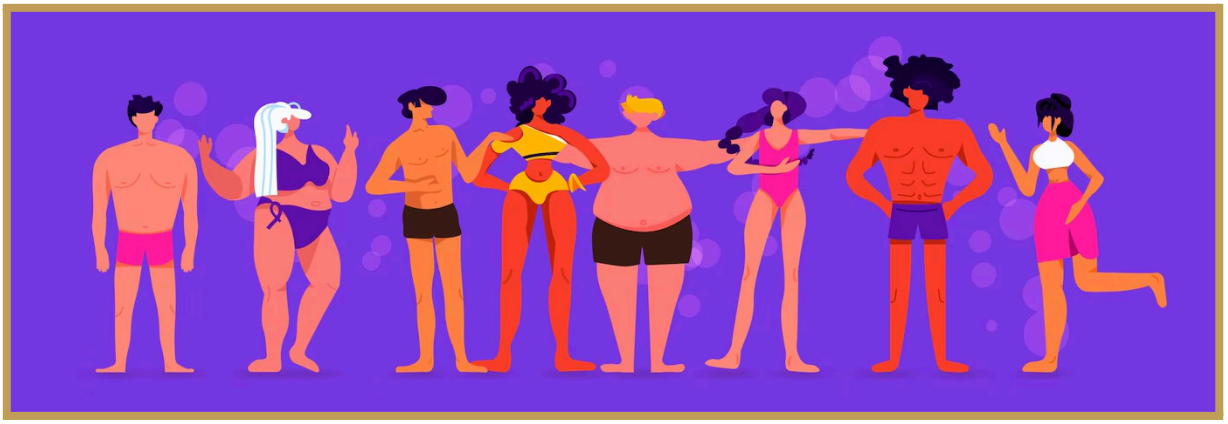


Factores que influyen en el peso de las personas.

Marcela Barrios

Data Science - Coderhouse
Junio 2024

Índice.

1 - Abstract con motivación y audiencia.....3
2 - Contexto y objetivos..... 3
3 - Información a utilizar..... 3
4 - Análisis general de los datos..... 4
5 - Comparación de las categorizaciones.....5
6 - ¿Cómo se dividen los datos respecto al género?..... 6
7 - Variables categóricas..... 7
7.1 - Family history with overweight: Historial familiar con sobrepeso..... 7
7.2 - FAVC: Frecuencia de consumo de alimentos altos en calorías por sem.....7
7.3 - CAEC: Consumo de alimentos entre comidas.....8
7.4 - SMOKE: Consumo de tabaco..... 8
7.5 - SCC: Control del consumo de calorías..... 8
7.6 - CALC: Consumo de alcohol..... 9
7.7 - MTRANS: Transporte utilizado..... 9
7.8 - NObeyesdad: Nivel de obesidad.....10
7.9 - Cat IMC: Categorías de IMC..... 10
8 - Variables numéricas..... 11
8.1 - Age: Edad..... 11
8.2 - Height: Altura..... 12
8.3 - Weight: Peso..... 12
8.4 - FCVC: Frecuencia de consumo de vegetales por semana..... 13
8.5 - NCP: Número de comidas principales por día..... 13
8.6 - CH2O: Consumo de agua diario..... 14
8.7 - FAF: Frecuencia de actividad física por semana.....14
8.8 - TUE: Tiempo de uso de dispositivos tecnológicos por semana..... 15
8.9 - IMC: Índice de Masa Corporal..... 15
9 - ¿Cómo se relaciona el peso con la altura?.....16
10 - ¿La población consultada tiene mayoritariamente algún grado de sobrepeso u obesidad?..... 16
11 - ¿Influyen los niveles familiares en el sobrepeso?..... 17
12 - ¿Cómo se distribuyen las distintas categorías de NObeyesdad y de cat_IMC respecto a las edades estudiadas?..... 18
13 - Análisis de la relación entre el medio de transporte más utilizado y el peso, diferenciando si consumen tabaco o no..... 19
14 - Normalización..... 19
15 - ¿Existe relación entre el IMC y los campos sin importar el peso y la altura?..... 19
16 - Los modelos de clasificación y las métricas.....21
17 - ¿Cómo es la clasificación respecto a cat_IMC?..... 22

17.1 - Árbol de decisión....22
17.2 - kNN.... 23
17.3 - Regresión logística....23
17.4 - Random Forest.... 24
17.5 - XGBoost.... 25
17.6 - CatBoost.... 25
17.7 - LightGBM.... 26
17.8 - Resumen.... 26
18 - ¿Cómo es la clasificación respecto a NObeyesdad?.... 27
18.1 - Árbol de decisión....27
18.2 - kNN.... 28
18.3 - Regresión logística....28
18.4 - Random Forest.... 29
18.5 - XGBoost.... 30
18.6 - CatBoost.... 30
18.7 - LightGBM.... 31
18.8 - Resumen.... 31
19 - ¿Conviene otra cantidad de agrupamientos?.... 32
20 - Validación de los modelos.... 32
21 - Factores que tienen mayor y menor relevancia tienen sobre NObeyesdad.... 33
22 - Conclusión.... 33

1 - Abstract con motivación y audiencia.

El estudio en cuestión busca analizar distintos hábitos y costumbres cotidianas de las personas, y poder identificar los factores que mayor relevancia tienen en su peso.

Esto es importante para poder gestionar distintas políticas en diversas áreas buscando prevenir enfermedades, e incentivando una mayor conciencia personal y familiar, buscando generar una mejor calidad de vida.

Existen distintas entidades que se deben involucrar en esta problemática y gestionar acciones al respecto desde distintos ángulos:

- Organismos de salud.
- Organismos educativos de distintos niveles, tanto públicos como privados.
- Empresas.

2 - Contexto y objetivos.

Contexto comercial: Se busca poder conocer mejor la problemática del sobrepeso, identificar y diferenciar las grandes segmentaciones de sobrepeso y obesidad, y poder generar mejores políticas y campañas de prevención para poder reducirla.

Definición de objetivo: Predecir los factores que más favorecen el sobrepeso de las personas en relación a la alimentación, actividad física, entre otras.

Contexto analítico: Utilizando la información del dataset en conjunto con técnicas de análisis de datos y distintos modelos, se buscará dar respuestas a las preguntas e hipótesis planteadas posteriormente.

3 - Información a utilizar.

Para poder desarrollar el estudio en cuestión se utiliza información obtenida de una encuesta web anónima a habitantes de México, Perú y Colombia, dentro de una franja etaria de entre 14 y 61 años.

De dicha encuesta se recopila información de **2111 personas** sobre los siguientes **17 campos**:

- Gender: Género.
- Age: Edad.
- Height: Altura.
- Weight: Peso.
- Family_history_with_overweight: Historial familiar con sobrepeso.
- FAVC: Frecuencia de consumo de alimentos altos en calorías por semana.
- FCVC: Frecuencia de consumo de vegetales por semana.

- NCP: Número de comidas principales por día.
- CAEC: Consumo de alimentos entre comidas.
- SMOKE: Consumo de tabaco.
- CH2O: Consumo de agua diario.
- SCC: Control del consumo de calorías.
- FAF: Frecuencia de actividad física por semana.
- TUE: Tiempo de uso de dispositivos tecnológicos por semana.
- CALC: Consumo de alcohol.
- MTRANS: Transporte utilizado.
- NObeyesdad: Nivel de obesidad.

Se decide agregar el campo IMC (Índice de Masa Corporal), el cual relaciona el peso con la altura: $IMC = \frac{peso}{altura*altura} = \frac{Weight}{Height*Height}$

Y posee las siguientes clasificaciones:

Clasificación	Rangos
Insuficiente	menor a 18.5
Normal	entre 18.5 y 24.9
Sobrepeso_I	entre 25.0 y 26.9
Sobrepeso_II	entre 27.0 y 29.9
Obesidad_I	entre 30.0 y 34.9
Obesidad_II	entre 35.0 y 39.9
Obesidad_III	mayor a 40.0

Tabla 1 - Categorías de IMC.

La idea de agregar este campo es poder comprobar si hay diferencia entre la clasificación original del dataset (“NObeyesdad”) y la clasificación conocida popularmente mediante dicha fórmula.

4 - Análisis general de los datos.

Inicialmente se realizaron distintos estudios, corrección de tipo de datos, análisis de nulos, duplicados y outliers, y se obtuvo lo siguiente:

Datos nulos: No posee.

Filas repetidas: Se detectan, pero en un rango inferior al 1,5% sobre el total, por lo que no se descartan debido a que la encuesta fue en tres países distintos de forma anónima y

podía ocurrir que existan coincidencias en valores mínimos.

Valores outliers: Se detectaron sólo en uno de los tres métodos utilizados, por lo que se decide no eliminarlos.

Columna	Tipo de dato
Gender	object
Age	int
Height	float64
Weight	float64
Family_history_with_overweight	object
FAVC	object
FCVC	float64
NCP	float64
CAEC	object
SMOKE	object
CH2O	float64
SCC	object
FAF	float64
TUE	float64
CALC	object
MTRANS	object
NObeyesdad	object
IMC	float64
cat_IMC	object

Tabla 2 - Tipo de dato de los campos.

5 - Comparación de las categorizaciones.

Se analiza la cantidad de datos de cada categoría, tanto de “NObeyesdad” y “cat_IMC”, y se compara cómo se relacionan entre ellas.

Tipo	NObeyesdad	cat_IMC
Insufficient_Weight/ insuficiente	272	267
Normal_Weight/ normal	287	303
Overweight_Level_I/ Sobrepeso_I	290	276
Overweight_Level_II/ Sobrepeso_II	290	292
Obesity_Type_I/ Obesidad_I	351	369
Obesity_Type_II/ Obesidad_II	297	334
Obesity_Type_III/ Obesidad_III	324	270

Tabla 3 - División de los datos en ambas categorías.

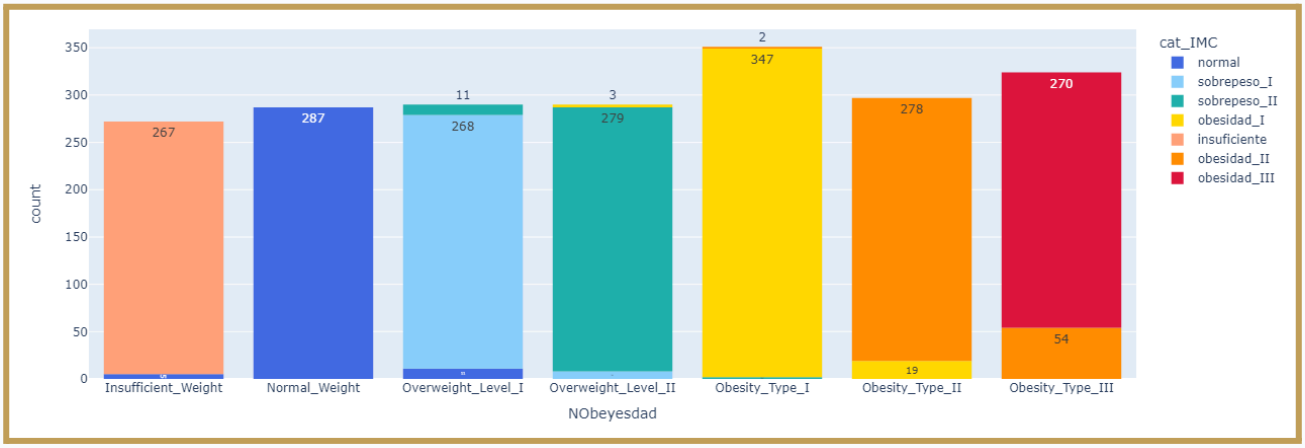


Gráfico 1 - Conteo de NObeyesdad separado por cat_IMC.

6 - ¿Cómo se dividen los datos respecto al género?

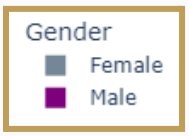
El relevamiento distingue solo dos categorías: Male y Female, que se dividen de la siguiente forma:

- Male: 1068.
- Female: 1043.

Teniendo una diferencia entre ellas menor al 1%.

7 - Variables categóricas.

