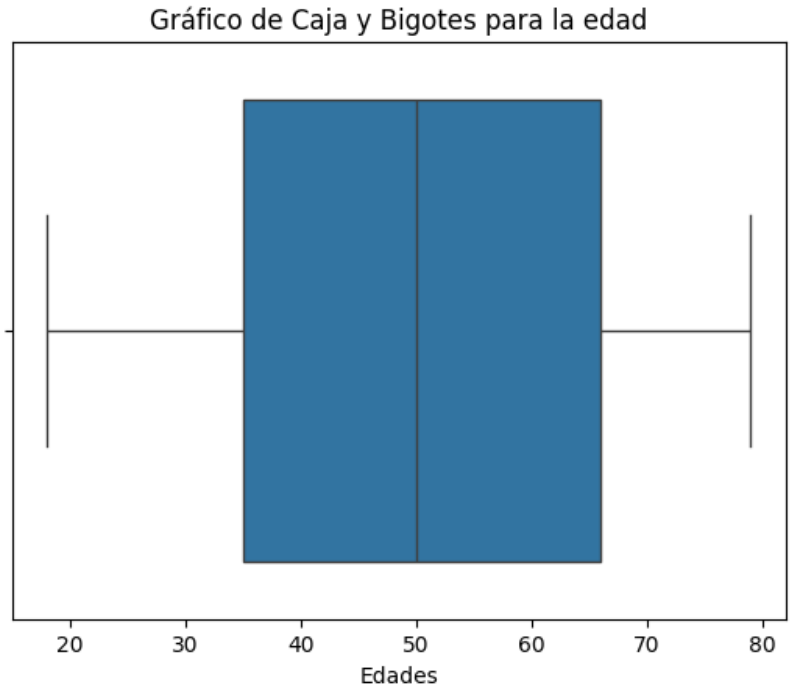

Gráfico 4. Gráfico caja y bigote por **EDAD**.

4. Modelo de toma de decisiones

De acuerdo a modelos de inteligencia artificial, previamente entrenados en modelamiento de secuencia y predicciones, se realizara mediante la información recolectada en el estudio del caso de uso planteado para el gimnasio MEGA HIT GYM, vamos a realizar un aplicativo que nos permita mediante la recolección de una captura manual o formulario, obtener información como la EDAD, PESO, ALTURA, BMI para predecir mediante esta información como se puede disminuir las posibilidad de generar algún tipo de obesidad en alguno de sus niveles.

Actualmente la base de datos que se utilizara para realizar el caso de uso es la siguiente:

De los datos vamos a dar valores numéricos a las columnas y valores que se representan con texto en este caso sería el Género y las categorías de la obesidad, esto dado que los modelos y herramientas que usaremos no los exige de esta forma. KNeighborsRegressor y MLPRegressor.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   1000 non-null   int64
 1   Gender                1000 non-null   object
 2   Height                1000 non-null   float64
 3   Weight                1000 non-null   float64
 4   BMI                   1000 non-null   float64
 5   PhysicalActivityLevel 1000 non-null   int64
 6   ObesityCategory       1000 non-null   object
dtypes: float64(3), int64(2), object(2)
memory usage: 54.8+ KB
```

Al valor **Genero** se le asignara el siguiente valor numérico:

1. Male
2. Female

Para Categorizar los niveles de obesidad asignaremos:

1. Normal weight
2. Obese
3. Overweight
4. Underweight

La nueva representación de los valores para nuestra información queda representada de la siguiente manera:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Age                    1000 non-null   int64
 1   Gender                 1000 non-null   int64
 2   Height                 1000 non-null   float64
 3   Weight                 1000 non-null   float64
 4   BMI                    1000 non-null   float64
 5   PhysicalActivityLevel  1000 non-null   int64
 6   ObesityCategory        1000 non-null   int64
dtypes: float64(3), int64(4)
memory usage: 54.8 KB
```

