**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA: CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----

A logo of hands holding a book and a candle

Description automatically generated

**MÔN: KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI:**

**NGHIÊN CỨU MỨC ĐỘ BÉO PHÌ DỰA TRÊN**

**THÓI QUEN HÀNG NGÀY VÀ TÌNH TRẠNG THỂ CHẤT**

**Mã môn học: NDS331085\_23\_2\_01**

**Thực hiện: Nhóm 4**

**Lê Dưỡng 21110412**

**Trần Văn Đại 21110875**

**Mai Đình Hồng Sơn 21110465**

**GVHD: Lê Minh Tân**

**Tp. Hồ Chí Minh, tháng 4 năm 2024**

Mục lục

**[1. Giới thiệu 3](#_Toc29660)**

[1.1 Giới thiệu bệnh béo phì 3](#_Toc15184)

[1.2 Giới thiệu về phạm vi dự án 3](#_Toc16375)

**[2. Mô tả dữ liệu 4](#_Toc27885)**

**[3. Thống kê và so sánh chung 6](#_Toc4477)**

**[4 - So sánh chuyên sâu 11](#_Toc19681)**

[4.1 Các biến liên quan đến nhân khẩu: 11](#_Toc23851)

[4.2 - Các thuộc tính liên quan đến ăn uống: 23](#_Toc27615)

[4.3 - Các thuộc tính liên quan đến tình trạng thể chất: 33](#_Toc11328)

**[5. Kết luận 41](#_Toc25622)**

**[6. Tài liệu tham khảo 43](#_Toc22104)**

# **1. Giới thiệu**

## 1.1 Giới thiệu bệnh béo phì

Béo phì là một căn bệnh ảnh hưởng đến sức khỏe sức khỏe của con người, không phân biệt giới tính, độ tuổi. Béo phì còn được coi là yếu tố nguy cơ phát triển các bệnh tim mạch, đái tháo đường, rối loạn lipid máu, cùng nhiều bệnh khác, thông qua các rối loạn chuyển hóa như kháng insulin. Nguy hiểm hơn khi trong những thập kỷ gần đây nó có xu hướng gia tăng, minh chứng là theo những số liệu đã được công bố: “Trên toàn thế giới từ năm 1980 đến năm 2008, chỉ số khối cơ thể trung bình (BMI) tăng thêm 0.4kg/m2 ở nam giới và 0.5kg/m2 ở phụ nữ mỗi thập kỷ. Ở Mỹ Latinh, mức tăng mỗi thập kỷ là 0.6kg/m2 ở nam giới và 1.4kg/m2 ở phụ nữ. Theo dự báo dựa trên thông tin từ Tổ chức Y tế Thế giới (WHO), và nếu xu hướng này tiếp tục, ước tính đến năm 2030, hơn 40% dân số thế giới sẽ bị thừa cân và hơn một phần năm sẽ bị béo phì ”. Đó là những con số đáng báo động.

Thông thường, béo phì được xác định bằng chỉ số khối cơ thể (BMI) và béo bụng (OA) thông qua chu vi vòng eo (WC). Mỗi quốc gia, thể trạng, cơ địa sẽ được phân theo từng nhóm người với mức chỉ số BMI khác nhau.

## 1.2 Giới thiệu về phạm vi dự án

Một số nhà nghiên cứu đã nỗ lực rất nhiều để xác định sớm các yếu tố ảnh hưởng đến việc phát sinh béo phì, thậm chí còn tạo ra các công cụ web như tính chỉ số khối cơ thể (BMI), tại đây bạn có thể tính toán mức độ béo phì của một người. Tuy nhiên, những công cụ này chỉ giới hạn trong việc tính toán chỉ số BMI, bỏ qua các yếu tố liên quan khác như liệu cá nhân có tiền sử gia đình béo phì hay không, thời gian dành cho thói quen tập thể dục, hồ sơ biểu hiện gen và các yếu tố khác. Ngoài việc được coi là yếu tố nguy cơ phát triển các bệnh khác thì phí sinh hoạt cao do béo phì có thể gây ra khó khăn tài chính đáng kể cho các cá nhân và gia đình mắc bệnh này.