Análisis de las nueve variables categóricas, tanto de forma total, como separando por el género con la siguiente identificación:



7.1 - Family history with overweight: Historial familiar con sobrepeso.

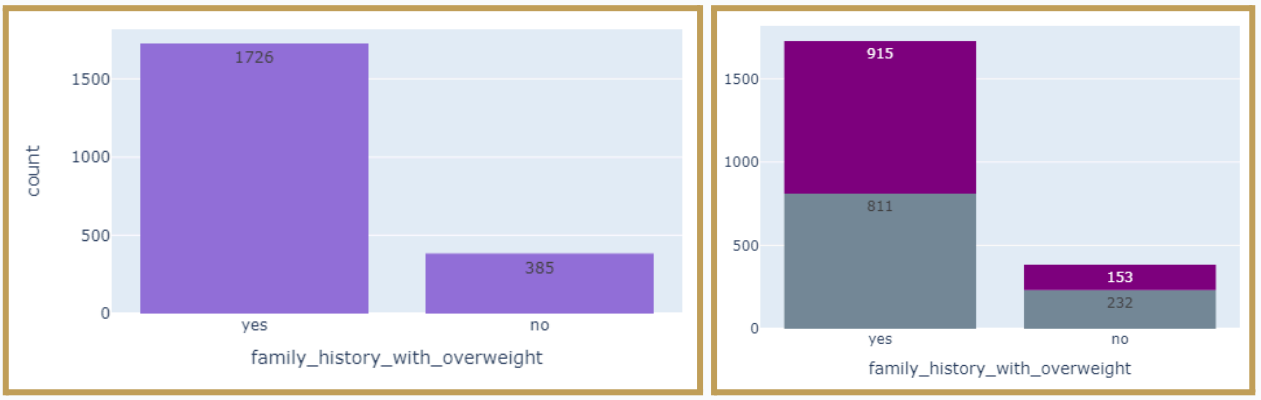


Gráfico 2 - Distribución por family_history_with_overweight: total y separado por género.

Aproximadamente el 82% si posee antecedentes, de los cuales el 53% son varones, mientras que para el caso de las personas sin antecedentes los varones son el 40%.

7.2 - FAVC: Frecuencia de consumo de alimentos altos en calorías por sem.

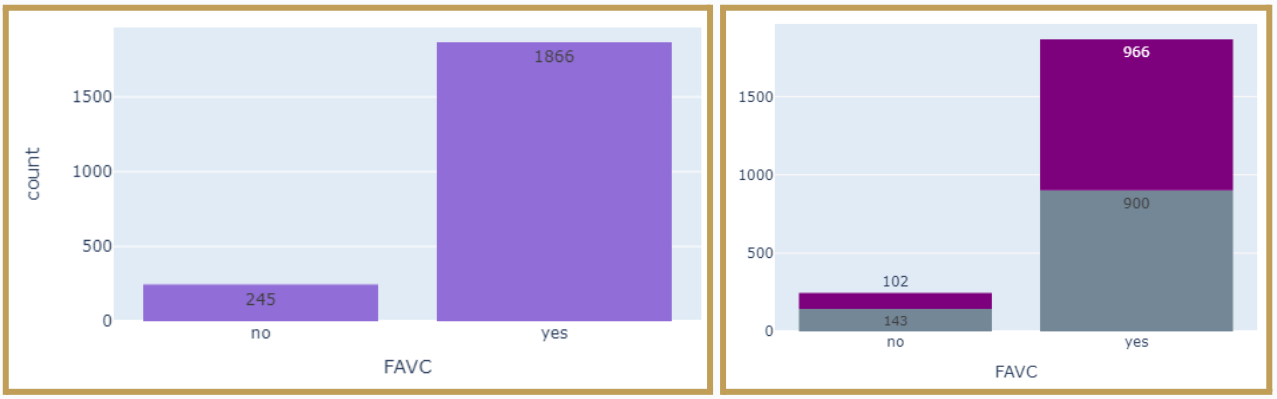


Gráfico 3 - Distribución por FAVC: total y separado por género.

Aproximadamente el 88% consume frecuentemente alimentos altos en calorías, de donde el 52% son varones y el resto son mujeres.

En el caso de la respuesta negativa el 58% son mujeres y el resto varones.

7.3 - CAEC: Consumo de alimentos entre comidas.

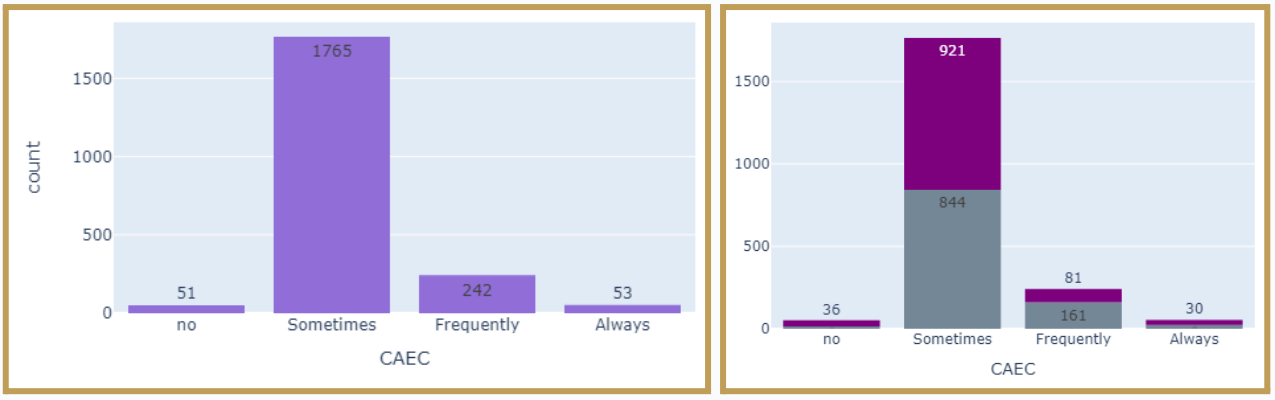


Gráfico 4 - Distribución por CAEC: total y separado por género.

La respuesta se divide en 2% no, 84% sometimes, 11%: frequently, y 3%: always.

7.4 - SMOKE: Consumo de tabaco.

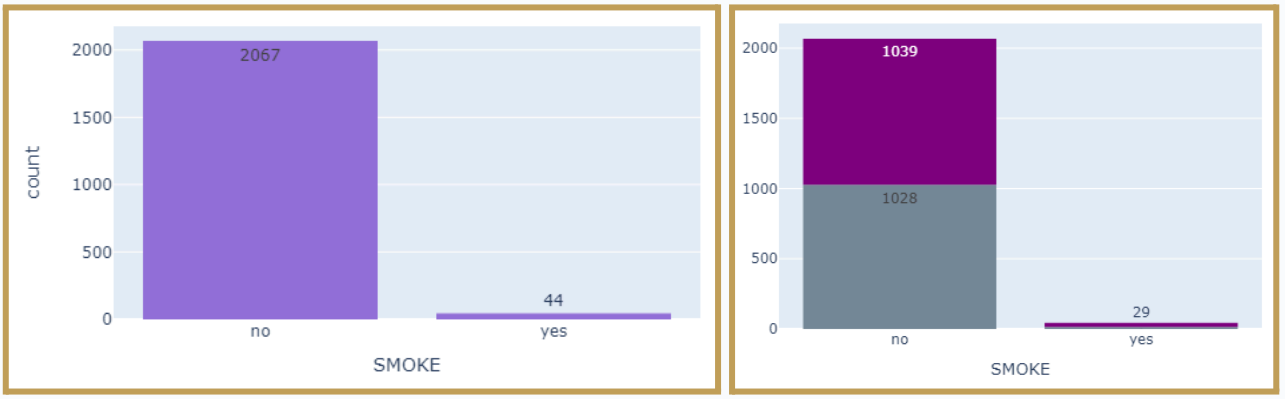


Gráfico 5 - Distribución por SMOKE: total y separado por género.

El 98% de las personas encuestadas no fuma, y del 2% restante, el 66% son varones.

7.5 - SCC: Control del consumo de calorías.

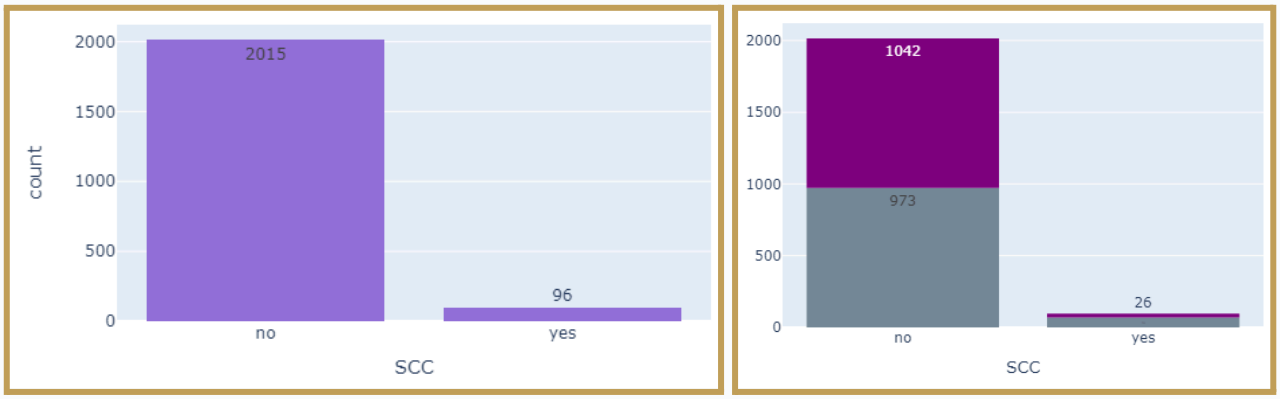


Gráfico 6 - Distribución por SCC: total y separado por género.

Solo el 5% controla las calorías que ingiere, del cual el 73% son mujeres.

7.6 - CALC: Consumo de alcohol.

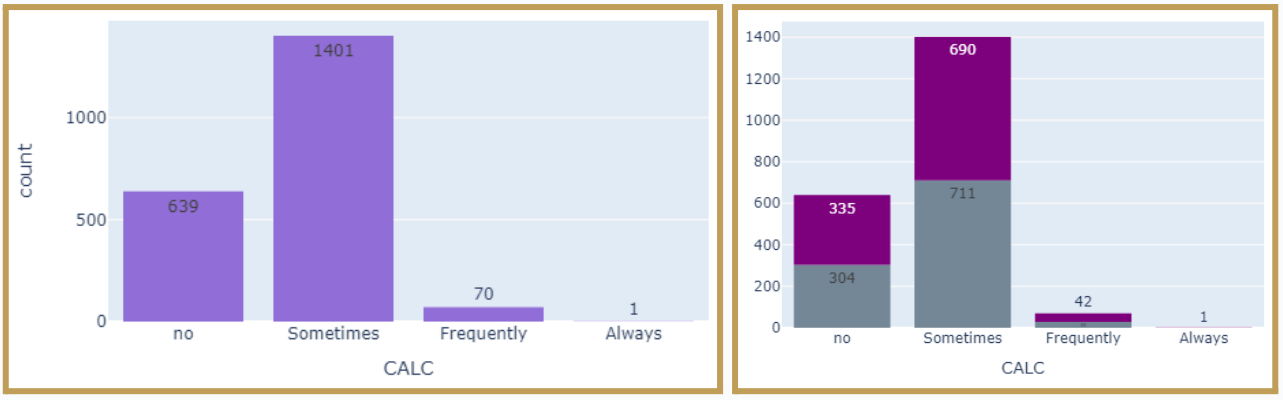


Gráfico 7 - Distribución por CALC: total y separado por género.

La mayor cantidad de respuestas son para 'no' y 'sometimes', con el 30% y 66% respectivamente. Mientras que el 4% restante lo consume frecuentemente o siempre.

7.7 - MTRANS: Transporte utilizado.