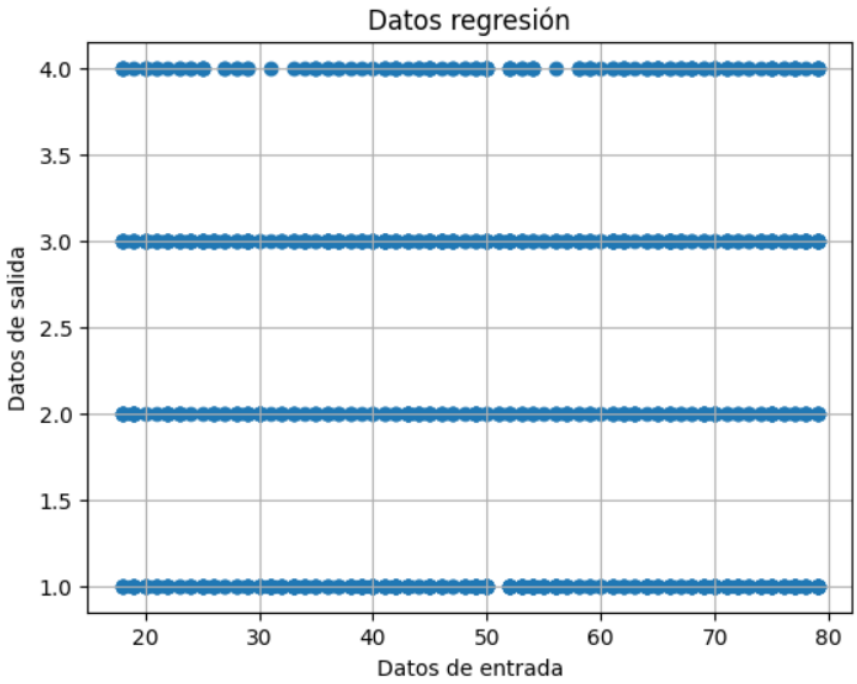

Gráfico 5. Tabla datos de regresión.

Usando la captura manual de información se realiza una predicción de la categoría de obesidad en la que se puede encontrar el usuario analizado, con estos datos podemos ir a la tabla de categorías y validar el nivel, con esta información poder realizar un plan de trabajo que permita que el usuario a prevenir y alejar de los niveles más graves de obesidad.

```
Nueva_entrada = np.zeros((1,6))
Nueva_entrada[0,0]=float(input('Ingrese su edad: '))
Nueva_entrada[0,1]=float(input('Ingrese su genero, para masculino 1 y femenino 2: '))
Nueva_entrada[0,2]=float(input('Ingrese su altura: '))
Nueva_entrada[0,3]=float(input('Ingrese su peso: '))
Nueva_entrada[0,4]=float(input('Ingrese su indice de masa corporal: '))
Nueva_entrada[0,5]=float(input('Ingrese su nivel de actividad fisica: '))

Proyeccion_1 = Modelo_1.predict(Nueva_entrada)
Proyeccion_2 = Modelo_2.predict(Nueva_entrada)

print('')
print('')
print('Según los datos ingresados, la proyección para definir la categoria de obesidad donde se puede incontrar, usando KNN serà: ',Proyeccion_1[0])
print('')
print('Según los datos ingresados, la proyección para definir la categoria de obesidad donde se puede incontrar: ',Proyeccion_2[0])
```

```
Ingrese su edad: 34
Ingrese su genero, para masculino 1 y femenino 2: 1
Ingrese su altura: 172
Ingrese su peso: 76
Ingrese su indice de masa corporal: 18
Ingrese su nivel de actividad fisica: 3


Según los datos ingresados, la proyección para definir la categoria de obesidad donde se puede incontrar, usando KNN serà:  2.6

Según los datos ingresados, la proyección para definir la categoria de obesidad donde se puede incontrar:  1.0833151281327233
```

5. Aprendizaje no supervisado

Mediante modelos de aprendizaje no supervisado vamos a realizar un modelo de recomendaciones de acuerdo a la información que se obtiene por medio de una captura manual o formulario con el fin de agrupar y segmentar usuarios del gimnasio MEGA HIT GYM, para recomendar y prevenir el padecimiento de obesidad en alguno de sus niveles.