Từ đó, mục tiêu của nghiên cứu này là đánh giá tác động của một loạt các yếu tố liên quan đến lối sống hàng ngày và hoạt động thể chất lên nguy cơ phát triển béo phì cùng với đó là tạo ra được một mô hình có thể dự đoán được mức độ béo phì của một người. Bằng cách này, chúng tôi hy vọng rằng nghiên cứu sẽ cung cấp thông tin hữu ích cho cộng đồng và giúp mọi người nhận thức rõ hơn về nguy cơ béo phì dựa trên các yếu tố chính trong lối sống của họ.

# **2. Mô tả dữ liệu**

[Obesity or CVD risk](https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/download?datasetVersionNumber=1) - Tác giả: Fabio Mendoza Palechor, Alexis de la Hoz Manotas

Nghiên cứu này sử dụng dữ liệu từ một cuộc khảo sát trên web được thực hiện trên 2111 cá nhân cư trú tại Mexico, Peru và Colombia. Độ tuổi của những người tham gia trải dài từ 14 đến 61 tuổi, đồng đều ở cả hai giới tính. Mặc dù dữ liệu cung cấp một mẫu đa dạng, nhưng định dạng khảo sát web ẩn danh thừa nhận hạn chế tiềm ẩn về tính tổng quát do sự thiên lệch chọn mẫu.

Tập dữ liệu bao gồm các biến khác nhau có liên quan đến việc nghiên cứu bệnh béo phì. Chúng bao gồm các chi tiết về thói quen ăn uống (tần suất tiêu thụ thực phẩm nhiều calo, lượng rau ăn vào, tần suất ăn, thói quen ăn vặt, lượng nước tiêu thụ và lượng rượu tiêu thụ), thói quen sinh hoạt (theo dõi lượng calo nạp vào, tần suất hoạt động thể chất, thời gian sử dụng công nghệ và phương tiện di chuyển thông thường) và thông tin nhân khẩu (giới tính, tuổi, chiều cao và cân nặng). Ngoài ra, biến "NObeyesdad" (Mức độ béo phì) phân loại mức độ béo phì của người tham gia với nhiều mức độ từ thiếu cân đến béo phì cấp độ 3.

**Mô tả thuộc tính**: Dữ liệu chứa 17 thuộc tính và 2111 bản ghi, chứa dữ liệu số và dữ liệu liên tục (n: danh nghĩa, s: định tính)

*+ Các thuộc tính liên quan tới nhân khẩu là:*

1. Gender (n/s) - giới tính: female (nữ), male (nam)

2. Age (s) - tuổi: giá trị số

3. Height (s) - chiều cao: giá trị số tính bằng mét (m)

4. Weight (s) - cân nặng: giá trị số tính bằng kilôgam (kg)

5. family history with overweight (n/s) - đã từng có thành viên trong gia đình bị thừa cân: yes (có), no (không)

*+ Các thuộc tính liên quan đến ăn uống:*

6. FAVC - Frequent consumption of high caloric food (n/s) - Thường xuyên tiêu thụ thực phẩm nhiều calo: yes (có), no (không)

7. FCVC - Frequency of consumption of vegetables (n/s) - Tần suất tiêu thụ rau: never (không bao giờ), sometimes (đôi khi), always (luôn luôn)

8. NCP - Number of main meals (n/s) - Số bữa ăn chính hàng ngày: Between 1 y 2 (giữa 1 và 2), three (ba), more than three (hơn ba)

9. CAEC - Consumption of food between meals (n/s) - Mức tiêu thụ thực phẩm giữa các bữa ăn: no (không), sometimes (đôi khi), frequently (thường xuyên), always (luôn luôn)

10. CH20 - Consumption of water daily (n/s) - Tiêu thụ nước hàng ngày: Less than a liter (dưới 1 lít), between 1 and 2 L (giữa 1 và 2 lít), more than 2 L (hơn 2 lít)

11. CALC - Consumption of alcohol (n/s) - Tiêu thụ đồ uống có cồn: I do not drink (tôi không uống), sometimes (đôi khi), frequently (thường xuyên), always (luôn luôn)

