

# Proyecto IA1

Daniel L. Aguilar Navas - 2230034  
Juan José Ardila Aragón - 2230035

# Contenido:

01

## Presentación del proyecto

Título  
Motivación  
Objetivo

02

## Presentación del dataset

Información general  
Columnas

03

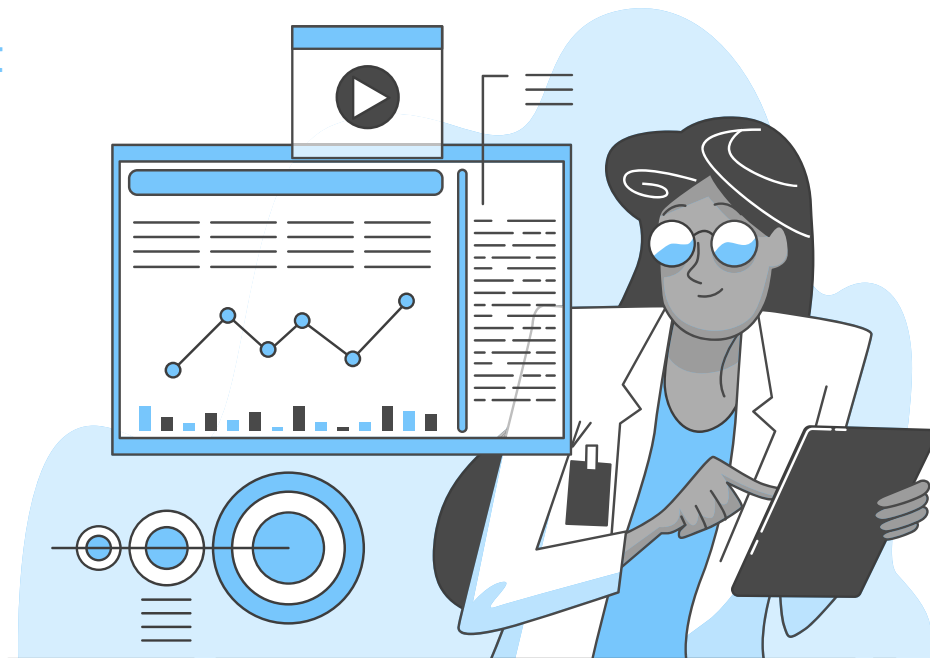
## Procesamiento

Procesamiento del dataset  
Estadísticas del dataset

04

## Algoritmo genético

Estructura del algoritmo



05

Clasificación por  
RandomForest

06

Clasificación por SVM

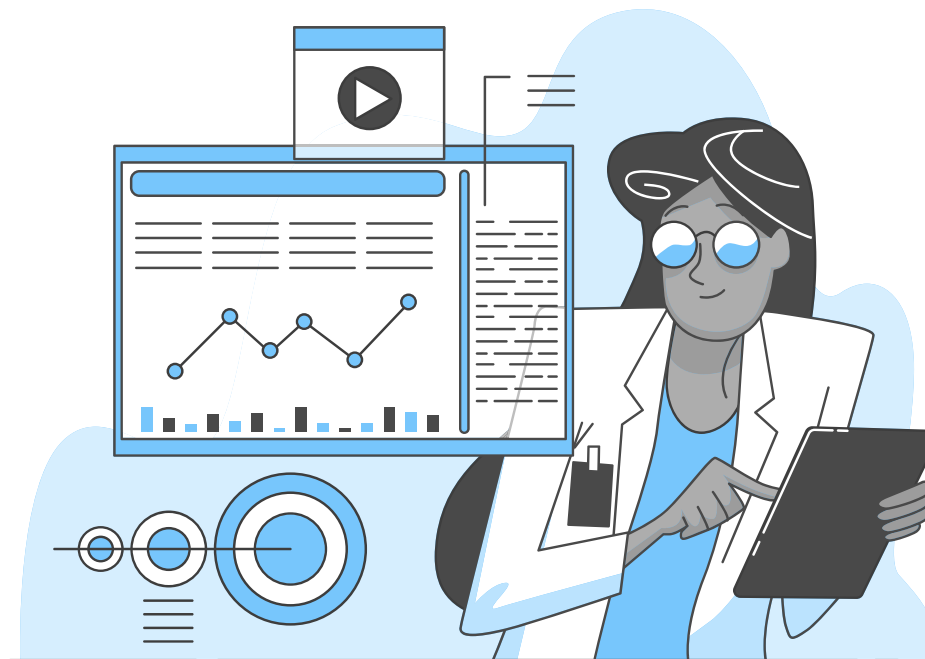
07

Regresión por DNN

08

Regresión por DT

# Contenido:



# Contenido:

09

No supervisado por PCA y  
TSNE

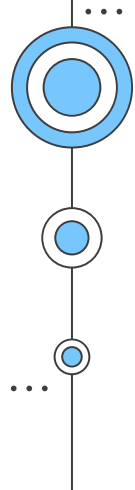
10

No supervisado por  
K-Means y Agglomerative  
Clustering

11

Comparaciones y  
Conclusiones

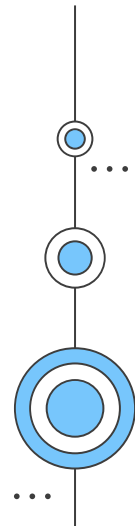




# 01

# Presentación

General del proyecto



# Predicción del IMC en la Población en base a sus hábitos



...

