Proyecto

May 26, 2025

1 Presentación

Integrantes:
- Antonio Anselmi, Miguel Maximiliano (20200118).

Resultado: No hay datos faltantes en ningun campo, y hay variables cualitativas y cuantitativas

¿Mientras mayor es la cantidad de comidas principales, menor es el control de consumo calórico?

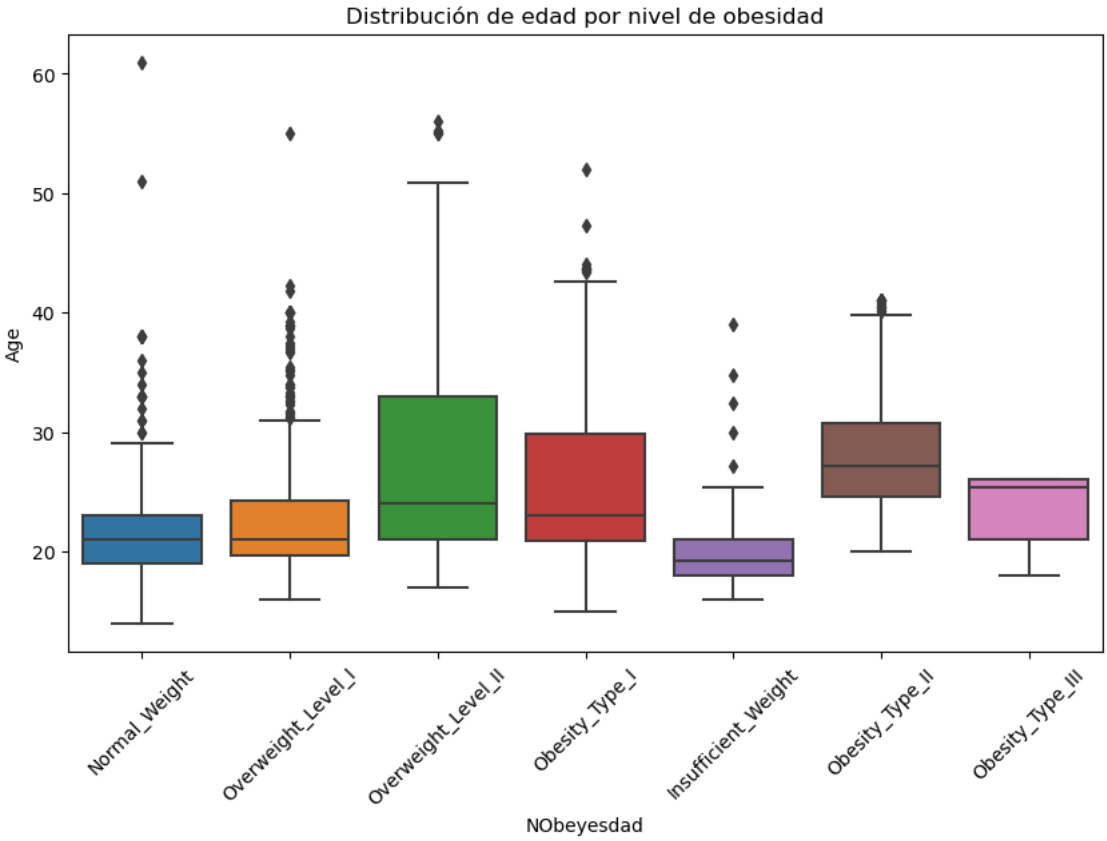
2 Análisis Exploratorio de Datos

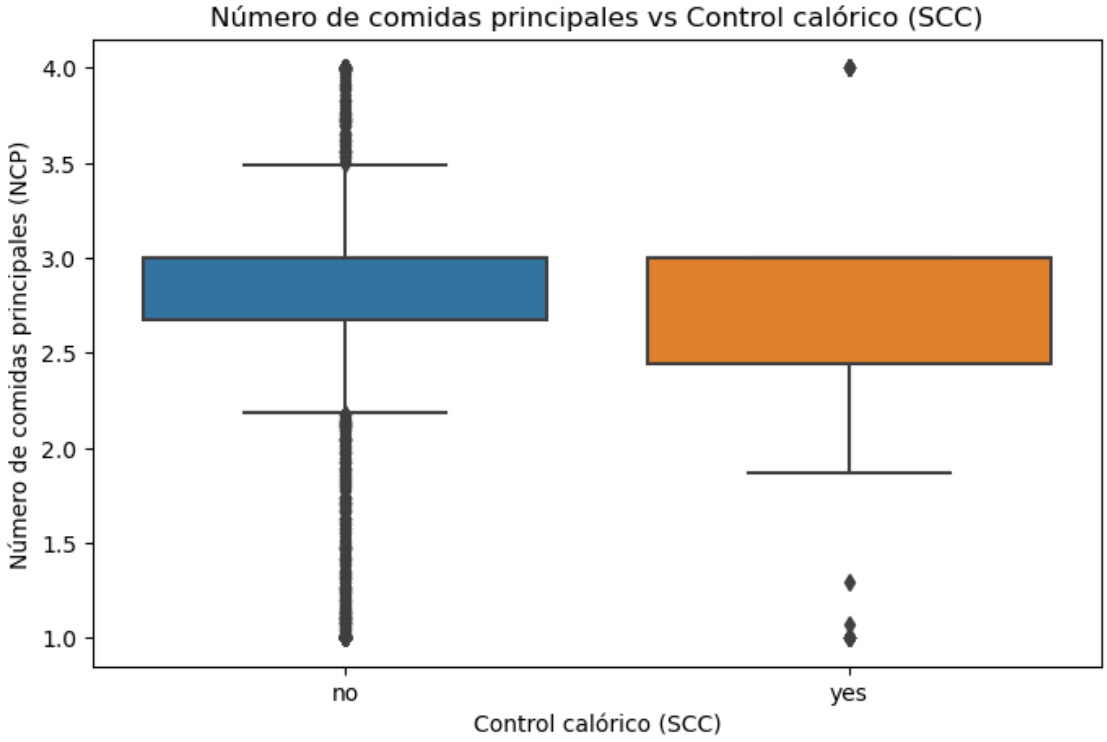
2.1 Descripción general del dataset

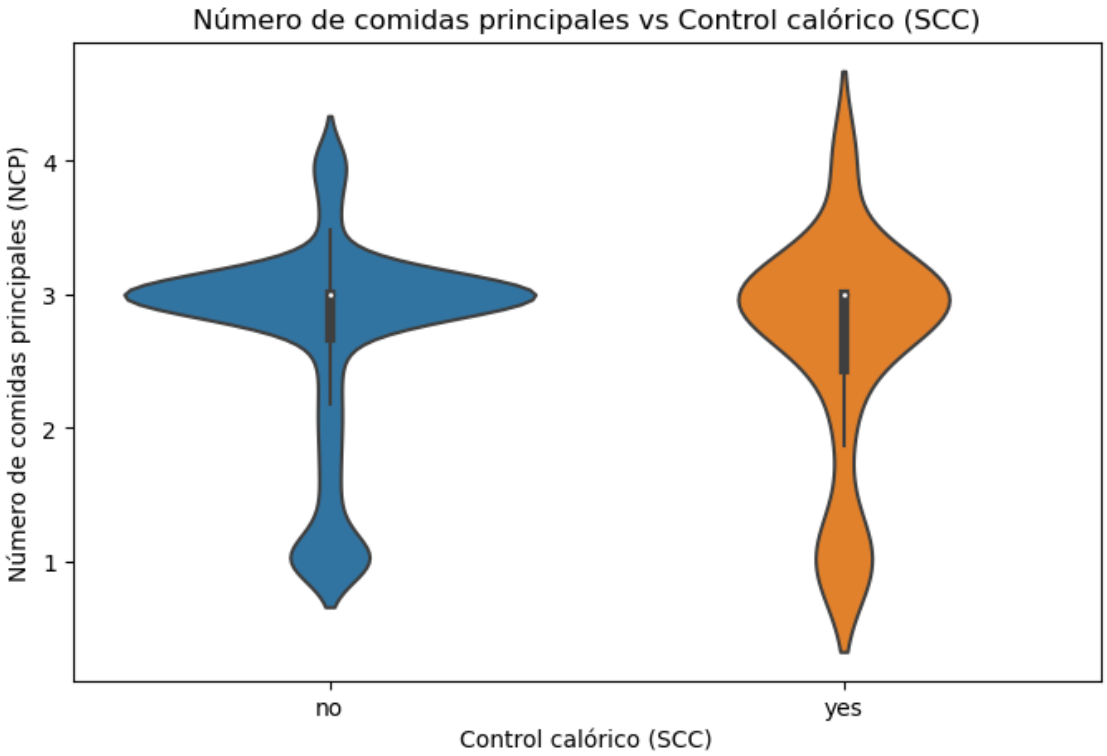
Instancias: 2111 personas
Características : 17 atributos + 1 variable objetivo
Objetivo : Clasificar el nivel de obesidad de una persona