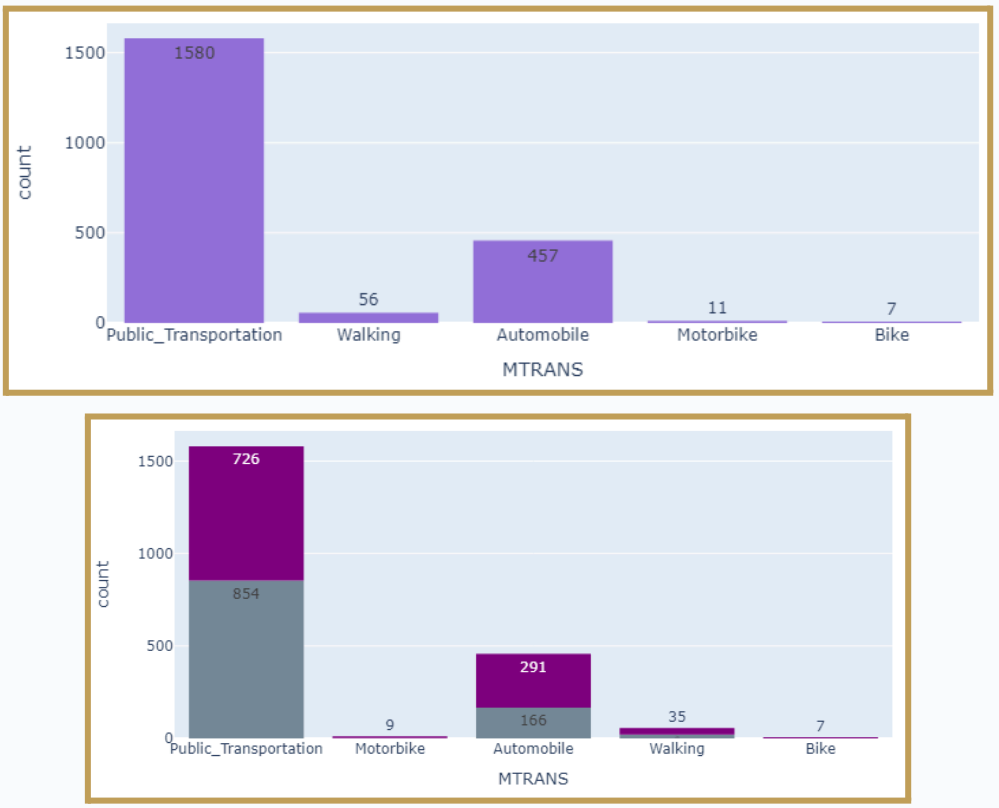


Gráfico 8 - Distribución por MTRANS: total y separado por género.

El 75% utiliza transporte público, el 22% utiliza moto o auto, y solo el 3% camina o anda en bicicleta. Dentro de este último porcentaje nos encontramos que el 63% son varones.

7.8 - NObeyesdad: Nivel de obesidad.

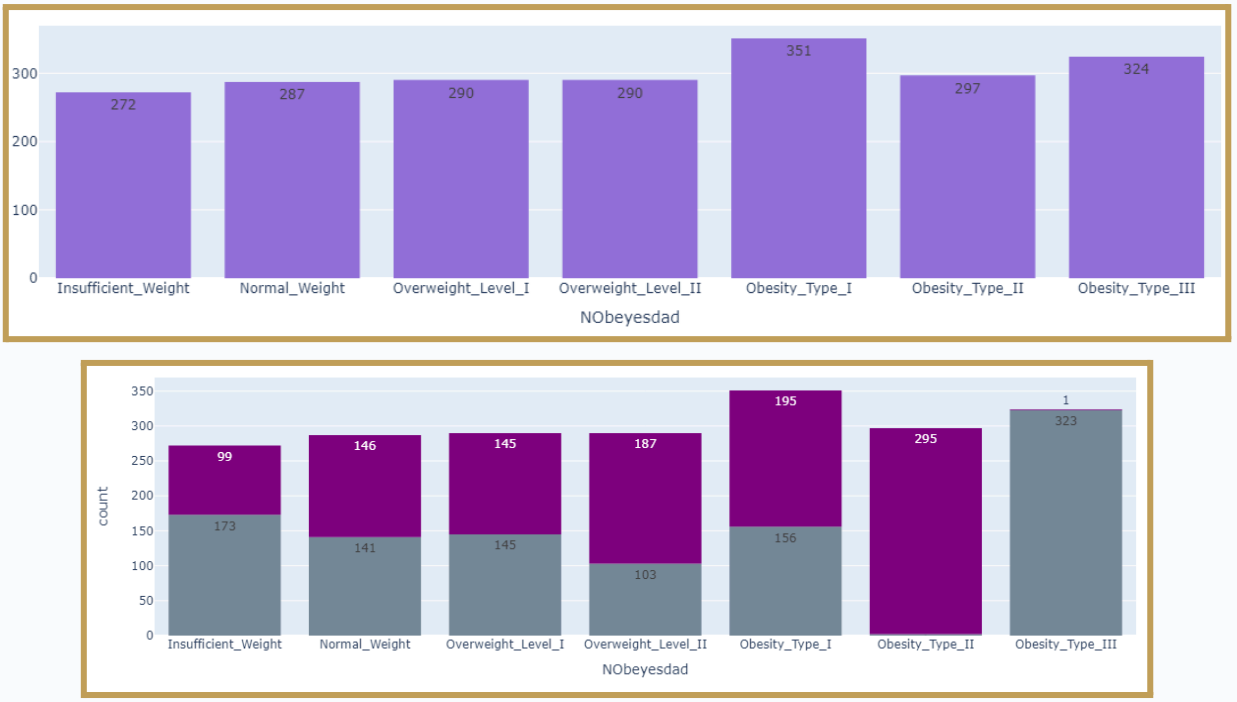


Gráfico 9 - Distribución por NObeyesdad: total y separado por género.

Se puede considerar que la muestra se divide parejamente, siendo 'Insufficient_Weight' el que menos casos posee con el 13%, y 'Obesity_Type_I' el de mayor cantidad con el 17%. Lo que se puede observar también es que 'Obesity_Type_II' es casi por completo de varones, mientras que 'Obesity_Type_III' es casi completo de mujeres.

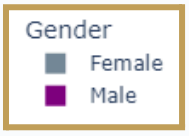
7.9 - Cat IMC: Categorías de IMC.

Gráfico 10 - Distribución por cat_IMC: total y separado por género.

También se puede considerar una distribución pareja respecto al total, la categoría que más casos tiene es 'obesidad_I' con el 18%. Y como particular se nota también que 'obesidad_III' está compuesto casi exclusivamente por mujeres.

8 - Variables numéricas.

Análisis de las nueve variables numéricas: mínimos, máximos, y gráficos univariados y separados por el género con la siguiente identificación:



8.1 - Age: Edad.

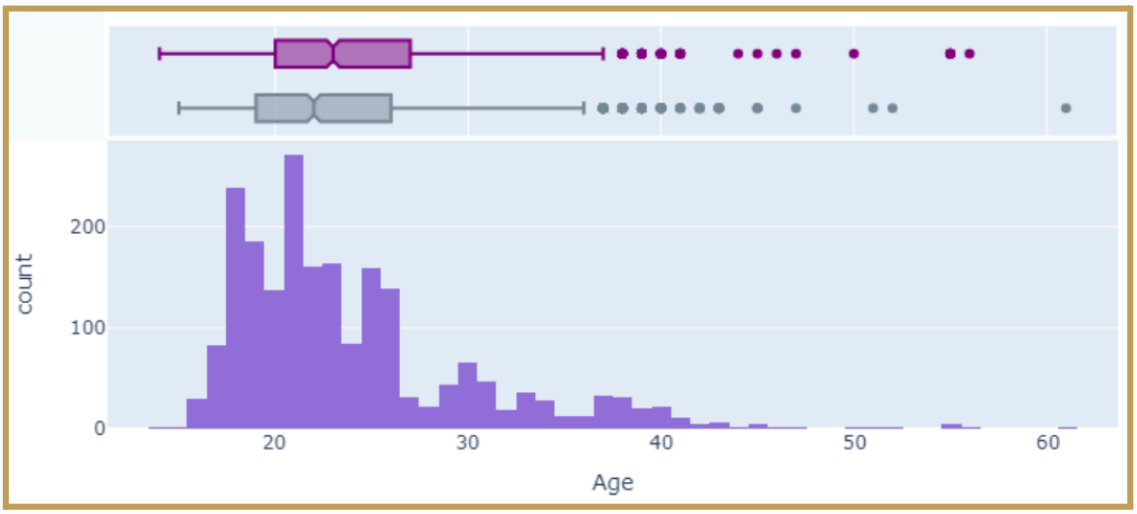


Gráfico 11 - Distribución por Age: total y separado por género.

Valores dentro del rango: 14 a 61 años.

La franja etaria más consultada está entre los 18 y los 27 años, disminuyendo considerablemente luego de los 40.

8.2 - Height: Altura.

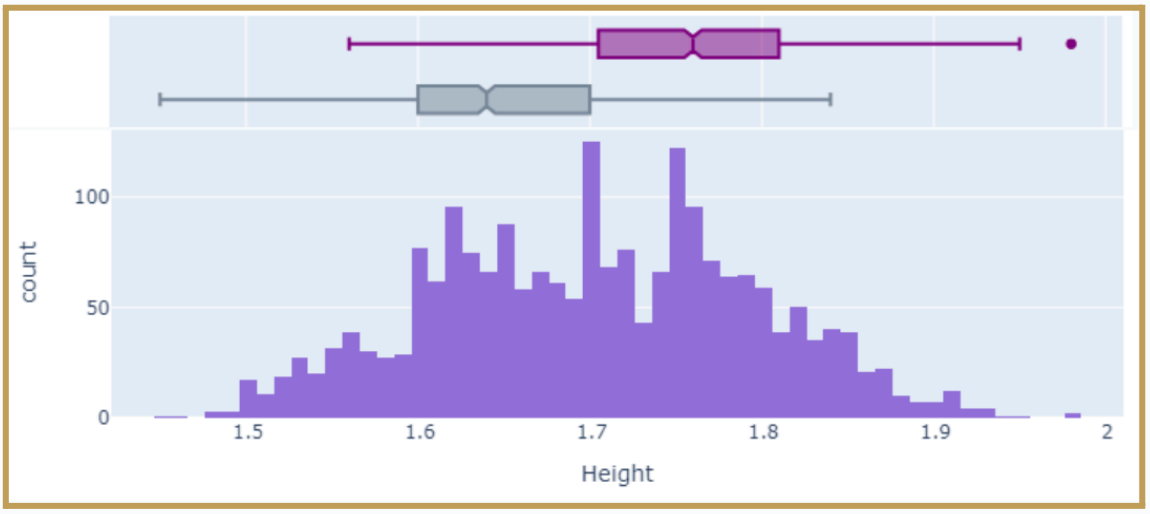


Gráfico 12 - Distribución por Height: total y separado por género.

Valores dentro del rango: 1.45 a 1.98 metros.

Las alturas que poseen mayores consultas se radican, aproximadamente, entre 1.60 y 1.80 metros. Siendo los varones considerablemente más altos respecto a las mujeres.

8.3 - Weight: Peso.

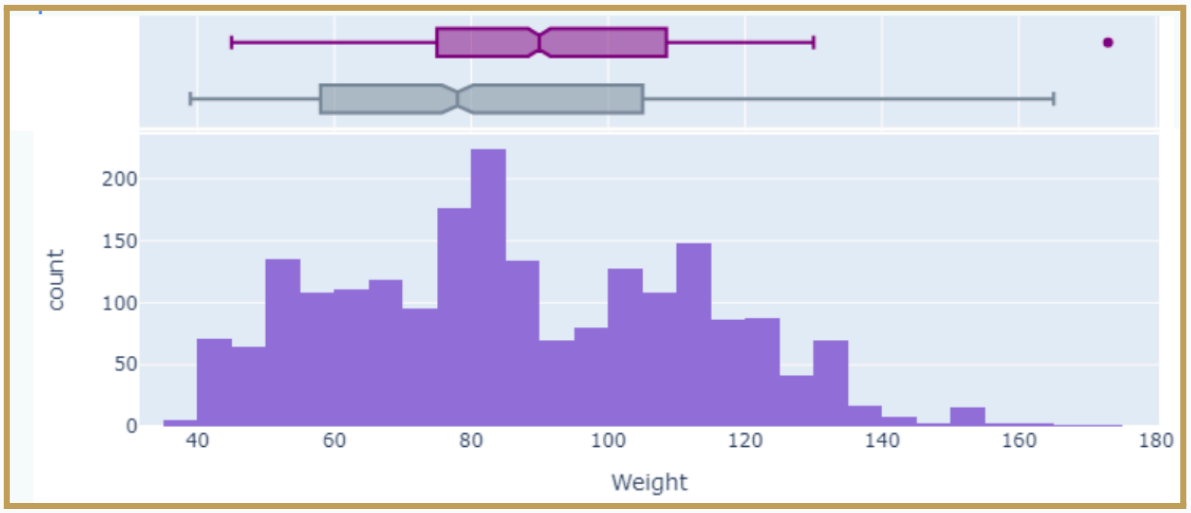


Gráfico 13 - Distribución por Weight: total y separado por género.