# Motivación:

**México:** El 40% de adolescentes y 38% de los adultos tienen obesidad. <sup>[1]</sup>

**Colombia:** Alrededor del 25% de adultos son obesos, con mayor incidencia en zonas urbanas. <sup>[2]</sup>

**Perú y Colombia:** La obesidad infantil está relacionada con trastornos metabólicos. <sup>[3]</sup>



# Objetivo:



**Diseñar y desarrollar** un modelo de aprendizaje automático para predecir el nivel de obesidad en función de los hábitos y características personales.

...





# 02

## Dataset

Info. general  
Columnas

# Obesity Prediction Dataset



Este conjunto de datos extraído de kaggle<sup>[4]</sup> incluye datos para la estimación de los niveles de obesidad en individuos de los países de México, Perú y Colombia

# Obesity Prediction Dataset

El dataset se encuentra compuesto por **17 columnas** y **2111 registros** que brindan información sobre el estado físico de la persona y diferentes hábitos que presenta en su diario vivir.



- Gender: Género
- Age: Edad
- Height: Altura (en metros)
- Weight: Peso (en kilogramos)
- family\_history: ¿Algún miembro de tu familia ha sufrido o sufre de sobrepeso?
- FAVC: ¿Consumes alimentos con alto contenido calórico frecuentemente?
- FCVC: ¿Sueles incluir verduras en tus comidas?
- NCP: ¿Cuántas comidas principales tienes al día?
- CAEC: ¿Consumes algún alimento entre comidas?
- SMOKE: ¿Fumas?
- CH2O: ¿Cuánta agua bebes diariamente?
- SCC: ¿Monitoreas las calorías que consumes diariamente?
- FAF: ¿Con qué frecuencia realizas actividad física?
- TUE: ¿Cuánto tiempo usas dispositivos tecnológicos como celulares, videojuegos, televisión, computadora y otros?
- CALC: ¿Con qué frecuencia consumes alcohol?
- MTRANS: ¿Qué medio de transporte utilizas normalmente?
- Obesity\_level (Target Column): Nivel de obesidad

# 03

## Procesamiento

Procesamiento del dataset  
Estadísticas del dataset

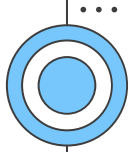
# Cómo se realizó el procesamiento:

1. Cambiar nombre de columnas
2. Verificación de datos nulos.
3. Insertar una columna respecto a IMC
4. Cambiar columnas que contienen “Strings” por números enteros que simbolizan un dato.
5. Calcular estadísticas de los datos.



# 04

## Algoritmo Genético



# Estructura del A.G.

Inicialización: Se selecciona una población aleatoria de individuos del dataset.

Fitness Function: Evalúa la salud de cada individuo con base en condiciones como IMC, consumo de calorías, ejercicio, etc.

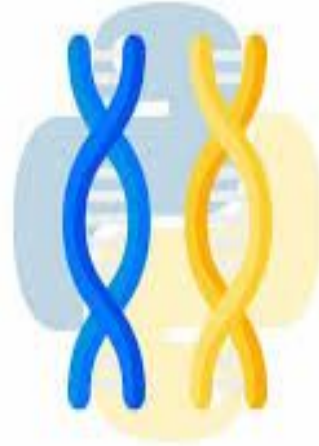
Selección: Se eligen los mejores individuos en cada generación.

Crossover: Se combinan características de los individuos seleccionados.

Mutación: Se introducen cambios aleatorios en algunos individuos.

Reducción de Población: Se mantiene la mejor mitad de los individuos para la siguiente generación.


Resultado: Se obtiene un vector con los índices de las mejores condiciones del dataset.





05

Clasificación por RF  
por tipo de  
obesidad





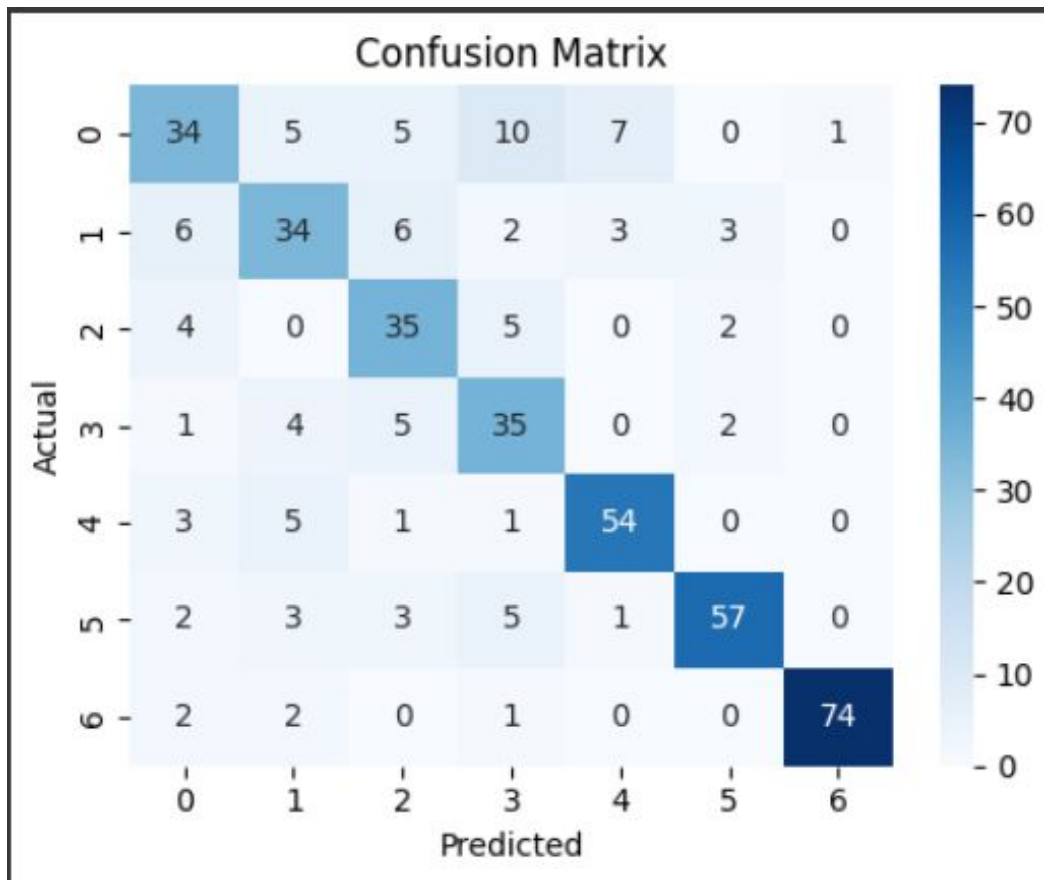
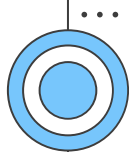