SCC	count
no	2015.0
yes	96.0

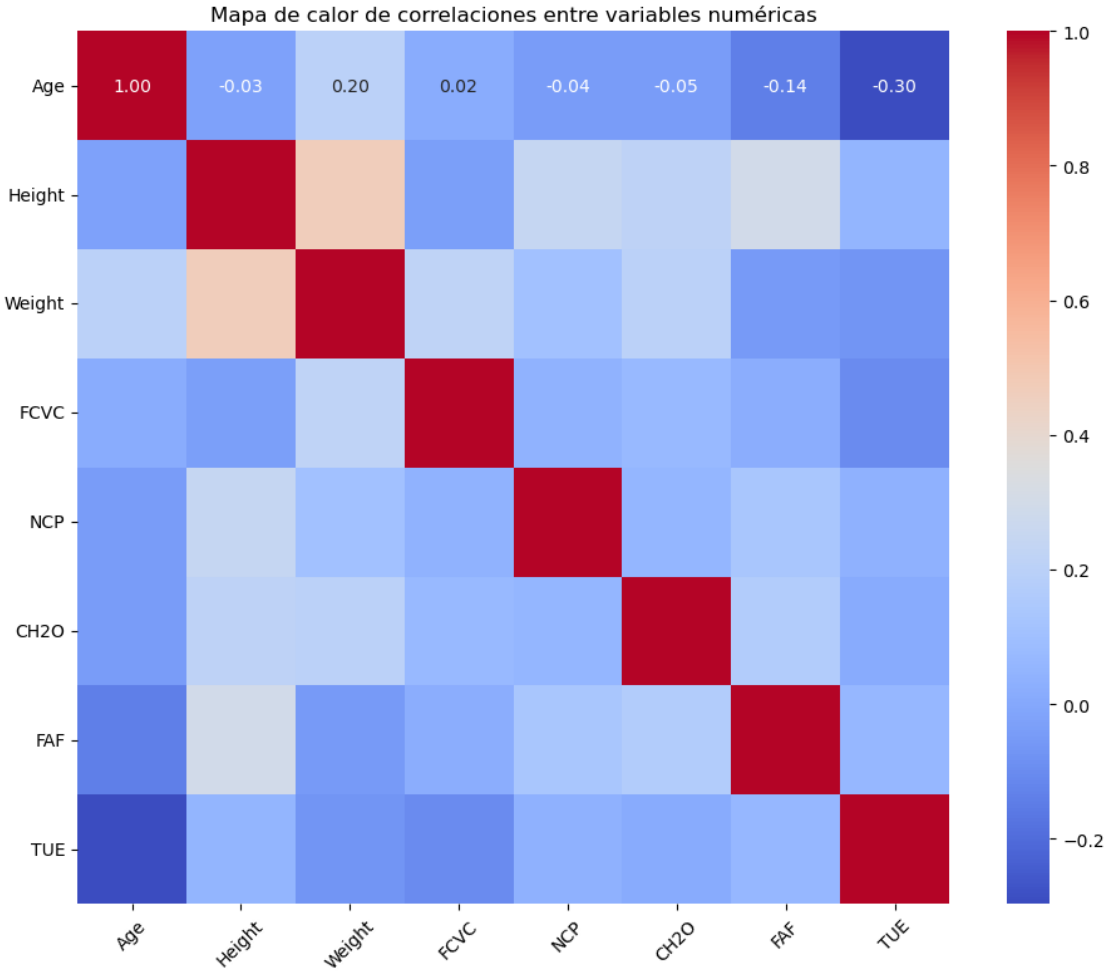
La mayoría de los entrevistados no controlan las calorías que consumen