Valores dentro del rango: 39.00 a 173.00 kilos.

Hay una amplia variedad de pesos, disminuyendo los casos cuando se superan los 140kg aproximadamente.

8.4 - FCVC: Frecuencia de consumo de vegetales por semana.

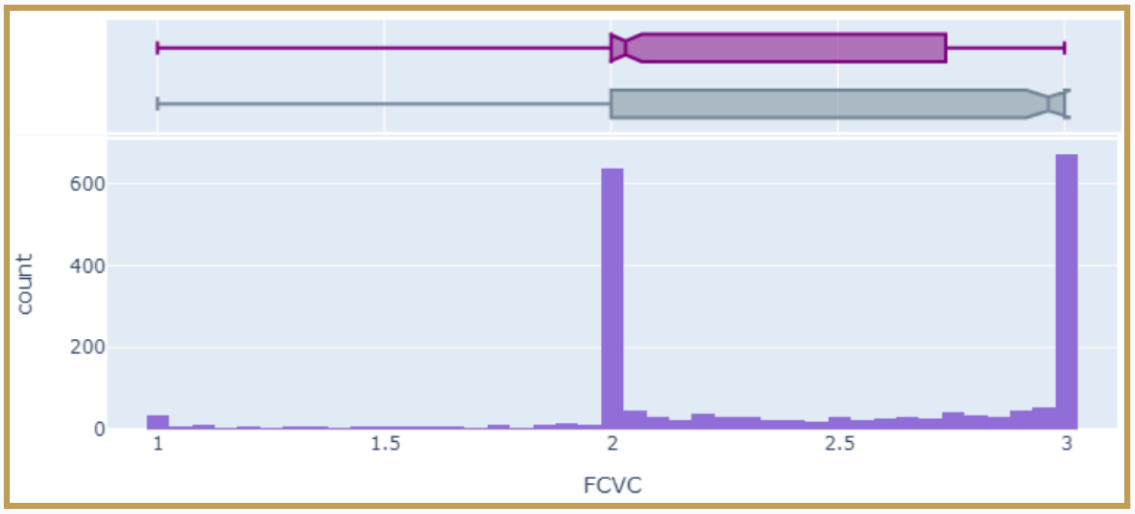


Gráfico 14 - Distribución por FCVC: total y separado por género.

Valores dentro del rango: 1.00 a 3.00

Adoptando que 1 = nunca, 2 = a veces, 3= siempre, se puede decir que gran parte consume vegetales semanalmente.

8.5 - NCP: Número de comidas principales por día.

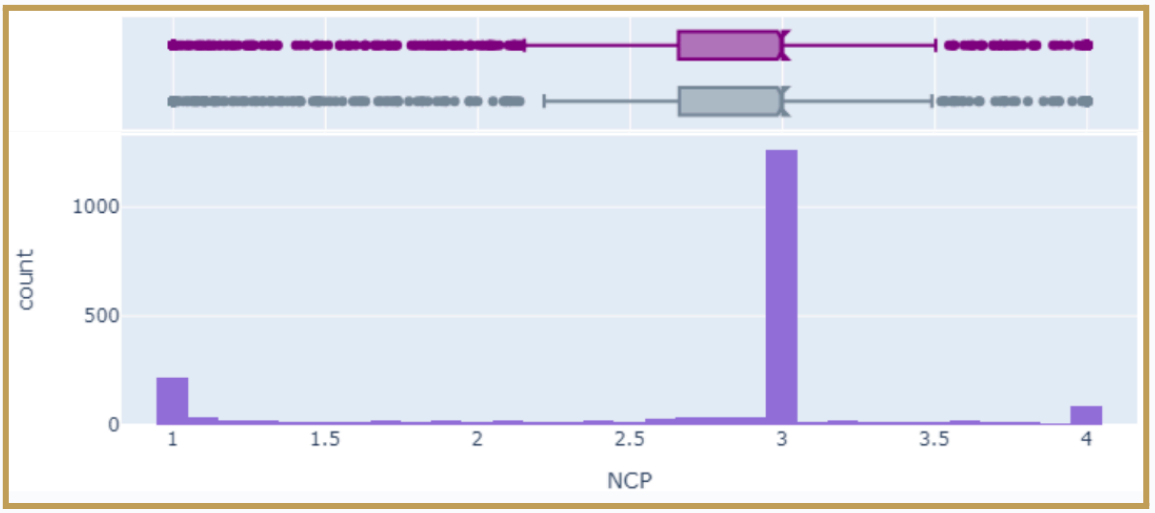


Gráfico 15 - Distribución por NCP: total y separado por género.

Valores dentro del rango: 1.00 a 4.00

La mayoría realiza 3 comidas principales, mientras que los extremos de 1 o 4 comidas también sobresalen en medidas muy inferiores.

8.6 - CH2O: Consumo de agua diario.

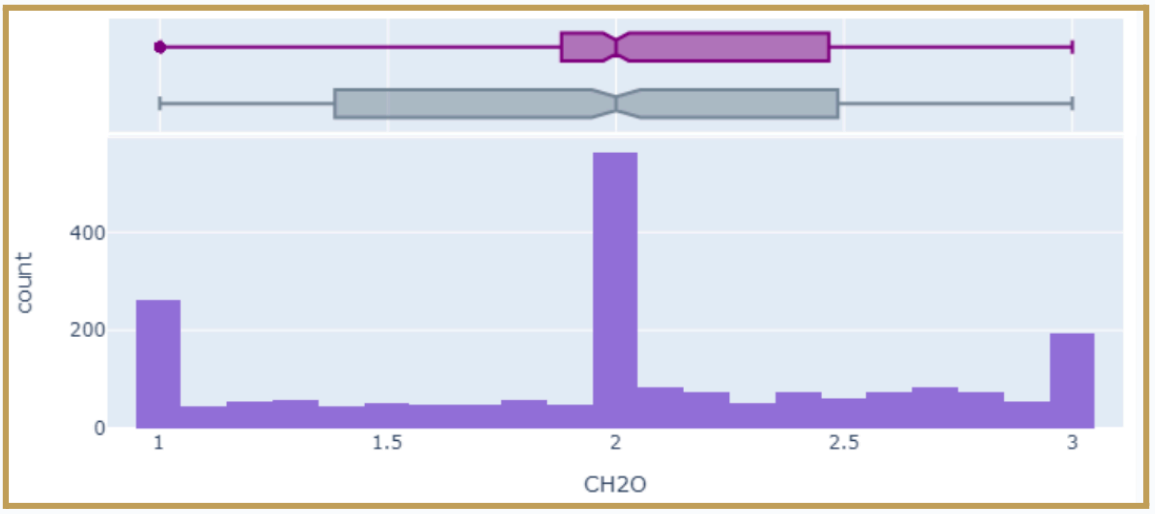


Gráfico 16 - Distribución por CH2O: total y separado por género.

Valores dentro del rango: 1.00 a 3.00

La mayoría de las personas tiene un consumo medio de agua.

8.7 - FAF: Frecuencia de actividad física por semana.

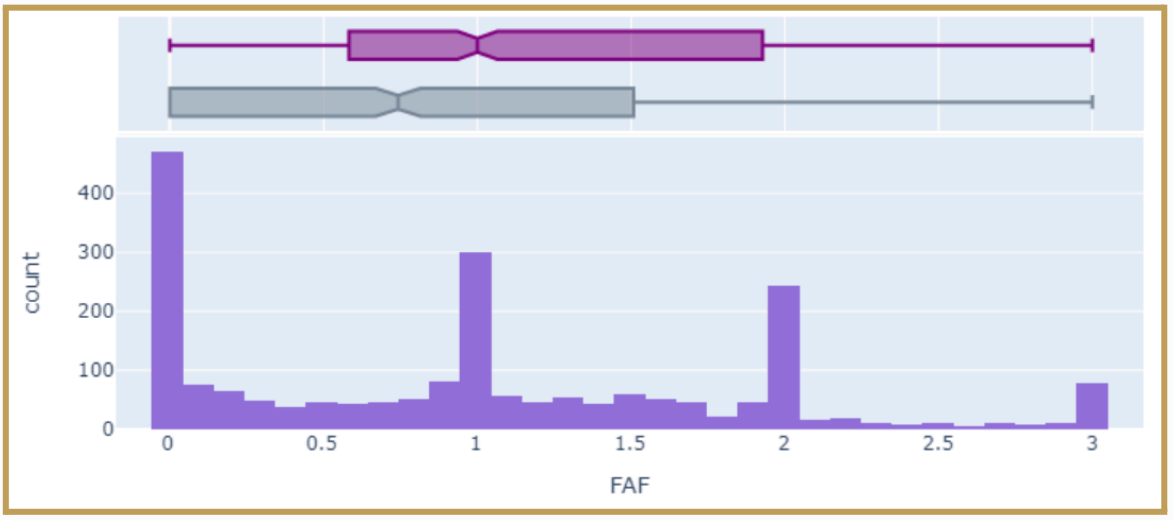


Gráfico 17 - Distribución por FAF: total y separado por género.

Valores dentro del rango: 0.00 a 3.00

Adoptamos que 0 = ninguna, 1 = baja, 2= moderada, 3 = alta, y se nota que un alto porcentaje no realiza actividad física, siendo mayor para los varones.

8.8 - TUE: Tiempo de uso de dispositivos tecnológicos por semana.

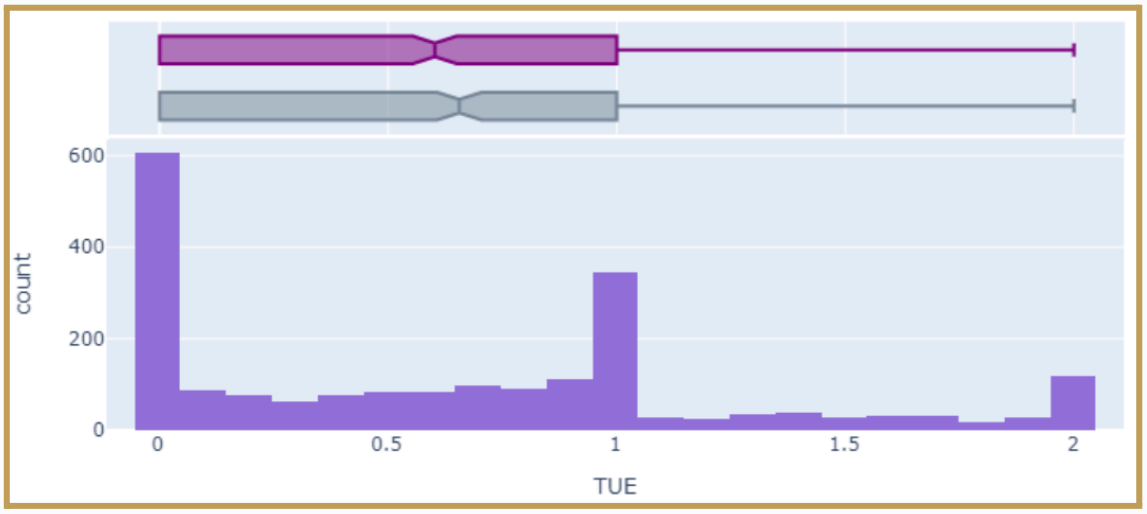


Gráfico 18 - Distribución por TUE: total y separado por género.