Al igual que el Modelamiento del punto anterior debemos realizar el ajuste a la fuente de datos que vamos a utilizar asignado valor numérico a las columnas en las cuales sus valores son de tipo texto:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   1000 non-null   int64
 1   Gender                1000 non-null   int64
 2   Height                1000 non-null   float64
 3   Weight                1000 non-null   float64
 4   BMI                   1000 non-null   float64
 5   PhysicalActivityLevel 1000 non-null   int64
 6   ObesityCategory       1000 non-null   int64
dtypes: float64(3), int64(4)
memory usage: 54.8 KB
```

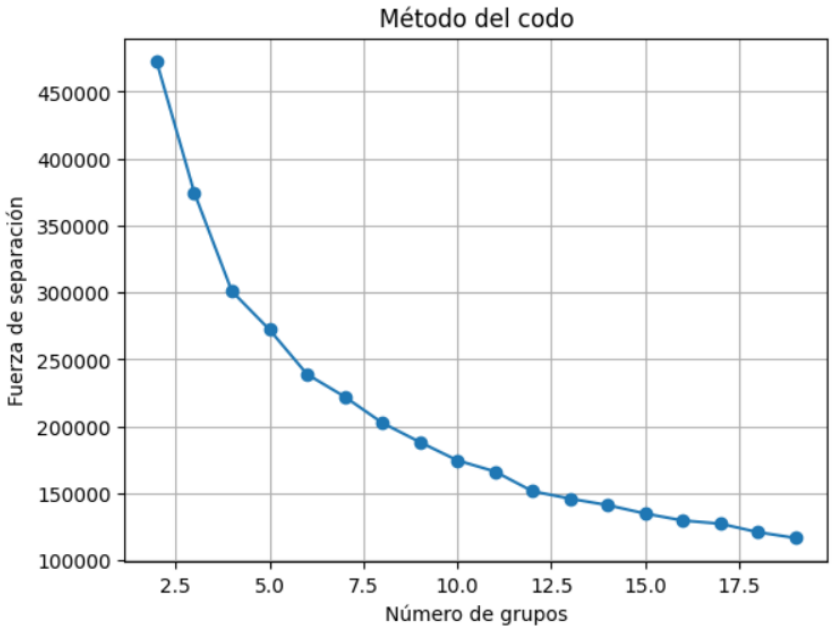

Gráfico 6. Tabla método de Codo agrupamiento de usuarios

A De acuerdo al grafico 4, se usará para el agrupamiento de valores superiores a 7.5 para usar información más precisa, esto significa que usaremos las 6 primeras agrupaciones.

Se usará una muestra de 10 registros para evidenciar el procesamiento de esta información y validar el comportamiento generado:

	Age	Gender	Height	Weight	BMI	PhysicalActivityLevel	ObesityCategory	Grupo
0	56	1	173.575262	71.982051	23.891783	4	1	2
1	69	1	164.127306	89.959256	33.395209	2	2	3
2	46	2	168.072202	72.930629	25.817737	4	3	2
3	32	1	168.459633	84.886912	29.912247	3	3	0
4	60	1	183.568568	69.038945	20.487903	3	1	1
5	25	2	166.405627	61.145868	22.081628	4	1	4
6	78	1	183.566334	92.208521	27.364341	3	3	3
7	38	1	142.875095	59.359746	29.078966	1	3	2
8	56	1	183.478558	75.157672	22.325577	4	1	1
9	75	1	182.974061	81.533460	24.353244	2	1	1

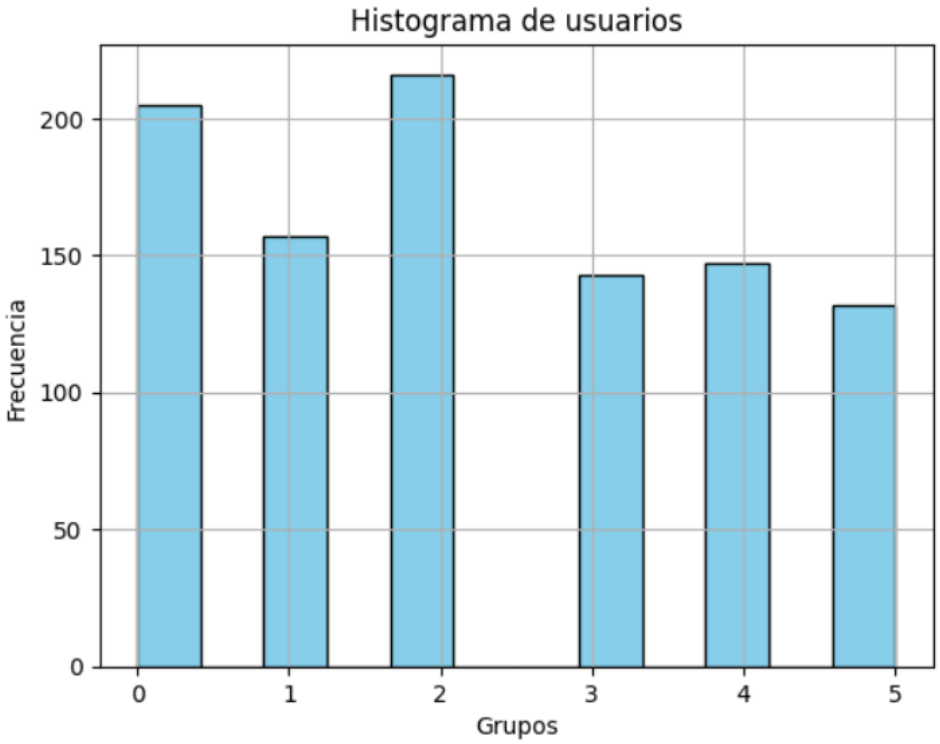

Gráfico 7. Tabla método de Codo agrupamiento de usuarios

Si analizamos el grafico generado determinar que tenemos 4 grupos con valores muy similares, y los grupos superiores también cuentas con una diferencia mínima un comportamiento que se establece mediante la selección de 6 grupos como punto de referencia.

Centroides:

	Age	Gender	Height	Weight	BMI	PhysicalActivityLevel	ObesityCategory
0	29.746341	1.478049	172.150899	84.366924	28.771734	2.604878	2.302439
1	68.547771	1.464968	176.934137	71.995685	23.134682	2.477707	1.681529
2	47.893519	1.537037	164.886326	69.646057	25.791622	2.527778	1.981481
3	65.888112	1.482517	166.114781	88.825283	32.362137	2.601399	2.356643
4	30.809524	1.469388	171.414700	56.298953	19.364488	2.503401	2.278912
5	65.916667	1.393939	169.810623	49.891796	17.519777	2.462121	2.833333

Analizando la información podemos encontrar que varios de los grupos que estamos modelando para ser más exactos en 4 grupos se encuentran en la categoría de obesidad, un dato preocupante que enciende las alarmas para generar planes de trabajo para combatir esta enfermedad y bajar los niveles en los diferentes grupos, si bien contamos con 2 grupos donde se encuentran un pero normal, también se pueden generar una estrategia para mantener estos excelentes niveles de salud y no descuidar la posibilidad de padecer esta enfermedad.