Distribución de edad por nivel de obesidad
Age
60
50
40
30
20
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III
NObeyesdad


Número de comidas principales vs Control calórico (SCC)
Número de comidas principales (NCP)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
no
yes
Control calórico (SCC)

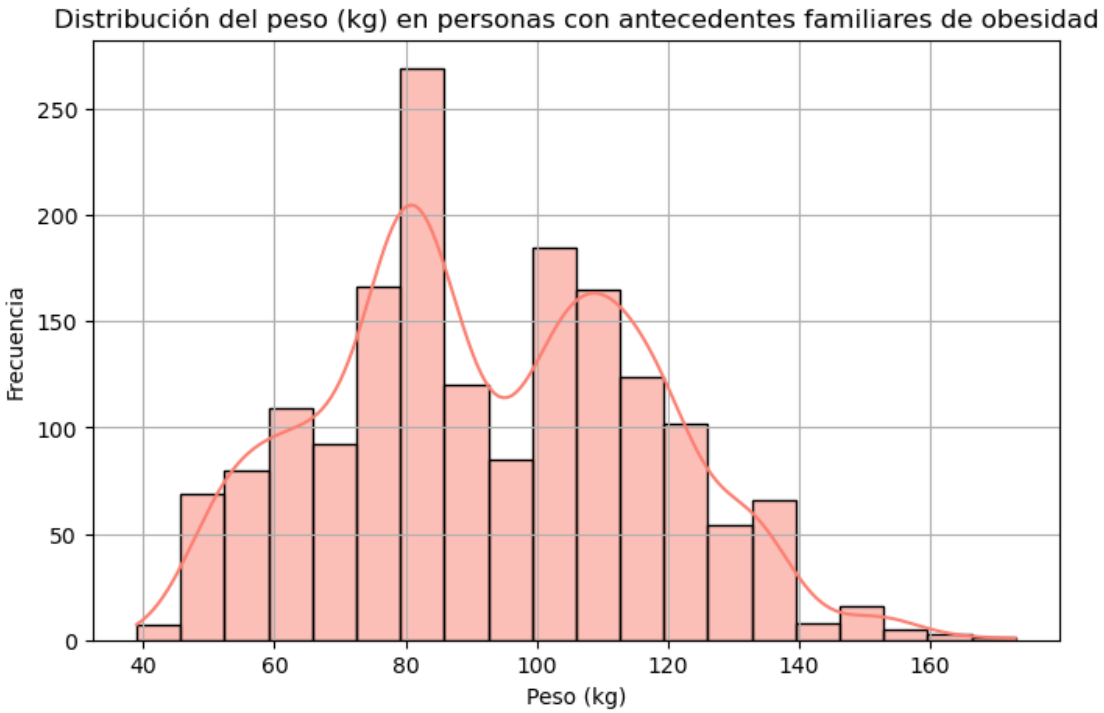


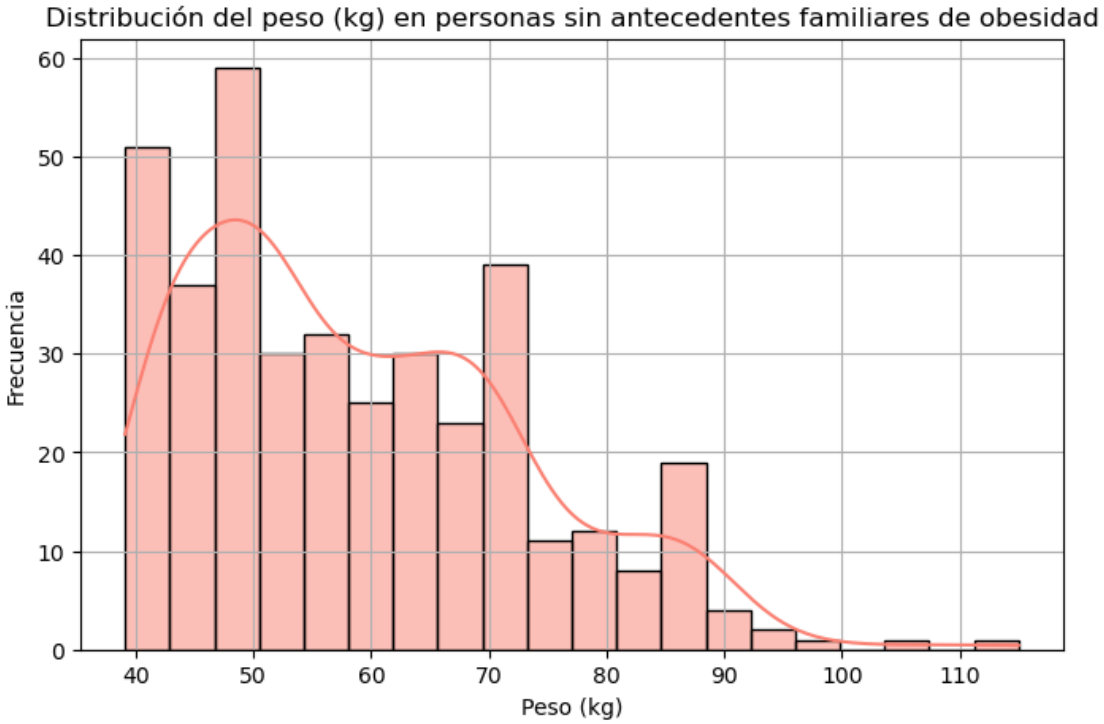
Podemos apreciar que las personas que **si controlan** su consumo de