Valores dentro del rango: 0.00 a 2.00

Adoptamos que 0 = bajo, 1 = medio, 2= alto, y se observa que la mayor cantidad de personas le da un uso entre medio y bajo a los dispositivos.

8.9 - IMC: Índice de Masa Corporal.

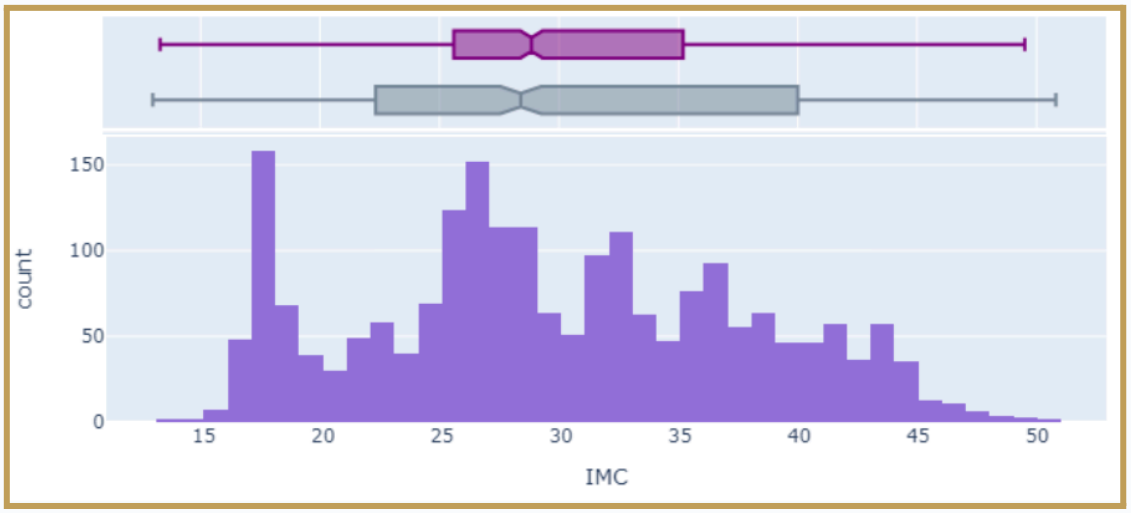


Gráfico 19 - Distribución por IMC: total y separado por género.

Valores dentro del rango: 13.00 a 50.80

La distribución no es pareja, pero se analizará mejor más adelante cuando se realicen estudios con mayor profundidad.

9 - ¿Cómo se relaciona el peso con la altura?.

Se calcula el Coeficiente Pearson para numérico vs numérico, y se obtiene aproximadamente un valor de 0.4623, que al encontrarse próximo a 0.5, se considera que las variables tienen una correlación positiva.

Además se puede observar la siguiente gráfica:

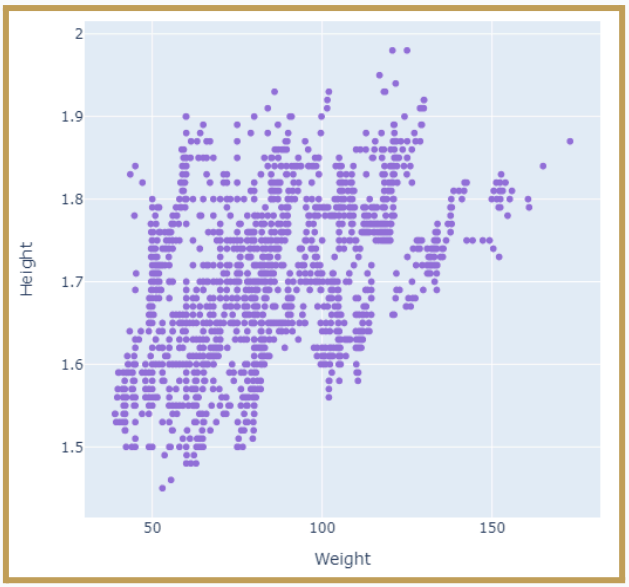


Gráfico 20 - Relación entre Weight vs Height.

10 - ¿La población consultada tiene mayoritariamente algún grado de sobrepeso u obesidad?.

Hipótesis nula (H_0): plantea que el IMC promedio para toda la población μ es el mismo que el IMC máximo dentro de la franja que se considera normal, siendo $\mu_0 = 24.9$. Se quiere probar si H_0 es incorrecto, es decir, si $\mu \neq \mu_0$.

Hipótesis alternativa (H_a): considera que mayoritariamente la población tiene algún grado de sobrepeso, es decir, $\mu > \mu_0$

Se toma un nivel de significancia (α) de 0.05, lo que significa que se está dispuesto a aceptar un 5% de probabilidad de estar equivocados al rechazar la hipótesis nula.

Aplicando el código correspondiente se obtiene P-valor≅$4.64e^{-143}$, lo cual es mucho menor que el α adoptado, por lo que se considera que es H_0 falso y se rechaza.

Se concluye que existe evidencia estadística para la alternativa H_a y que el IMC de la población estudiada está por arriba del IMC considerado normal.

A continuación se encuentra la gráfica indicando la media anterior equivalente a $\mu_0 = 24.9$, y la moda, la media y la mediana reales, donde se observa la asimetría positiva.

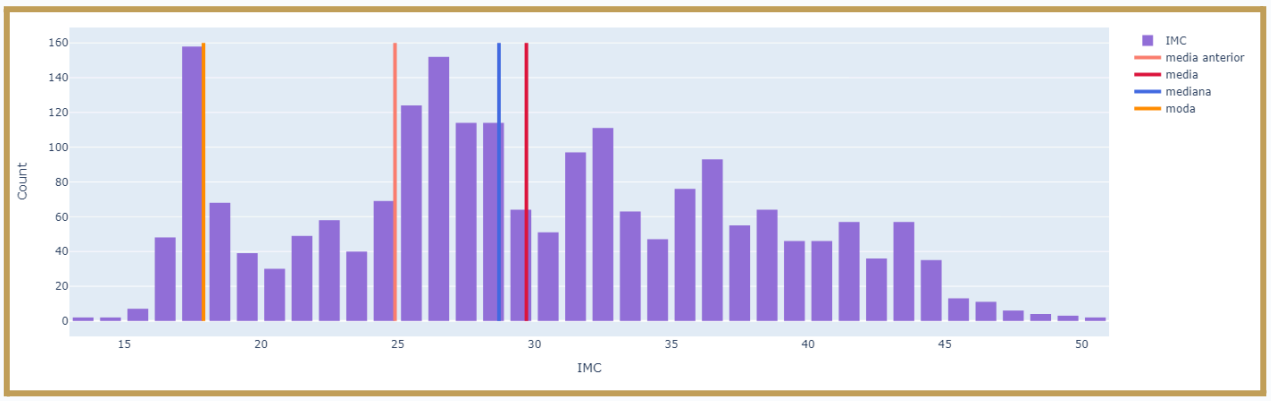


Gráfico 21 - Distribución por IMC indicando métricas estadísticas.

11 - ¿Influyen los niveles familiares en el sobrepeso?.

Hipótesis nula (H_0): plantea que $\mu_1 = \mu_2$, siendo μ_1 la media del IMC de la población con antecedentes familiares de sobrepeso, y μ_2 la media del IMC de la población sin antecedentes.

Hipótesis alternativa (H_a): plantea que $\mu_1 \neq \mu_2$, en oposición a H_0.

Se toma un nivel de significancia (α) de 0.05, lo que significa que se está dispuesto a aceptar un 5% de probabilidad de estar equivocados al rechazar la hipótesis nula.

Aplicando el código correspondiente se obtiene P-valor≅$1.01e^{-181}$, lo cual es mucho menor que el α adoptado, por lo que se considera que es H_0 falso y se rechaza.

Se concluye que existe evidencia estadística para la alternativa H_a y que el IMC de ambas poblaciones estudiadas es diferente.

En la gráfica se observa que las personas con antecedentes tienen un IMC más elevado.

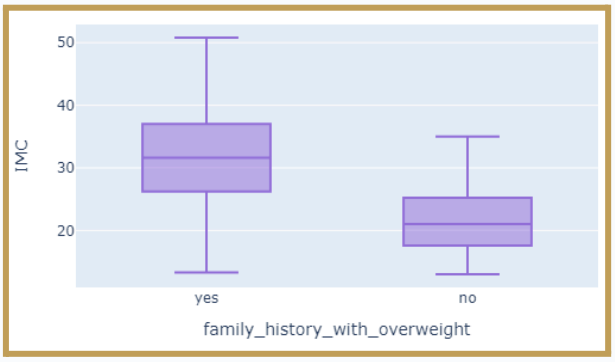
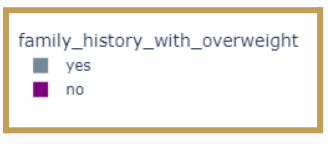


Gráfico 22 - IMC respecto a los antecedentes familiares de sobrepeso.

A su vez, se analiza respecto a la categorización original, contemplando que:



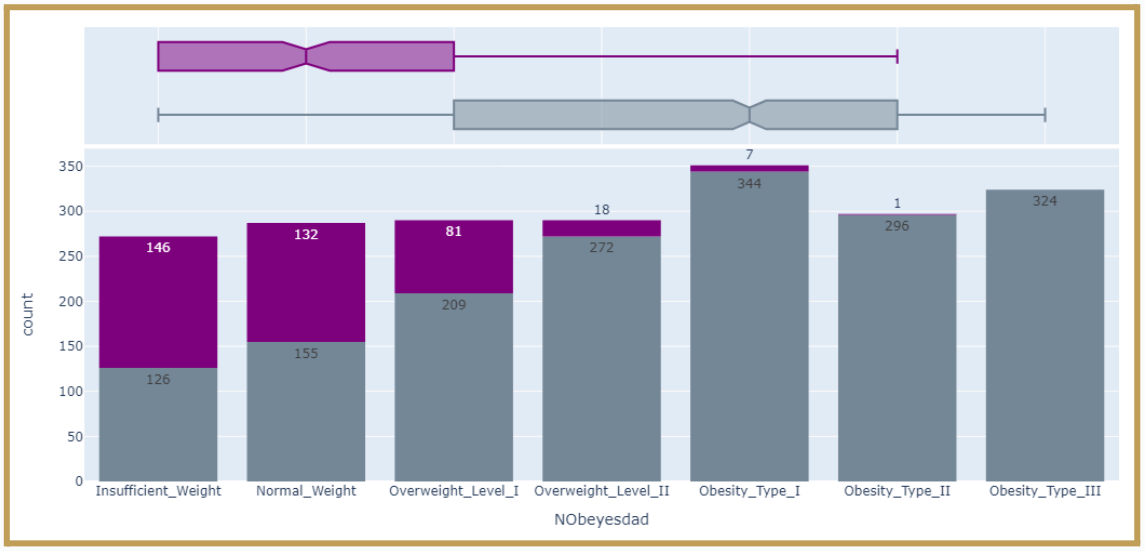


Gráfico 23 - NObeyesdad respecto a los antecedentes familiares de sobrepeso.

Se observa que las personas que poseen algún grado de sobrepeso u obesidad, casi en su totalidad, poseen antecedentes familiares. Mientras que en los dos primeros casos se dan de forma más pareja los casos con y sin antecedentes.

12 - ¿Cómo se distribuyen las distintas categorías de NObeyesdad y de cat_IMC respecto a las edades estudiadas?.

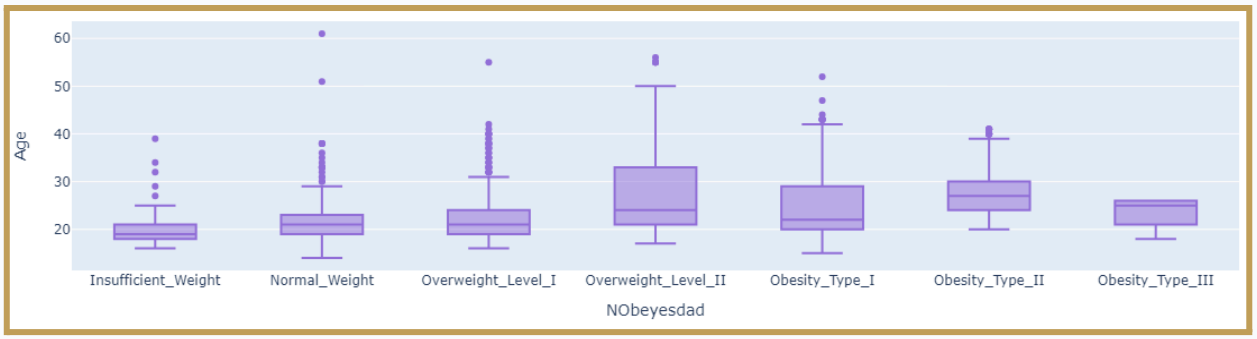


Gráfico 24 - Niveles de NObeyesdad respecto a la edad.