# Random forest

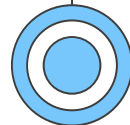
Train-test 80-20

criterion="gini" evaluando que tan mezcladas están las clases en un nodo

Profundidad del árbol	Train Accuracy	Train F1 Score	Train Recall	Test Accuracy	Test F1 Score	Test Recall
3	0.56 ± 0.01	0.51 ± 0.01	0.56 ± 0.01	0.54 ± 0.04	0.49 ± 0.04	0.54 ± 0.04
5	0.66 ± 0.01	0.62 ± 0.01	0.66 ± 0.01	0.62 ± 0.04	0.58 ± 0.04	0.62 ± 0.04
7	0.77 ± 0.01	0.76 ± 0.01	0.77 ± 0.01	0.69 ± 0.03	0.67 ± 0.04	0.69 ± 0.03
<b>10</b>	<b>0.91 ± 0.01</b>	<b>0.91 ± 0.01</b>	<b>0.91 ± 0.02</b>	<b>0.74 ± 0.02</b>	<b>0.74 ± 0.03</b>	<b>0.74 ± 0.02</b>
15	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.02	0.76 ± 0.02	0.75 ± 0.02	0.76 ± 0.02



...

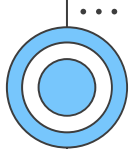


...



# 06

**Clasificación por  
SVM por tipo de  
obesidad**



# SVM

## Con sigmoid

Reporte de clasificación tras K-Fold con SVC:

	precision	recall	f1-score	support
0.0	0.50	0.43	0.46	287
1.0	0.58	0.27	0.37	290
2.0	0.52	0.25	0.34	290
3.0	0.50	0.60	0.54	351
4.0	0.60	0.64	0.62	272
5.0	0.54	0.81	0.65	297
6.0	0.75	0.99	0.86	324
accuracy			0.58	2111
macro avg	0.57	0.57	0.55	2111
weighted avg	0.57	0.58	0.55	2111

...

## Con ajuste polinomico


📄 Reporte de clasificación tras K-Fold con SVC:

	precision	recall	f1-score	support
0	0.49	0.43	0.45	235
1	0.52	0.37	0.43	237
2	0.54	0.35	0.42	235
3	0.53	0.62	0.57	292
4	0.69	0.65	0.67	207
5	0.59	0.86	0.70	233
6	0.84	0.98	0.90	249
accuracy			0.61	1688
macro avg	0.60	0.61	0.59	1688
weighted avg	0.60	0.61	0.59	1688



07

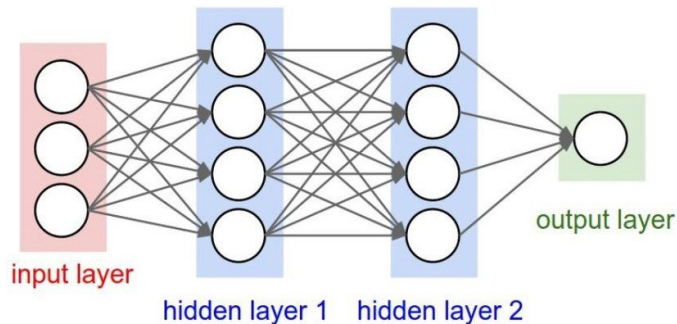
## Regresión de IMC por DNN





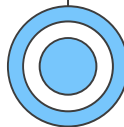
# Dense Neural Network

Para entrenar el modelo se usaron **3 capas relu de 512,128,64** parámetros y por último, **una capa densa** para predicción de datos.