calorías, tienen un mayor rango en la decisión del **Número de comidas principales**. En cambio, los que **no controlan** su consumo de calorías, tienden **en mayoría** , comer **solamente** las 3 comidas principales del día (valor impuesto por la sociedad).



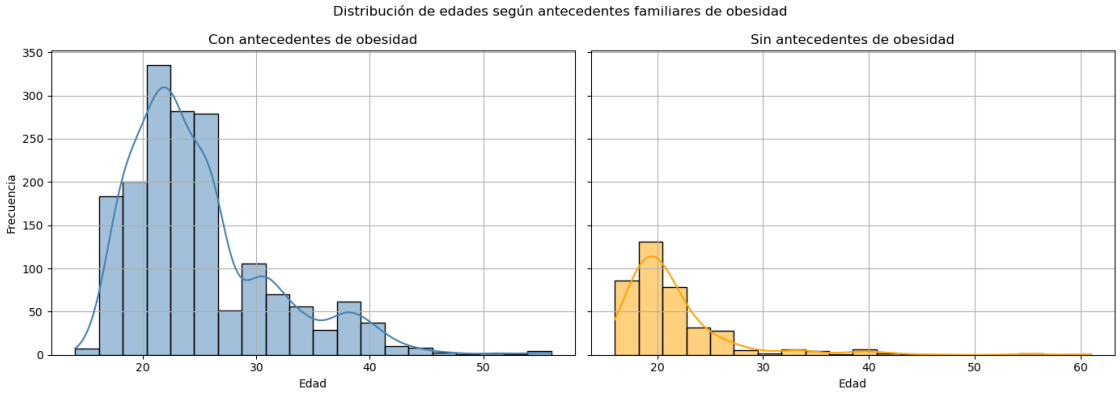

```
  Variable 1 Variable 2  Correlación
7     Height     Weight     0.463136
```

Hay **correlación alta** entre el Peso y la talla de la persona entrevistada.