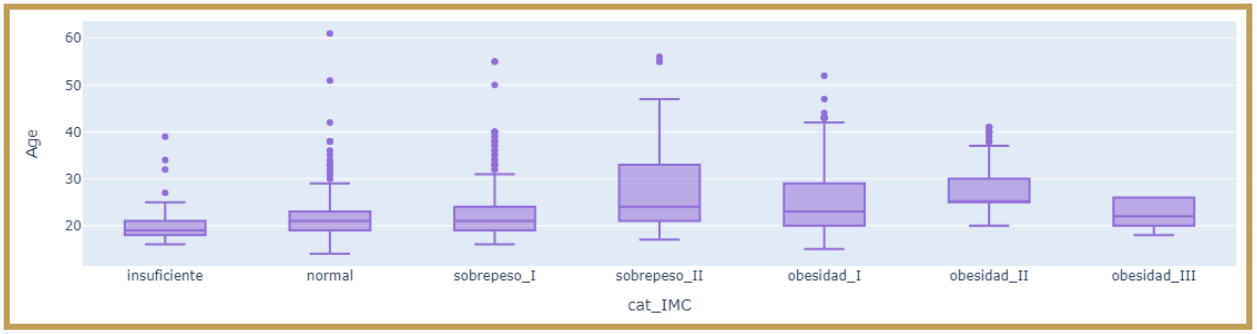


Gráfico 25 - Niveles de cat_IMC respecto a la edad.

Para ambos casos, las categorías de los extremos se desarrollan mayoritariamente en edades más tempranas comparadas con las demás.

13 - Análisis de la relación entre el medio de transporte más utilizado y el peso, diferenciando si consumen tabaco o no.

Partiendo de la distinción sobre smoke:

Gráfico 26 - Relación entre el tipo de transporte más utilizado y el peso.

Las personas consultadas que utilizan mayoritariamente transporte público o automóviles son las de peso más elevado, y las que porcentualmente más consumen tabaco respecto a las que no lo hacen.

14 - Normalización.

Se realizó el proceso de normalización distinguiendo entre los distintos tipos de variables para el uso de las correspondientes librerías:

- Categóricas binarias: Label Encoder.
- Multiclases ordinales: Ordinal Encoder.
- Multiclases nominales: Get Dummies.

15 - ¿Existe relación entre el IMC y los campos sin importar el peso y la altura?.

Aclaración: no se considerarán Height, Weight, NObeyesdad y cat_IMC porque tienen relación directa con el IMC, y se quiere conocer como actúan las otras variables.

Se utilizó el modelo Linear Regression, y se obtuvo un R cuadrado aproximadamente igual a 0.4653, quedando muy por debajo del 0.8 que lo cataloga como alto.

Además se completó el estudio con un análisis gráfico, y estas son las condiciones:

- **Linealidad**: La relación entre las variables independientes y la variable dependiente debe ser lineal.

- **Independencia de errores**: Los errores deben ser independientes, esto significa que el error asociado con una observación no debe influir en el error de otra observación.
- **Homocedasticidad**: Los errores deben tener varianza constante, es decir, que la dispersión de los errores debe ser la misma para todos los valores de las variables independientes.
- **Normalidad de la distribución de errores**: Los errores deben seguir una distribución normal.
- **Ausencia de valores atípicos influyentes**: Verificar la presencia de valores atípicos., porque estos pueden sesgar las estimaciones de los coeficientes de regresión.

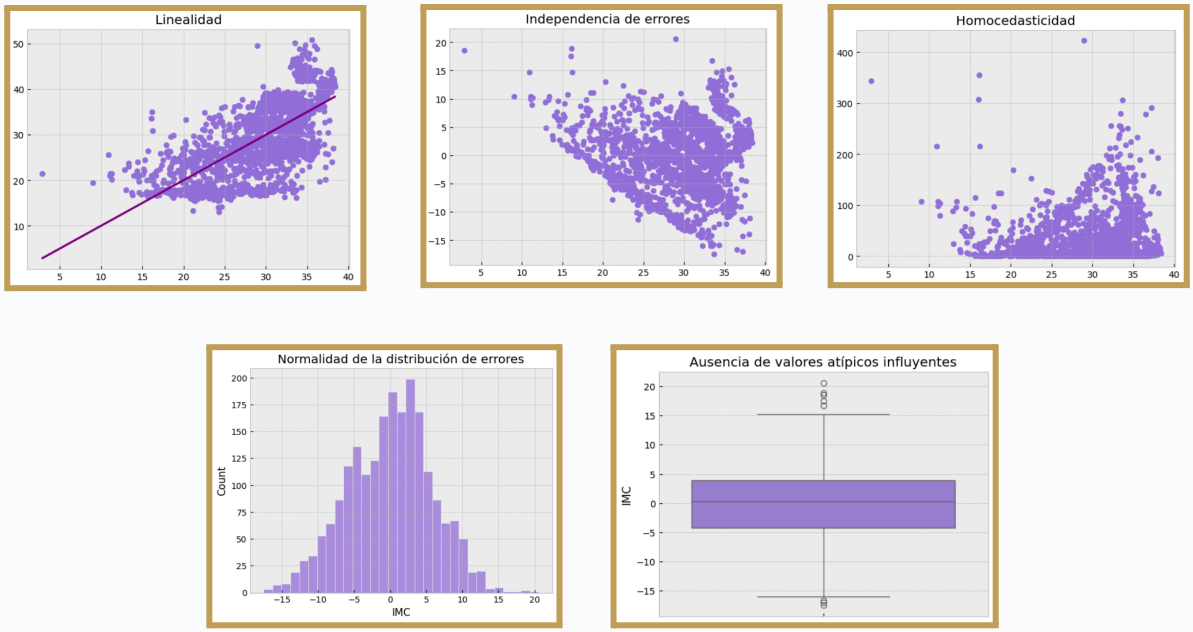


Gráfico 27 - Análisis Regresión Lineal.

Observaciones:

- **Linealidad**: Hay una leve tendencia lineal, pero la dispersión es grande entre los puntos y la recta de ajuste, el modelo no es bueno.
- **Independencia de errores**: Los puntos no se encuentran del todo dispersos.
- **Homocedasticidad**: Los puntos no se concentran del todo en una línea recta inferior, sino que están dispersos en algunos sectores.
- **Normalidad de la distribución de errores**: La distribución se podría considerar normal.
- **Ausencia de valores atípicos influyentes**: Posee valores atípicos, y en algunos casos alejados.

Como conclusión se puede mencionar que no sirve utilizar un modelo de Regresión lineal dado que el R cuadrado es bajo, y 4 de las 5 gráficas no arrojan resultados aptos, y más importante aún, no es posible relacionar el valor del IMC con el resto de las variables.

16 - Los modelos de clasificación y las métricas.

En primera instancia se eliminan los campos comprometidos, definimos las variables dependiente e independiente, se separan los datos en training y testing en una relación 80-20, y se escalan las variables numéricas con Standard Scaler.

Los modelos de clasificación que se utilizarán serán:

- Árbol de decisión.
- k-NN.
- Regresión logística.
- Random Forest.
- XGBoost.
- CatBoost.
- LightGBM.

Y se calcularán las siguientes métricas:

- **Confusion Matrix (Matriz de confusión)**: Tabla que muestra los Verdaderos Positivos (TP), Falsos Positivos (FP), Verdaderos Negativos (TN) y Falsos Negativos (FN). Proporciona una imagen clara del rendimiento del modelo más allá de la simple exactitud.
- **Accuracy (Exactitud)**: Proporciona la fracción de predicciones correctas entre el total de casos. Es útil cuando las clases están balanceadas, pero puede ser engañosa en conjuntos de datos desbalanceados.

$$Accuracy = \frac{Nro\ de\ predicciones\ correctas}{Total\ de\ predicciones}$$

- **Precision (Precisión)**: Mide la proporción de predicciones positivas que fueron realmente correctas. Es importante cuando el costo de un Falso Positivo es alto.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensibilidad)**: Mide la proporción de casos positivos reales que fueron identificados correctamente. Es crucial cuando es esencial detectar todos los casos positivos (por ejemplo, en diagnósticos médicos).

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**: Combina la precisión y la sensibilidad en una sola métrica mediante su media armónica. Es útil cuando se busca un equilibrio entre Precisión y Sensibilidad.

$$F1 - Score = 2x\frac{Precision\ x\ Recall}{Precision + Recall}$$

17 - ¿Cómo es la clasificación respecto a cat_IMC?.

Aclaración: se elimina el campo IMC porque tiene relación directa, pero se dejan Height y Weight, que aunque también su relación es directa nos va a servir para poder compararlo con la clasificación de NObeyesdad.

Vale recordar que ya se verificó que no existe relación entre IMC y las variables que no son parte de su formulación.

17.1 - Árbol de decisión.

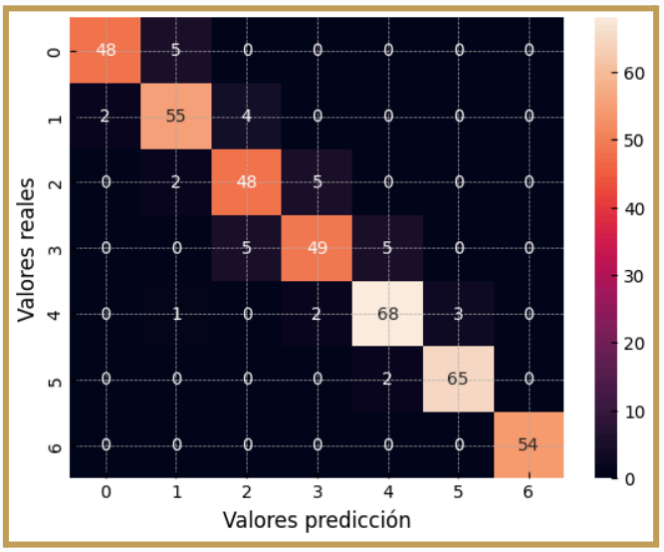


Gráfico 28 - Confusion matrix - Árbol de decisión.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	1.0	**0.9148936170212766**
Presicion	1.0	0.9160957365402301
Recall	1.0	0.9142290916535132
F1-Score	1.0	0.9148810028780933

Tabla 4 - Métricas Árbol de decisión.

17.2 - kNN.

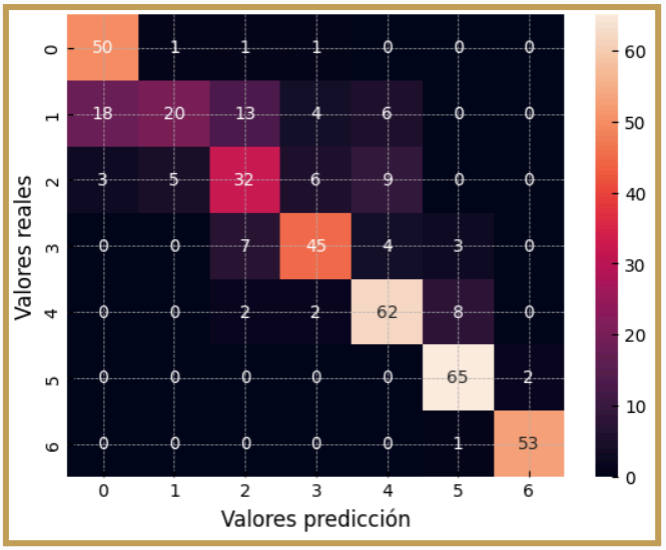


Gráfico 29 - Confusion matrix - kNN.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.8518957345971564	**0.7730496453900709**
Presicion	0.8511314521172842	0.7720515255263978
Recall	0.852184716967804	0.772180528307105
F1-Score	0.8432751397705319	0.7560750744132382

Tabla 5 - Métricas kNN.