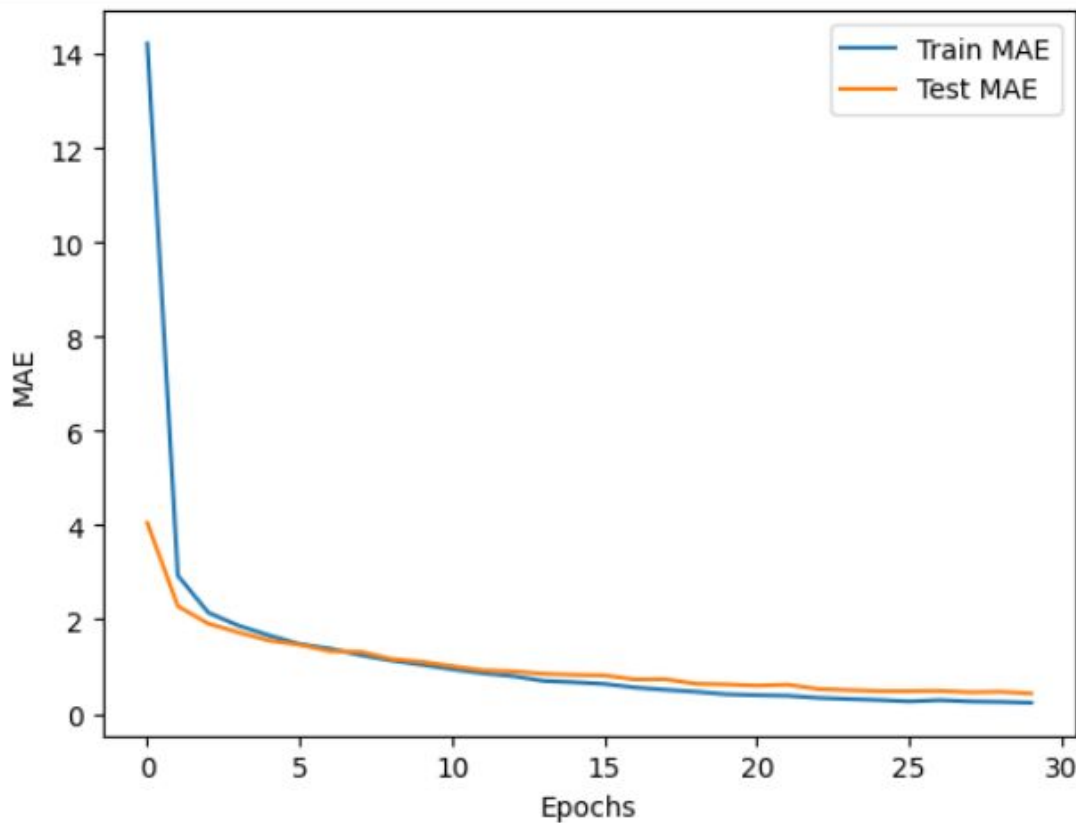


```
Predicciones (primeros 5): [19.109716 39.29602 23.070292 16.962923 29.497086]  
Reales (primeros 5): [17.41536553 42.03995319 17.53104456 18.17867036 24.16326531]
```

...



# Grafica train vs test MAE





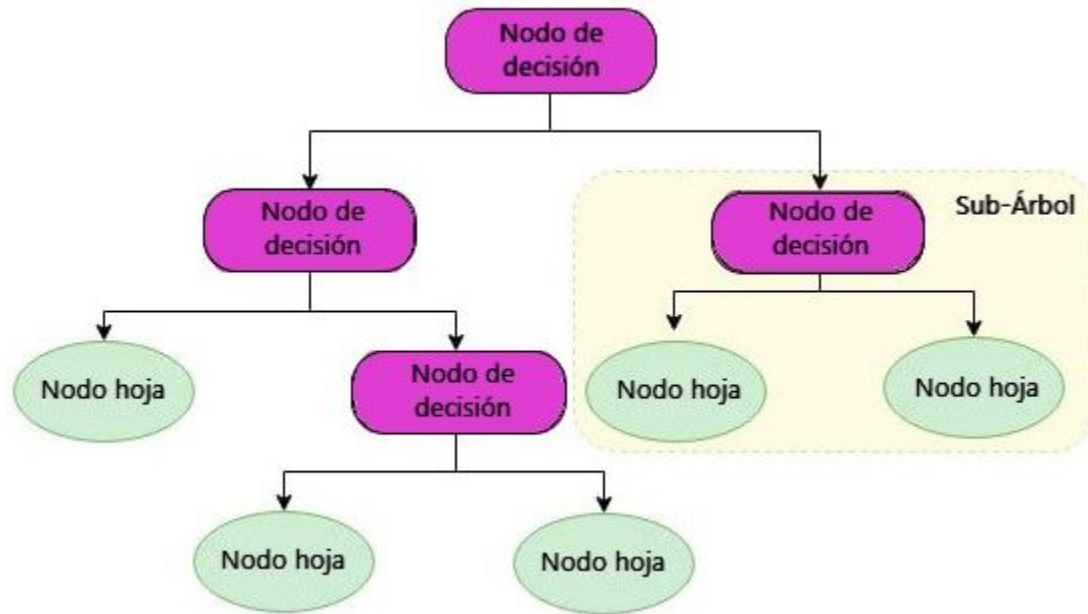


08

# Regresión de IMC por DT



# Árbol de Decisión para regresión:



```
Predicciones de IMC con datos de prueba: [18.66783428 19.1349481 27.67775922]  
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:2732: UserWarning: X has  
warnings.warn()
```