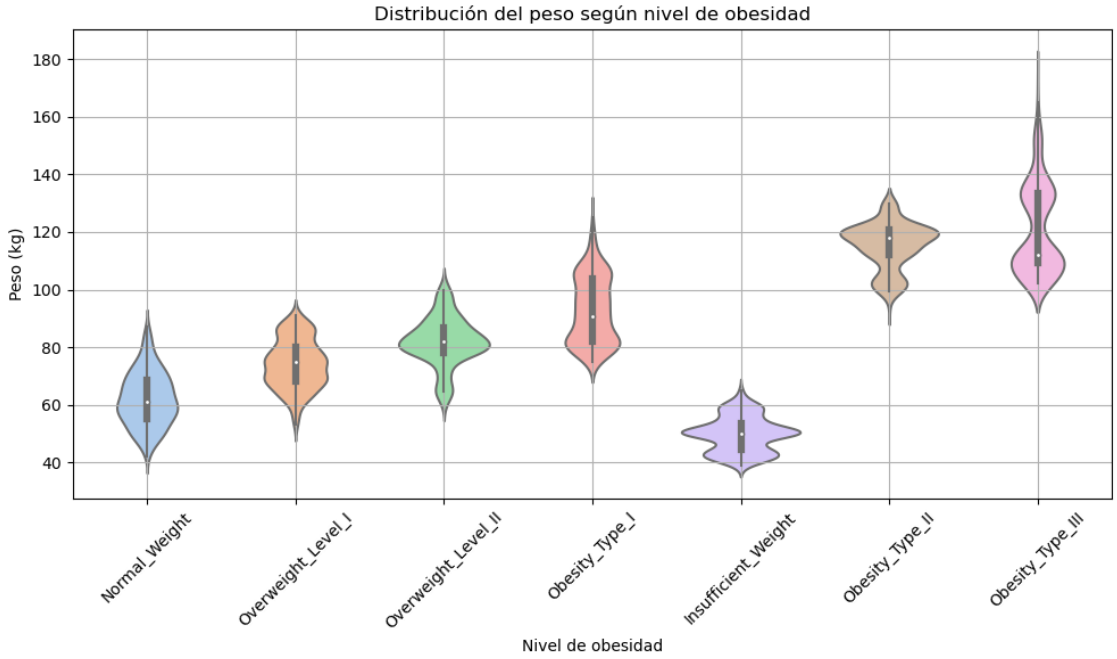

Distribución del peso (kg) en personas con antecedentes familiares de obesidad
Frecuencia
Peso (kg)
0
50
100
150
200
250
40
60
80
100
120
140
160


Distribución del peso (kg) en personas sin antecedentes familiares de obesidad
Frecuencia
Peso (kg)
0
10
20
30
40
50
60
40
50
60
70
80
90
100
110

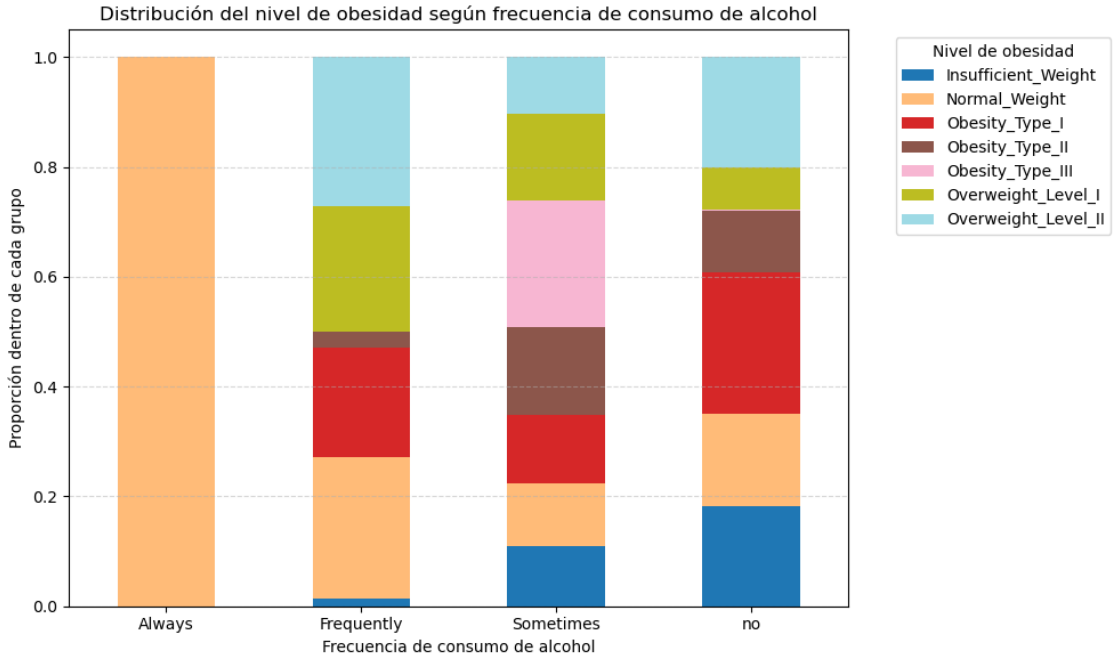
La distribución del peso de las personas que **NO tienen antecedentes familiares con obesidad** , tienen un peso con **asimetría positiva**, es decir que tienden a tener menor peso a comparación de los que **SI tienen antecedentes familiares con obesidad**. Sin embargo es raro que el peso de 40kg sea muy frecuente en el grupo sin antecedentes. Veamos:



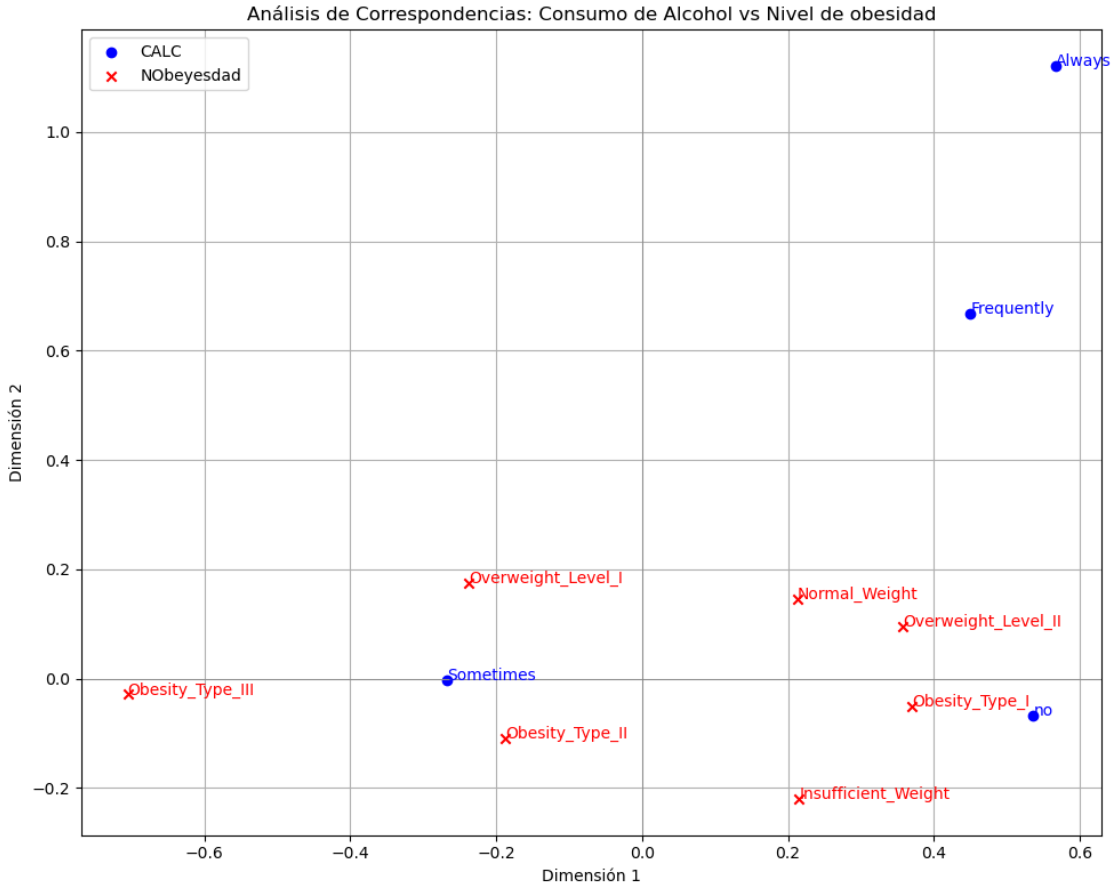
Podemos observar que hay más caso de personas **con antecedentes familiares de obesidad** en este dataset. Tambien hay mayor variación de edad en este grupo respecto al otro.



El **peso** es una **variable importante** para entrenar nuestro modelo de clasificación. Podemos observar que hay diferencias entre los pesos de los niveles



Estadístico Chi^2: 338.578
Grados de libertad: 18
Valor p: 0.0
Hay una relación significativa entre consumo de alcohol y nivel de obesidad (se rechaza H0)



Interesante que los niveles de obesidad más altos no esten necesariamente asociados con el consumo de alcohol en alta frecuencia.

```
NObeyesdad
Insufficient_Weight    0.128849
Normal_Weight          0.135955
Obesity_Type_I         0.166272
Obesity_Type_II        0.140692
Obesity_Type_III       0.153482
Overweight_Level_I     0.137376
Overweight_Level_II    0.137376
Name: NObeyesdad, dtype: float64
```

El target esta completamente **balanceado**.