*+ Các thuộc tính liên quan đến tình trạng thể chất là:*

12. SCC - Calories consumption monitoring (n/s) - theo dõi lượng calo tiêu thụ: yes (có), no(không)

13. FAF - Physical activity frequency (n/s) - tần suất hoạt động thể chất: I do not have (tôi không có), 1 or 2 days (1 hoặc 2 ngày), 2 or 4 days (2 hoặc 4 ngày), 4 or 5 days (4 hoặc 5 ngày)

14. TUE - Time using technology devices (n/s) - thời gian sử dụng các thiết bị công nghệ: 0–2 hours (0 đến 2 giờ), 3–5 hours (3 đến 5 giờ), more than 5 hours (hơn 5 giờ)

15. MTRANS - Transportation used (n/s) - phương tiện di chuyển được sử dụng: automobile(ô tô), motorbike (xe máy), bike (xe đạp), public transportation (phương tiện công cộng),walking (đi bộ)

16. SMOKE(n/s) - hút thuốc: yes (có), no (không)

17.Cuối cùng, tất cả dữ liệu đã được gắn nhãn và biến NObeyesdad được tạo với các giá trị:

Insufficient Weight: Không đủ cân

Normal Weight: Cân nặng bình thường

Overweight Level I: Thừa cân mức độ I

Overweight Level II: Thừa cân mức độ II

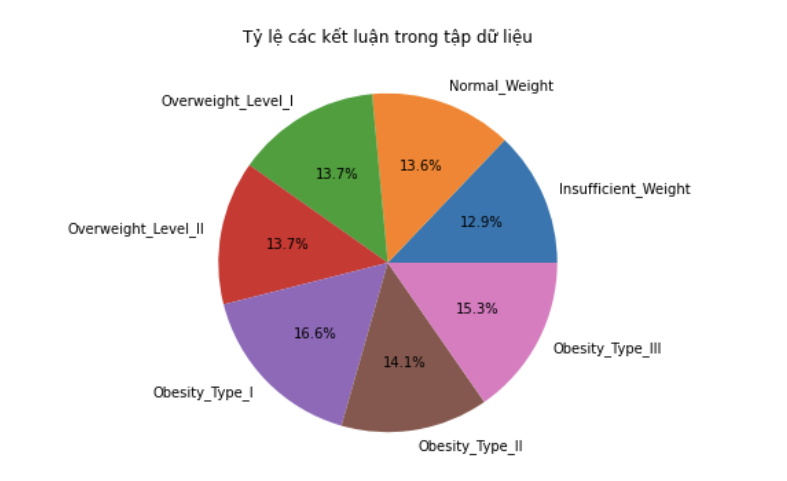
Obesity Type I: Béo phì loại I

Obesity Type II: Béo phì loại II

Obesity Type III: Béo phì loại III

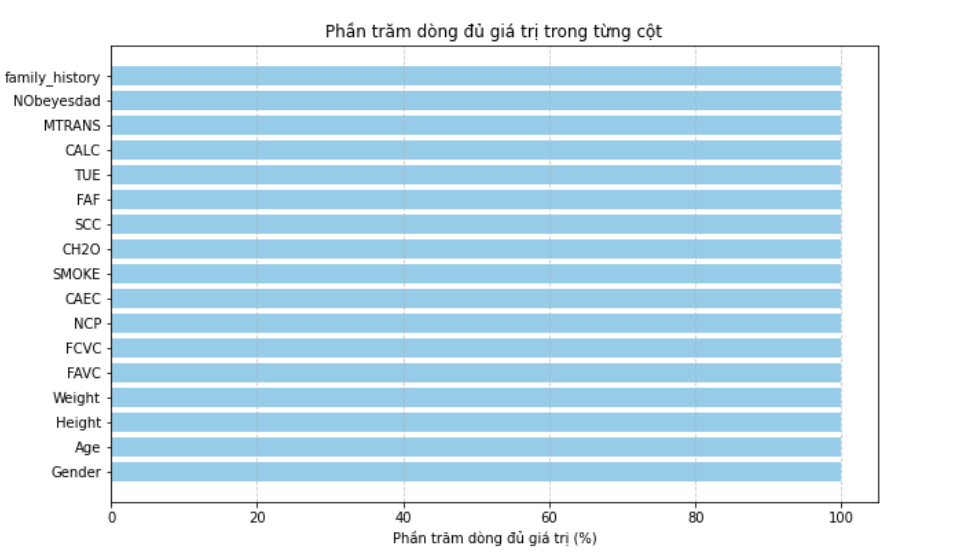
# 3. Thống kê và so sánh chung

* Tỷ lệ kết quả ở trong tập dữ liệu:



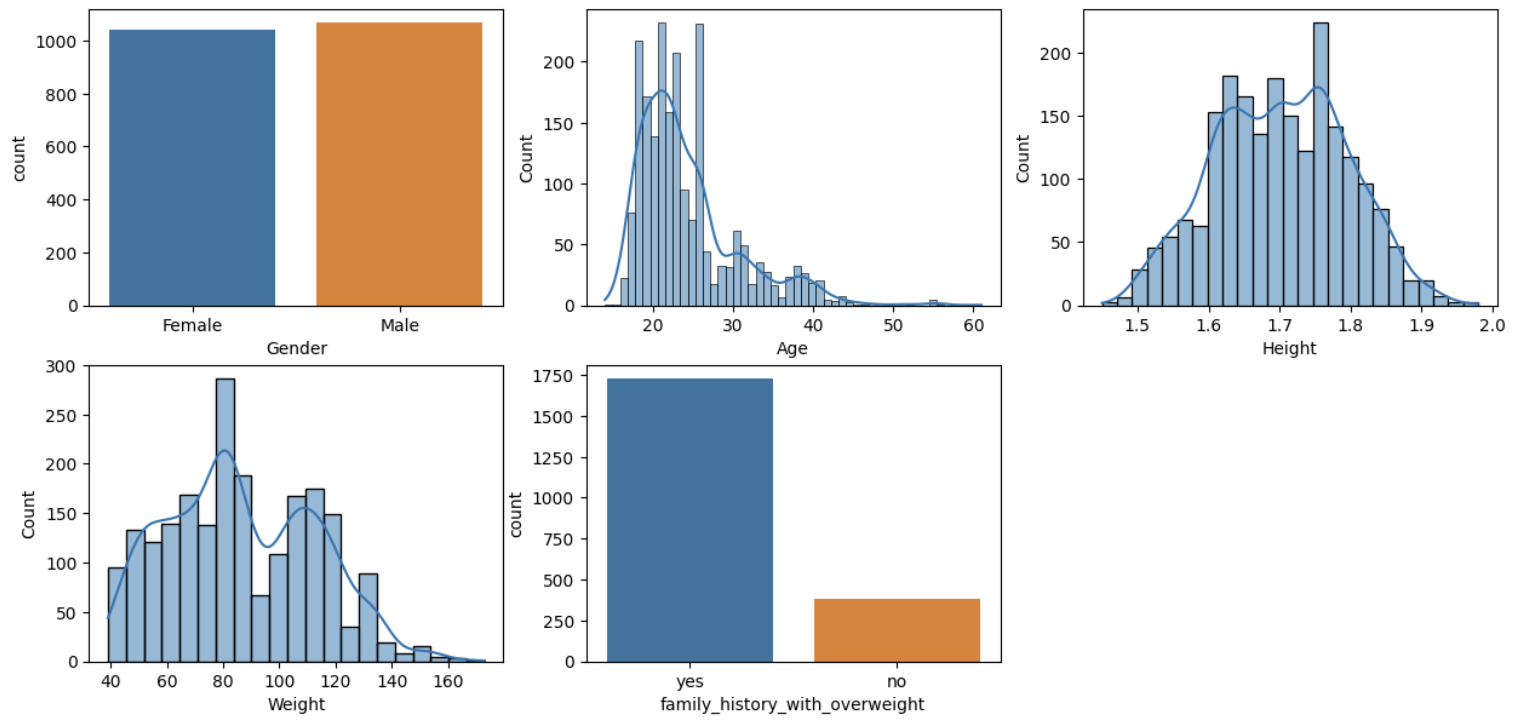
Biểu đồ cho thấy tỷ lệ giữa các kết luận chiếm những phần tương đối cân bằng, dẫn đến việc dùng để huấn luyện cho các mô hình sẽ đảm bảo được tính chính xác hơn chứ không có xu hướng kết quả dự đoán ra thiên về một kết luận nào đó.