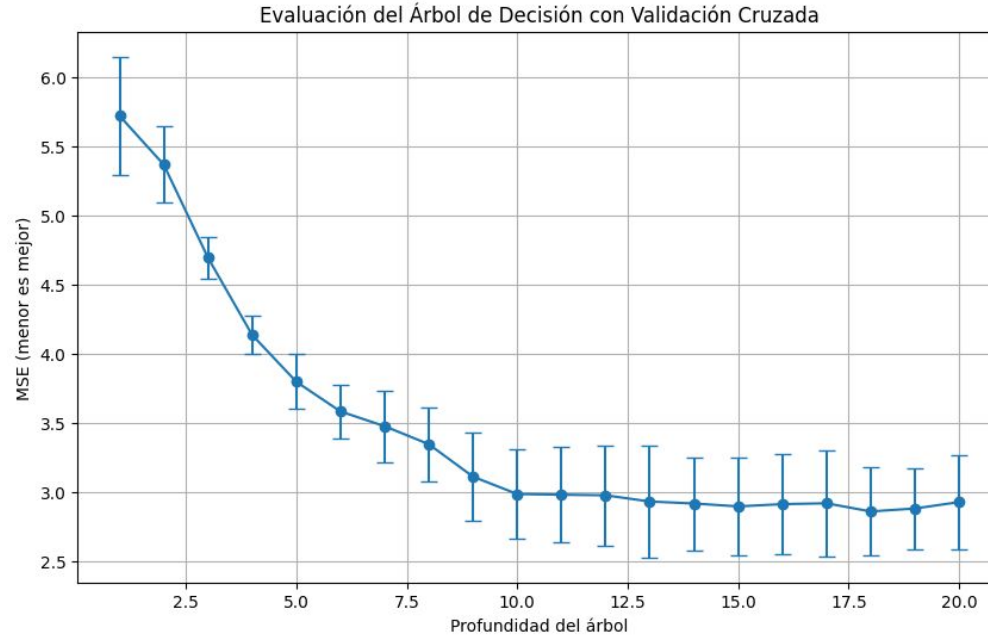
# Variación de la profundidad del DT:

```
Profundidad: 1 -> MAE: 5.718 (+/- 0.42787)
Profundidad: 2 -> MAE: 5.365 (+/- 0.27632)
Profundidad: 3 -> MAE: 4.693 (+/- 0.14970)
Profundidad: 4 -> MAE: 4.135 (+/- 0.13519)
Profundidad: 5 -> MAE: 3.797 (+/- 0.19740)
Profundidad: 6 -> MAE: 3.582 (+/- 0.19213)
Profundidad: 7 -> MAE: 3.476 (+/- 0.25854)
Profundidad: 8 -> MAE: 3.346 (+/- 0.26488)
Profundidad: 9 -> MAE: 3.111 (+/- 0.31895)
Profundidad: 10 -> MAE: 2.986 (+/- 0.32032)
Profundidad: 11 -> MAE: 2.981 (+/- 0.34192)
Profundidad: 12 -> MAE: 2.977 (+/- 0.36276)
Profundidad: 13 -> MAE: 2.932 (+/- 0.40381)
Profundidad: 14 -> MAE: 2.918 (+/- 0.33592)
Profundidad: 15 -> MAE: 2.897 (+/- 0.35425)
Profundidad: 16 -> MAE: 2.913 (+/- 0.36164)
Profundidad: 17 -> MAE: 2.919 (+/- 0.38172)
Profundidad: 18 -> MAE: 2.860 (+/- 0.31706)
Profundidad: 19 -> MAE: 2.881 (+/- 0.29248)
Profundidad: 20 -> MAE: 2.928 (+/- 0.34083)
```

Por medio de validación cruzada  
y kfold = 10

**Mejor profundidad según CV: 10 con  
MSE: 21.743**

# Variación de la profundidad del DT:



**Mejor profundidad según CV: 18 con  
MAE: 2.860**

Métrica	DNN	DT (k fold)
MAE	3.7987277	2.860

El mejor modelo fue el Decision Tree. 😎🌲

...