```
#Tomando un cliente nuevo para asignarle un grupo según el modelo anterior
Cliente_New = np.zeros((1,7))
Cliente_New[0,0] = float(input('Ingrese la edad: '))
Cliente_New[0,1] = float(input('Ingrese el género (1 masculino - 2 femenino):'))
Cliente_New[0,2] = float(input('Digite la altura: '))
Cliente_New[0,3] = float(input('Digite el peso: '))
Cliente_New[0,4] = float(input('Digite el indice de masa musuclar (BMI): '))
Cliente_New[0,5] = float(input('Ingrese la actividad fisica: '))
Cliente_New[0,6] = float(input('Ingrese su categoria de la enfermedad: '))
Etiqueta_Cliente = Modelo_Cluster.predict(Cliente_New)
print('Según los datos del usuario, el grupo es: ',Etiqueta_Cliente)
print(' ')

if Etiqueta_Cliente == 0 or Etiqueta_Cliente == 2 :
  print('Excelente mantenga sus buenos habitos de salud, hidratacion y actividad fisica')
if Etiqueta_Cliente == 1 or Etiqueta_Cliente == 3 or Etiqueta_Cliente == 4 or Etiqueta_Cliente == 5:
  print('Debe mejorar sus habitos de salud, puede que este en riesgo de padecer alguna enfermedad relacionada con la obesidad')
```

```
Ingrese la edad: 68
Ingrese el género (1 masculino - 2 femenino):2
Digite la altura: 160
Digite el peso: 80
Digite el indice de masa musuclar (BMI): 25
Ingrese la actividad fisica: 1
Ingrese su categoria de la enfermedad: 2
Según los datos del usuario, el grupo es:  [3]

Debe mejorar sus habitos de salud, puede que este en riesgo de padecer alguna enfermedad relacionada con la obesidad
```

Mediante el sistema de inteligencia artificial que se modelo y entreno con la información recolectada en el estudio de obesidad en el gimnasio MEGA HIT GYM, se creó un formulario que permite ilustrar de manera muy sencilla un sistema de recomendación para los usuarios así pueden realizar un seguimiento preventivo para dejar de padecer obesidad o mejorar los niveles actuales donde se encuentra, con la información suministrada por el levantamiento de la información necesaria.

Referencias bibliográficas

Castrillón, O. D., Sarache, W., & Castaño, E. (2017). Sistema Bayesiano para la Predicción de la Diabetes. *Información tecnológica*, *28*(6), 161-168.

Muñiz, C. C., León-García, P. E., Diaz, A. S., & Hernández-Pérez, E. (2023). Predicción de diabetes mellitus basada en el índice triglicéridos y glucosa. Medicina Clínica, 160(6), 231-236.

Fernández, S. C., Martínez, M. M., Montero, C. D. F., Rodríguez, I. G., Arenas, Á. V., & Calvo, M. O. (2021). Modelos predictivos de diabetes gestacional, un nuevo modelo de predicción. *Medicina de Familia. SEMERGEN*, *47*(8), 515-520.

Ascona, Y. D. A. (2019, June). Métodos de aprendizaje supervisado para la predicción de diabetes. In *8va Jornada Cientifica de estudiantes-FIA*.

Ascona, Y. D. A. (2019, June). Métodos de aprendizaje supervisado para la predicción de diabetes. In *8va Jornada Cientifica de estudiantes-FIA*.

May, O. A. C., Koo, J. J. P., Kinani, J. M. V., & Encalada, M. A. Z. (2018). Construcción De Un Modelo De Predicción Para Apoyo Al Diagnóstico De Diabetes (Construction of a Prediction Model To Support the Diabetes Diagnosis). *Pistas Educativas*, *40*(130).

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016). Tratamiento masivo de datos utilizando técnicas de Machine Learning.

Vázquez, A. M. (2018). Introducción a machine learning.

Portela González, J. (2024). Machine learning: aprendizaje no supervisado.