3 Modelamiento

Probaremos los Algoritmos **KNN** , **Random Forest**, **Regresión Logística** y lo tunearemos cada modelo mediante **validación cruzada** de 10 capas, usaremos **F1-Score** para elegir el mejor modelo ya que pondera **Precisión** y **Recall** para no subestimar categorías.

Tambien repartiremos los datos en proporción de **70:30** para **Entrenamiento** y **Validación** respectivamente.

KNN:

- Se comparará el entrenamiento con **3,5,7** vecinos.
- Se comparará el entrenamiento con pesos diferentes.

Random Forest:

- Se comparará el entrenamiento bosques de **100 y 200** arboles de decisión.
- Se variará la máxima profundidad 10 y 20.

Regresión Logística:

- Se varía el inverso C de regularización. - Se prueban los dos tipos de regularización **Ridge & Lasso**.

```
 RandomForest
Mejores hiperparámetros: {'model__max_depth': 20, 'model__n_estimators': 100}
Mejor F1 promedio en validación cruzada: 0.9504
Evaluación en test set:
                     precision    recall  f1-score   support

Insufficient_Weight       1.00      0.93      0.96        82
      Normal_Weight       0.71      0.90      0.79        86
     Obesity_Type_I       0.97      0.94      0.96       106
    Obesity_Type_II       1.00      0.99      0.99        89
   Obesity_Type_III       1.00      0.99      0.99        97
 Overweight_Level_I       0.86      0.82      0.84        87
Overweight_Level_II       0.96      0.87      0.92        87

           accuracy                           0.92       634
          macro avg       0.93      0.92      0.92       634
       weighted avg       0.93      0.92      0.92       634
```

```
 KNN
Mejores hiperparámetros: {'model__n_neighbors': 3, 'model__weights': 'distance'}
Mejor F1 promedio en validación cruzada: 0.8452
Evaluación en test set:
                     precision    recall  f1-score   support

Insufficient_Weight       0.84      0.96      0.90        82
      Normal_Weight       0.73      0.41      0.52        86
     Obesity_Type_I       0.89      0.95      0.92       106
    Obesity_Type_II       0.95      0.97      0.96        89
   Obesity_Type_III       0.99      1.00      0.99        97
 Overweight_Level_I       0.69      0.76      0.72        87
Overweight_Level_II       0.74      0.79      0.77        87

           accuracy                           0.84       634
          macro avg       0.83      0.83      0.83       634
```

```
       weighted avg       0.84      0.84      0.83       634


 LogisticRegression
Mejores hiperparámetros: {'model__C': 10.0, 'model__penalty': 'l1'}
Mejor F1 promedio en validación cruzada: 0.7884
Evaluación en test set:
                     precision    recall  f1-score   support

Insufficient_Weight       0.98      1.00      0.99        82
      Normal_Weight       0.71      0.66      0.69        86
     Obesity_Type_I       0.65      0.74      0.69       106
    Obesity_Type_II       0.91      0.96      0.93        89
   Obesity_Type_III       1.00      0.99      0.99        97
 Overweight_Level_I       0.60      0.64      0.62        87
Overweight_Level_II       0.59      0.46      0.52        87

           accuracy                           0.78       634
          macro avg       0.78      0.78      0.78       634
       weighted avg       0.78      0.78      0.78       634


 Mejor modelo: RandomForest con F1 CV = 0.9504
```

El mejor modelo que conseguimos fue el **Random Forest** de maxima profundidad 20 y 100 arboles con un F1-score promedio de 0.9504, por lo tanto es el mejor clasificando.

```
Modelo guardado como modelo_obesidad.pkl

[NbConvertApp] Converting notebook Proyecto.ipynb to pdf
[NbConvertApp] Support files will be in Informe_files\
[NbConvertApp] Making directory .\Informe_files
[NbConvertApp] Writing 28029 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | b had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 391093 bytes to Informe.pdf
```