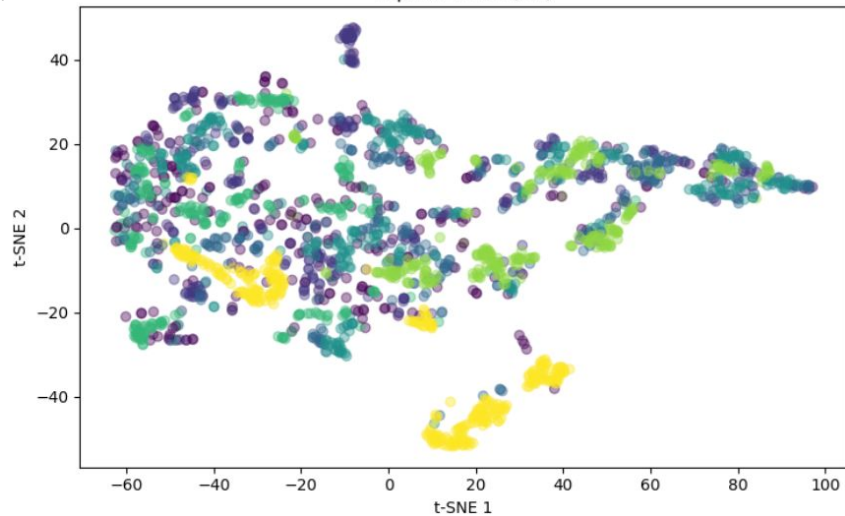
09

No supervisado por  
PCA y TSNE

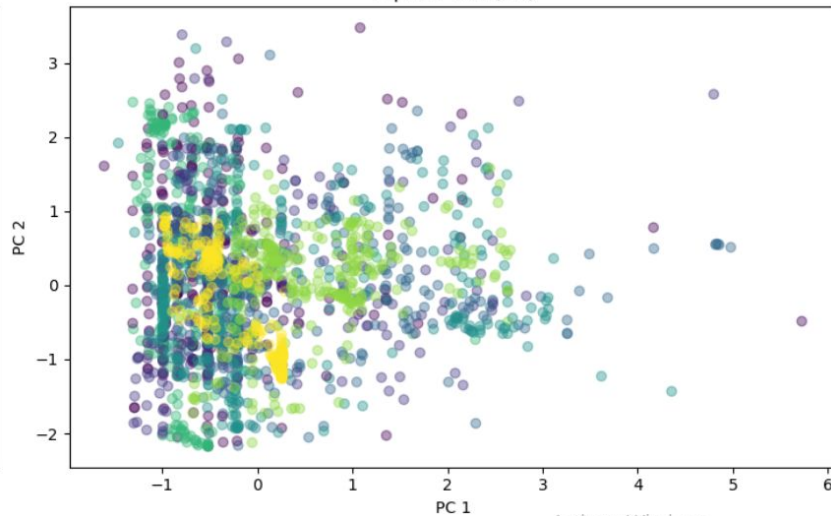


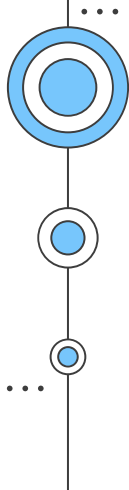
# No supervisados

Espacio t-SNE (2D)

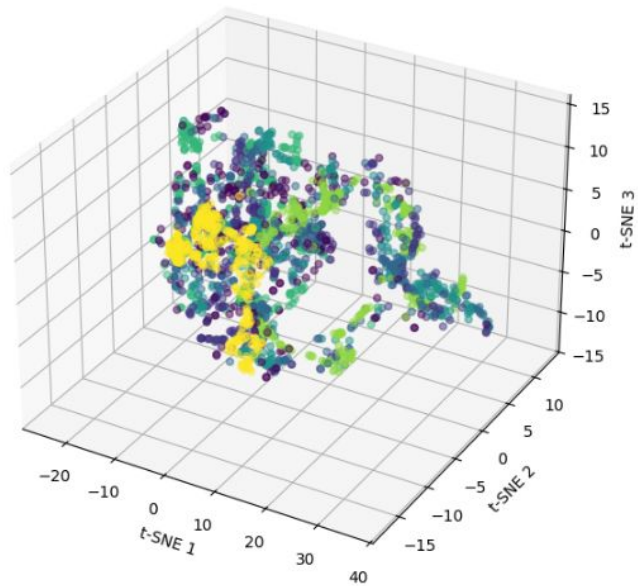


Espacio PCA (2D)

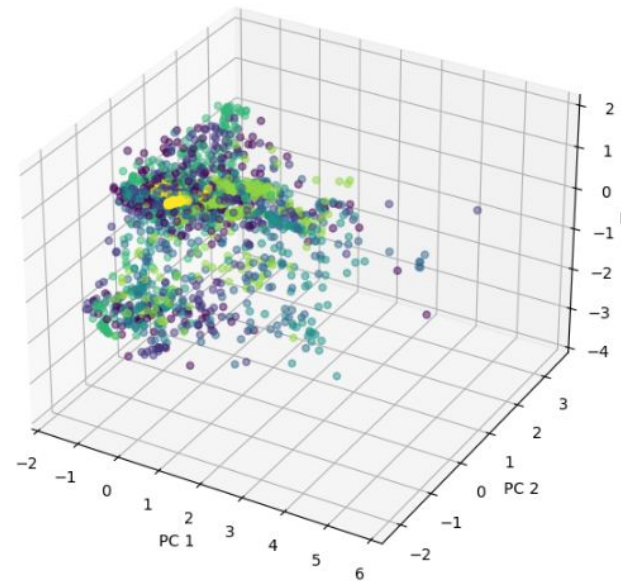




Espacio t-SNE (3D)



Espacio PCA (3D)



...