17.3 - Regresión logística.

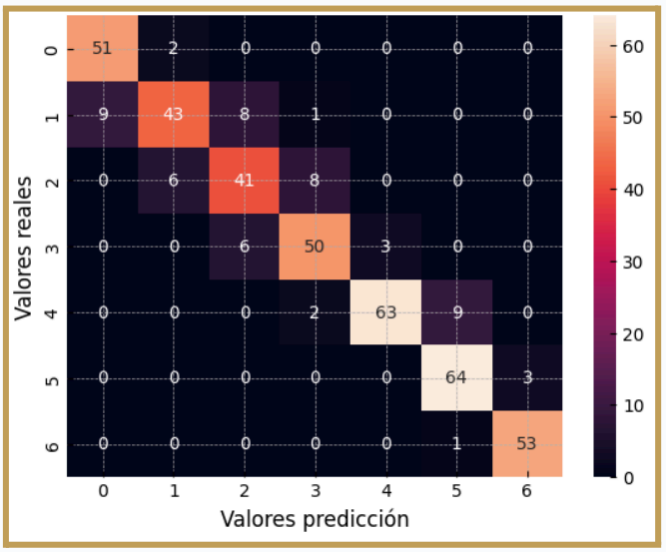


Gráfico 30 - Confusion matrix - Regresión logística.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.8690758293838863	**0.8628841607565012**
Presicion	0.8665915384459123	0.8605861174775625
Recall	0.8706409862490592	0.8640215813904742
F1-Score	0.8673885179796291	0.8601053815682483

Tabla 6 - Métricas Regresión logística.

17.4 - Random Forest.

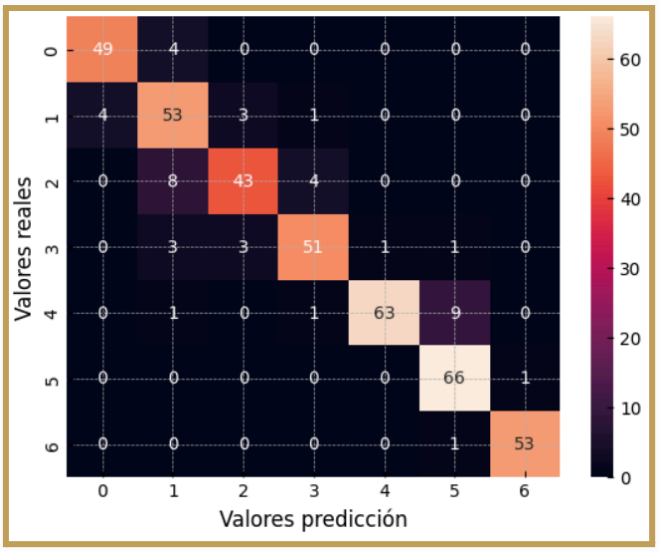


Gráfico 31 - Confusion matrix - Random Forest.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9034360189573459	**0.8936170212765957**
Presicion	0.9069473430706998	0.8982759207219347
Recall	0.902905689726224	0.8939304545829841
F1-Score	0.9031933997686752	0.8939054236330127

Tabla 7 - Métricas Random Forest.

17.5 - XGBoost.

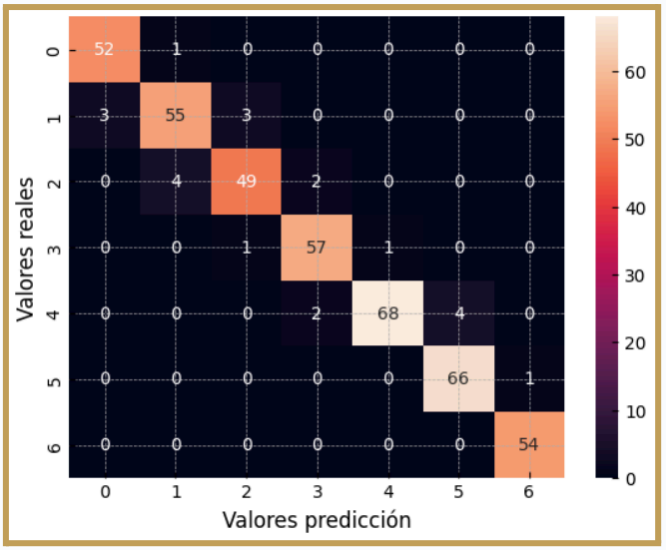


Gráfico 32 - Confusion matrix - XGBoost.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	1.0	**0.9479905437352246**
Presicion	1.0	0.9473226163669054
Recall	1.0	0.9491108216204184
F1-Score	1.0	0.9478342240312374

Tabla 8 - Métricas XGBoost.

17.6 - CatBoost.

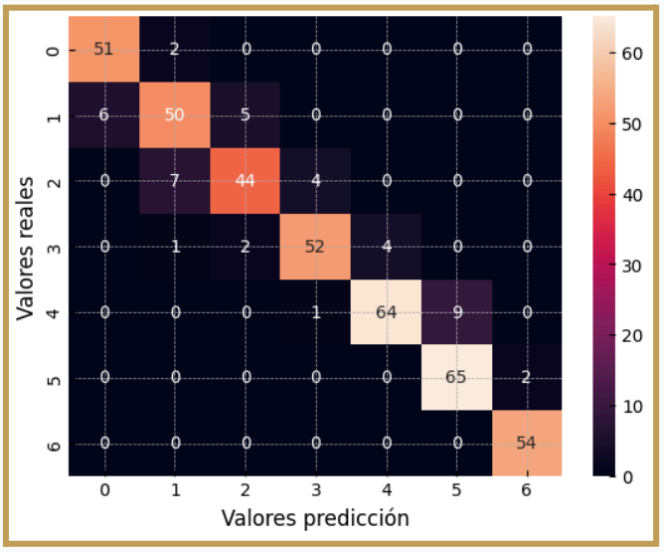


Gráfico 33 - Confusion matrix - CatBoost.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9146919431279621	**0.8983451536643026**
Presicion	0.9148785674382741	0.8981337912120751
Recall	0.9159439670164903	0.8997580475557908
F1-Score	0.9151605371412306	0.897953122824703

Tabla 9 - Métricas CatBoost.

17.7 - LightGBM.

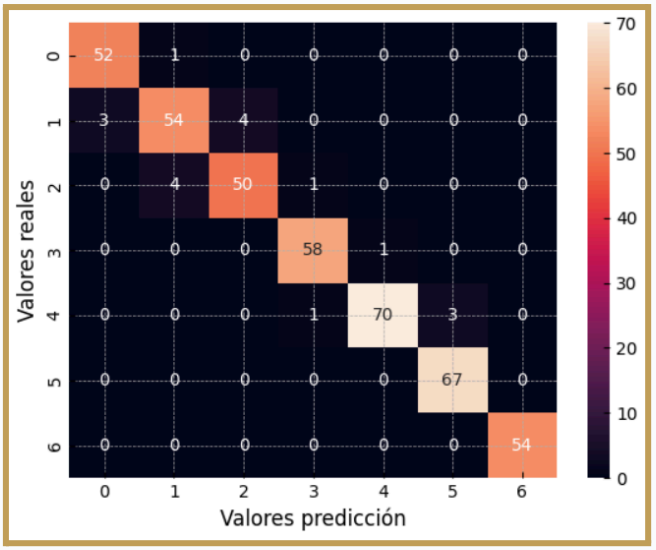


Gráfico 34 - Confusion matrix - LightGBM

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9146919431279621	**0.9574468085106383**
Presicion	0.9148785674382741	0.9566228179194111
Recall	0.9159439670164903	0.9577808113722178
F1-Score	0.9151605371412306	0.9569719289643189

Tabla 10 - Métricas LightGBM.

17.8 - Resumen.

Como las clases están balanceadas, se compararon los Accuracy de los conjuntos de testeo de los distintos modelos, y se observa que en general se obtienen buenos resultados, pero que los mejores son el Árbol de decisión, XGBoost y LightGBM.

18 - ¿Cómo es la clasificación respecto a NObeyesdad?.

Se elimina cat_IMC para que no afecte el resultado.

18.1 - Árbol de decisión.

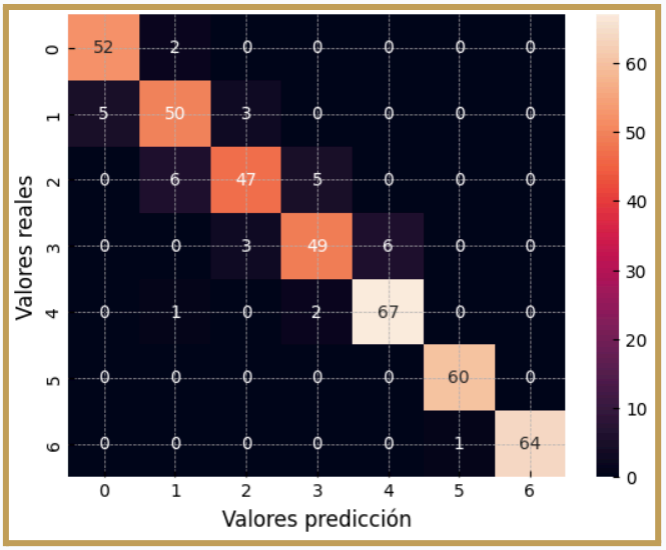


Gráfico 35 - Confusion matrix - Árbol de decisión.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	1.0	**0.9196217494089834**
Presicion	1.0	0.9175636511797644
Recall	1.0	0.9174232262902214
F1-Score	1.0	0.9170257567943114

Tabla 11 - Métricas Árbol de decisión.

18.2 - kNN.

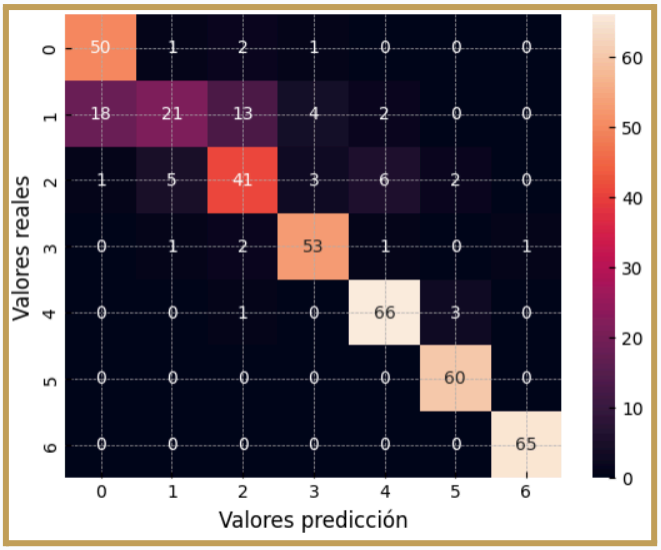


Gráfico 36 - Confusion matrix - kNN.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.8755924170616114	**0.8416075650118203**
Presicion	0.8767105478840298	0.8323329717626441
Recall	0.87138748068444	0.835934527067532
F1-Score	0.8622576054907717	0.8222432094686019

Tabla 12 - Métricas kNN.

18.3 - Regresión logística.

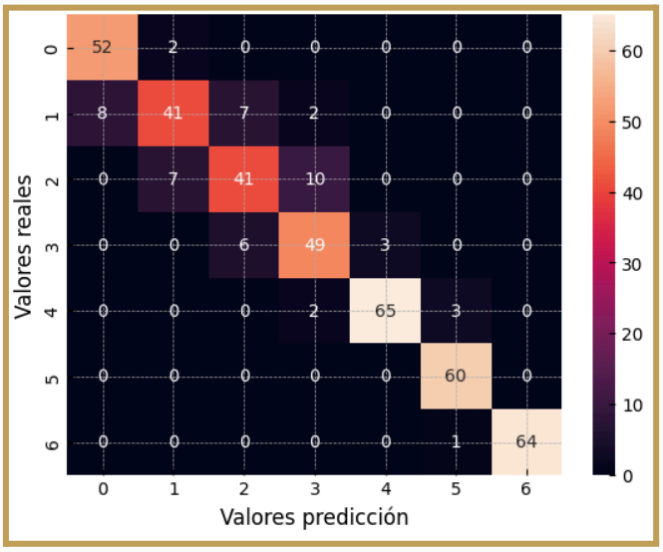


Gráfico 37 - Confusion matrix - Regresión logística.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9046208530805687	**0.8794326241134752**
Presicion	0.8738694366635543	0.8738694366635543
Recall	0.9028091531575234	0.8763957808292784
F1-Score	0.9015319561707937	0.8736598795050036

Tabla 13 - Métricas Regresión logística.