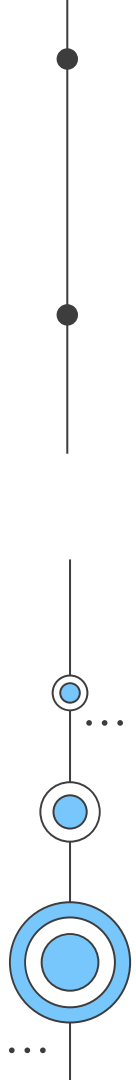


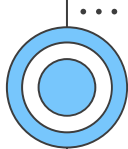




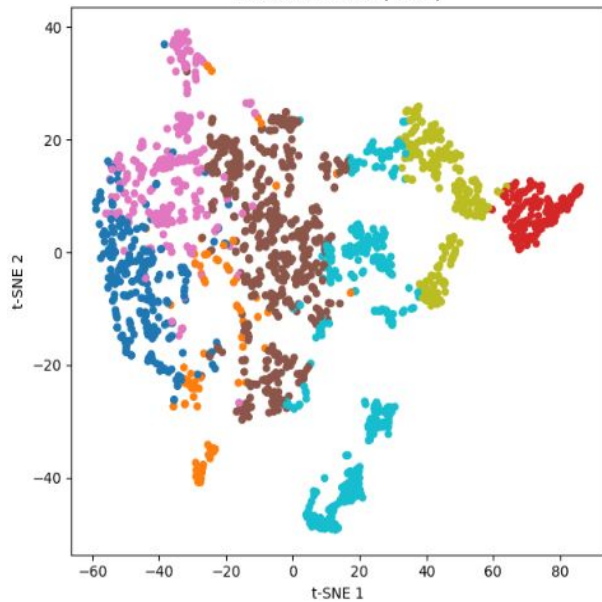
# 10

No supervisado por  
K-Means y  
Agglomerative  
Clustering

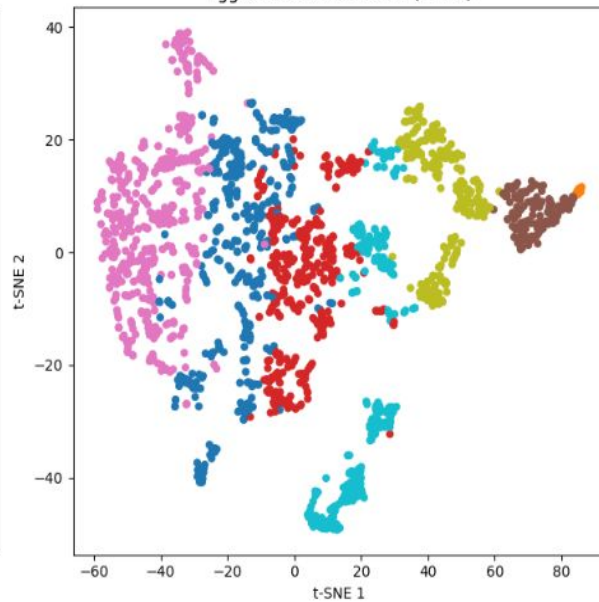




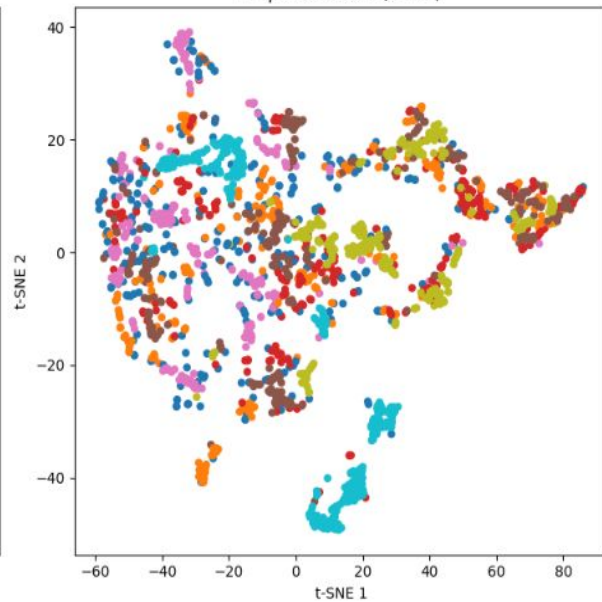
KMeans alineado (t-SNE)



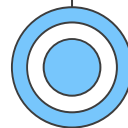
Agglomerative alineado (t-SNE)

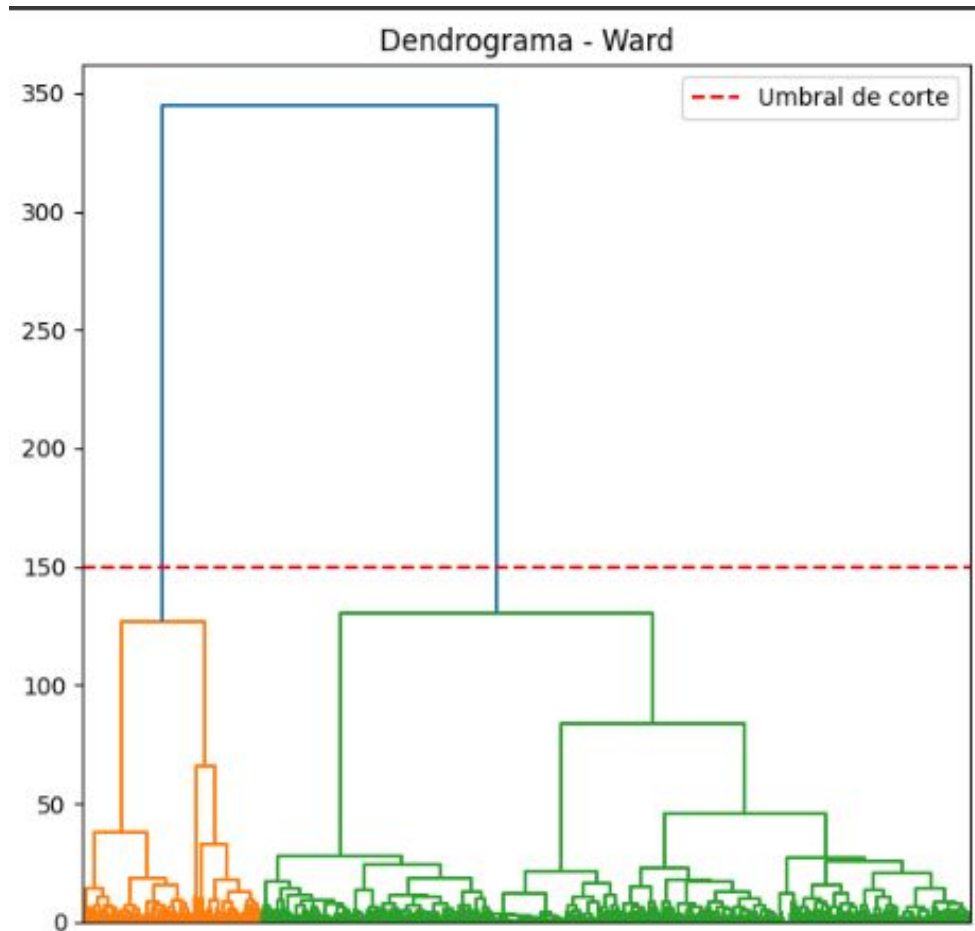
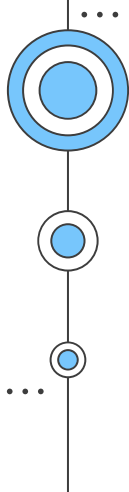


Etiquetas reales (t-SNE)

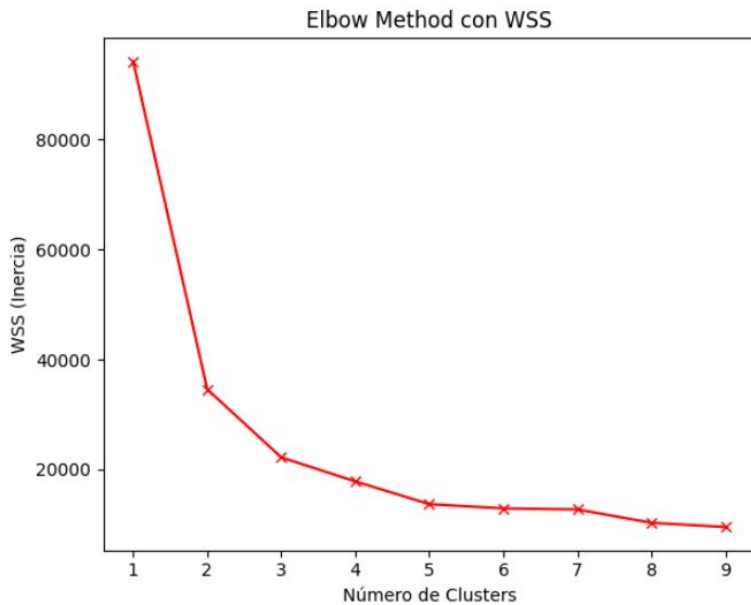
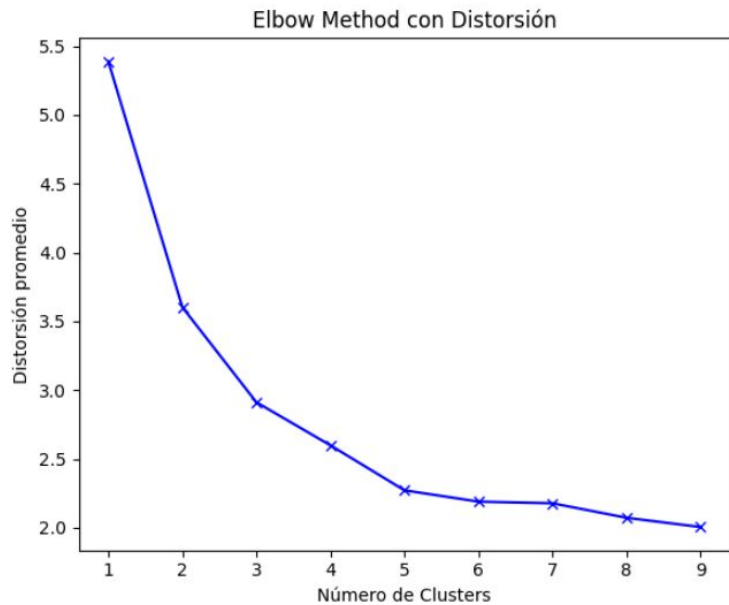


...






# k-means




# 11

## Comparaciones y Conclusiones



Modelo	Tipo	Accuracy	F1-Score	Recall
Random Forest(K-fold=10)	Supervisado	<b>0.76 ± 0.02</b>	<b>0.75 ± 0.02</b>	<b>0.76 ± 0.02</b>
SVM (K-Fold=10)	Supervisado	0.61 ± 0.03	0.59 ± 0.03	0.61 ± 0.03
K Means	No supervisado	0.32	0.30	0.32
Agglomerative Clustering	No supervisado	0.31	0.29	0.31



El mejor modelo fue el Random Forest. 🤘🌲

...

# Referencias:

1)Zafra-Tanaka, J.H., Braverman, A., et al. (2023). City features related to obesity in preschool children: A cross-sectional analysis of 159 cities in Latin America. The Lancet Regional Health.

[https://www.thelancet.com/journals/lanam/article/PIIS2667-193X\(23\)00032-7/fulltext](https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(23)00032-7/fulltext)

2)Castro, P.A., & Spijker, J. (2024). Adult Obesity in Colombia from the Sociodemographic and Public Health Perspective: A Scoping Review. Revista Gerencia y Políticas de Salud.

3)Loayza-Castro, J.A., Vera-Ponce, V.J., et al. (2024). Maternal obesogenic environment and its association with childhood obesity in Peru: A 9-year analysis. MedRxiv.

4) Dataset : <https://www.kaggle.com/datasets/ruchikakumbhar/obesity-prediction>  
5)colab:

[https://colab.research.google.com/drive/1ovEYbf1fTEQXc\\_Gr8\\_8g0EBOyGSsPbrk?usp=drive\\_link](https://colab.research.google.com/drive/1ovEYbf1fTEQXc_Gr8_8g0EBOyGSsPbrk?usp=drive_link)