18.4 - Random Forest.

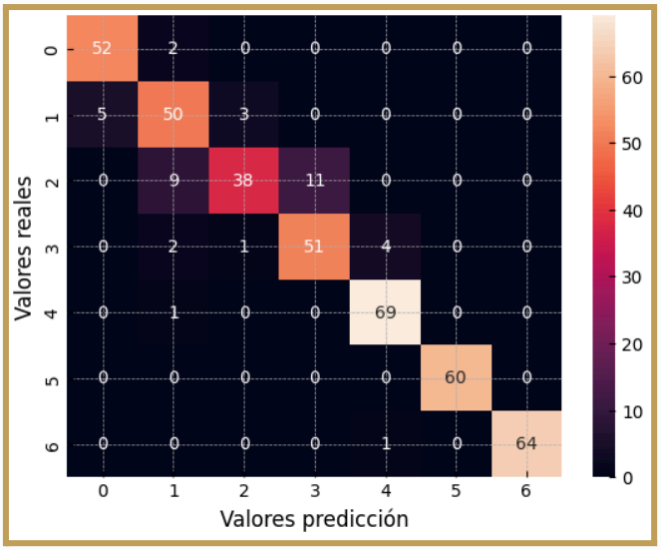


Gráfico 38 - Confusion matrix - Random Forest.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9419431279620853	**0.9078014184397163**
Presicion	0.943019318418907	0.9076150977300019
Recall	0.9403216341526445	0.9042634796329378
F1-Score	0.9408003226848726	0.9024557804904736

Tabla 14 - Métricas Random Forest.

18.5 - XGBoost.

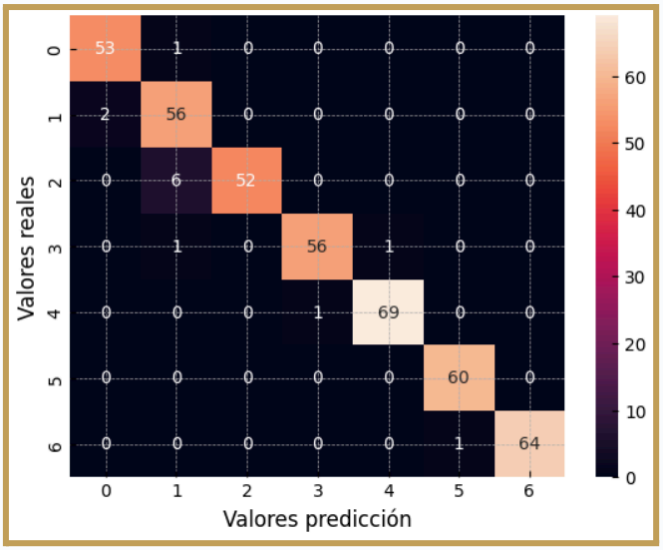


Gráfico 39 - Confusion matrix - XGBoost.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	1.0	**0.9692671394799054**
Presicion	1.0	0.9700590495826538
Recall	1.0	0.9684853369582435
F1-Score	1.0	0.9685107607083011

Tabla 15 - Métricas XGBoost.

18.6 - CatBoost.

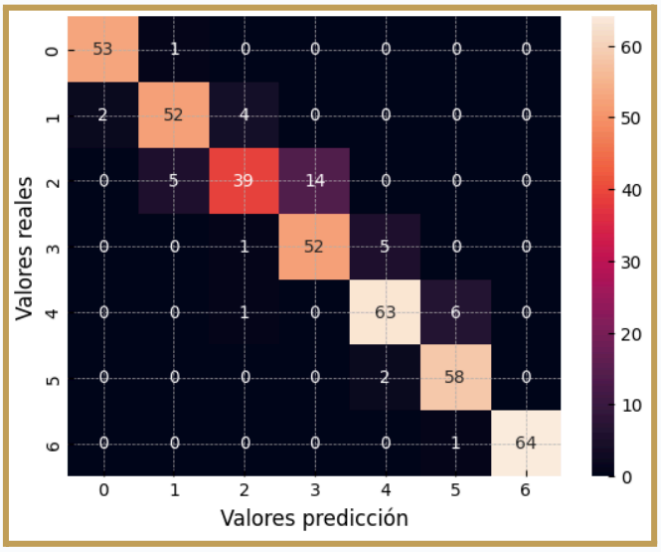


Gráfico 40 - Confusion matrix - CatBoost.

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9360189573459715	**0.900709219858156**
Presicion	0.9353699829467118	0.9010058906610633
Recall	0.9345976825327178	0.8997543963061204
F1-Score	0.9345624339985984	0.8978954401701474

Tabla 16 - Métricas CatBoost.

18.7 - LightGBM.

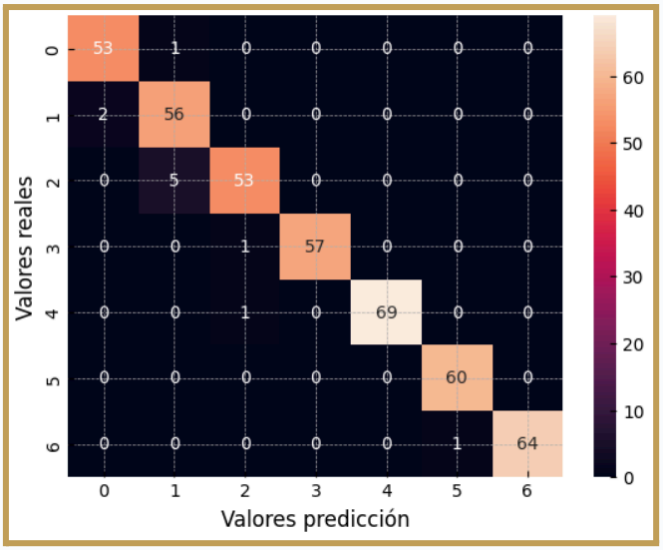


Gráfico 41 - Confusion matrix - LightGBM

	Conjunto entrenamiento	Conjunto testeo
Accuracy	0.9360189573459715	**0.9739952718676123**
Presicion	0.9353699829467118	0.9734435844430556
Recall	0.9345976825327178	0.9734114453326276
F1-Score	0.9345624339985984	0.9731367424751438

Tabla 17 - Métricas LightGBM.

18.8 - Resumen.

Al igual que en el caso anterior, como las clases están balanceadas, se compararon los Accuracy de los conjuntos de testeo de los distintos modelos, y se observa que en general se obtienen buenos resultados, pero que los mejores son el Árbol de decisión, XGBoost y LightGBM.

19 - ¿Conviene otra cantidad de agrupamientos?.

Se parte con la eliminación de IMC y cat_IMC para que no afecten el resultado. Y se prueba con los distintos algoritmos:

- **K-means**: algoritmo de clustering (aprendizaje no supervisado) que divide un conjunto de datos en grupos distintos.
- **Agglomerative Clustering**: algoritmo de agrupamiento jerárquico que construye una jerarquía de grupos.
- **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise): algoritmo de agrupamiento basado en la densidad.

Se analizan los resultados que se obtuvieron y se concluye que se podrían considerar menos agrupamientos que los iniciales, pero esto no sería lo ideal para poder generar un estudio más detallado a sabiendas de los problemas de salud que traen aparejados.

20 - Validación de los modelos.

Si se comparan las dos clasificaciones, NObeyesdad y cat_IMC, se puede observar que en lo particular en ambos casos los mejores accuracy se obtuvieron utilizando los mismos modelos, pero en lo general la clasificación original arroja mejores resultados.

	cat_IMC	**NObeyesdad**
Árbol de decisión	0.914894	0.919622
XGBoost	0.947991	0.969267
LightGBM	0.957447	**0.973995**

Tabla 18 - Comparación de modelos.

A raíz del análisis superior se trabajará con LightGBM y la clasificación original:

- **Con cross_val_score**:
  K-Fold Cross-Validation Accuracy: 0.9662
  Stratified K-Fold Cross-Validation Accuracy: 0.9633

- **Con un optimizador**:
  K-Fold Cross-Validation Accuracy: 0.9728
  Stratified K-Fold Cross-Validation Accuracy: 0.9680

- **Semillero**:
  Precisión final (votación de predicciones): 0.9693

Se puede considerar que los resultados obtenidos son buenos, dado que los valores son altos, estando los accuracy por arriba del 0.96, próximos al valor inicial 0.973995.

21 - Factores que tienen mayor y menor relevancia tienen sobre NObeyesdad.

Se hace uso de **Importancia de las Características** (Feature Importance) e **Importancia por Permutación** (Permutation Importance), y se nota que para ambos casos las variables con más peso se repiten, aunque en distinto orden: Weight, Gender, Age, FCVC, Height. De igual forma ocurre para las de menor peso, pudiendo reconocer por ejemplo SMOKE y las distintas categorías de MTRANS.

En función de las **Características Polinomiales** y con el estudio de **Reducción de la dimensionalidad con PCA** se obtiene la siguiente gráfica:

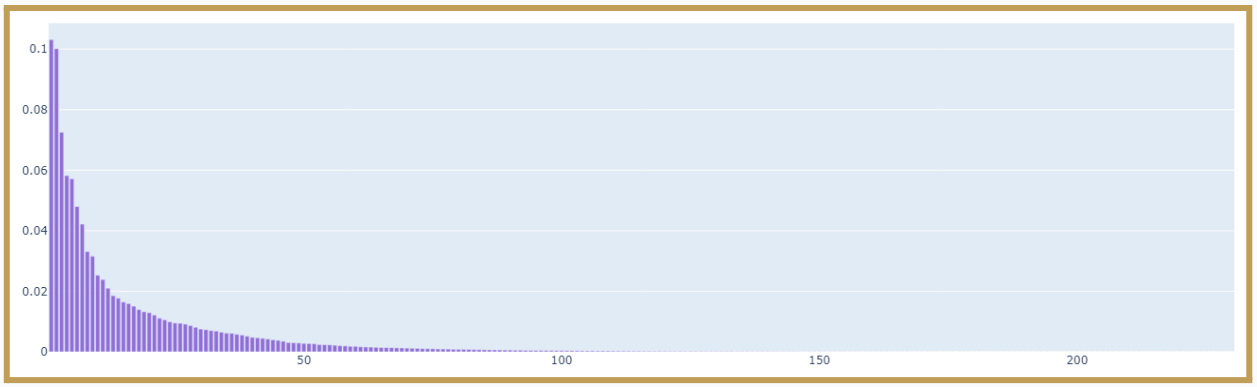


Gráfico 42 - Fracción de varianza que aporta cada componente y la información acumulada.

Se aprecia que con las primeras componentes del gráfico se acumula un alto porcentaje de las varianzas acumuladas, por lo que se puede reducir la dimensionalidad eliminando las restantes. (con los primeros 30 casos se consigue aproximadamente el 80%).

22 - Conclusión.

Respecto a la información original se puede comentar brevemente que las personas encuestadas se dividen parejamente en mujeres y varones, y en su mayoría poseen algún tipo de sobrepeso u obesidad.

Además, si bien no se puede predecir el valor del IMC respecto a los demás campos distintos de Weight y Height, cuando estos datos si se usan los niveles obtenidos por los modelos de clasificación son elevados, aunque levemente inferiores a los de la clasificación original.

A raíz de esto podemos optar por suprimir la inserción de dichos campos (IMC y cat_IMC), y trabajar directamente con NObeyesdad.

El modelo que mejor predice la clasificación es LightGBM, con un accuracy de 0.973995.