Dữ liệu không bị mất mát ở bất cứ bản ghi (dòng) nào, chúng ta có thể thấy rõ hơn thông qua biểu đồ sau:

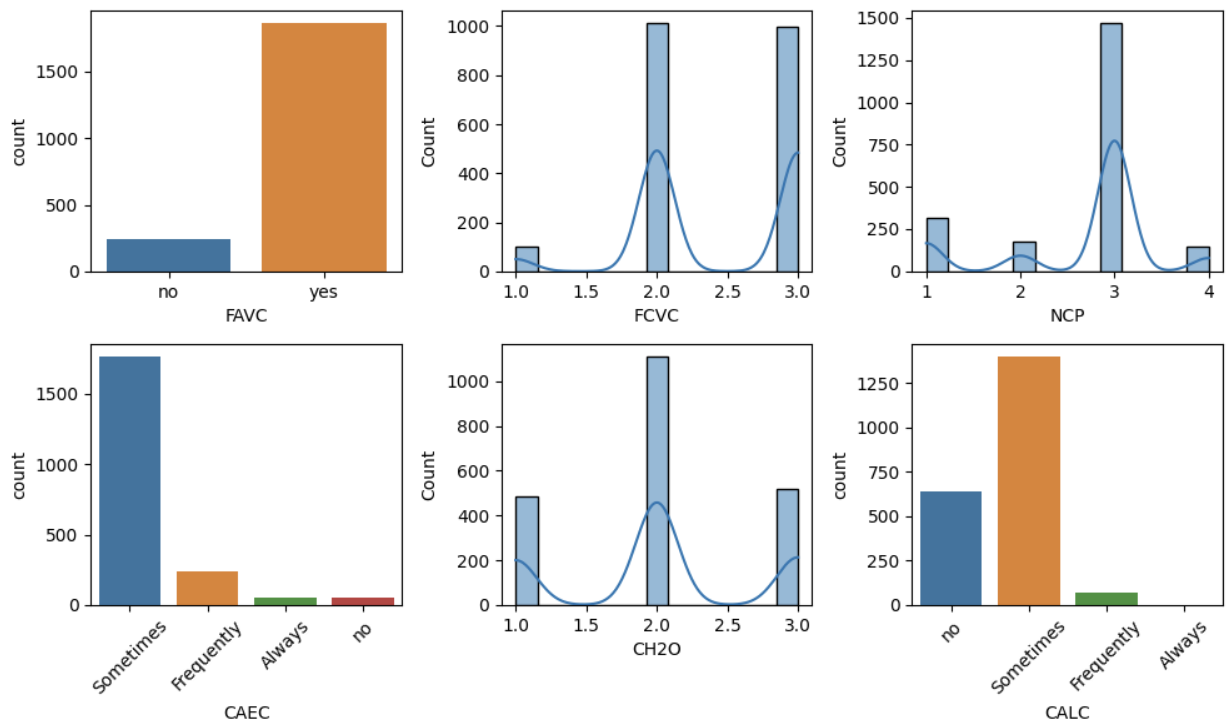


Điều này giúp nghiên cứu giảm thiểu việc phải xử lý những giá trị bị thiếu (nếu bị thiếu có thể loại bỏ đi các dòng bị thiếu nhưng với cách làm thể làm mất tính cân bằng của một số thuộc tính trong dữ liệu dẫn tới việc phân tích chuyên sâu về các thuộc tính đó không mang lại nhiều thông tin hiệu quả hoặc một cách có thể sẽ tốt hơn là sử dụng phương pháp mô phỏng Monte Carlo để tìm và thay thế vào các giá trị bị thiếu đó).

Đối với các thuộc tính liên quan tới nhân khẩu, ta có biểu đồ về sự phân bố các giá trị như sau:

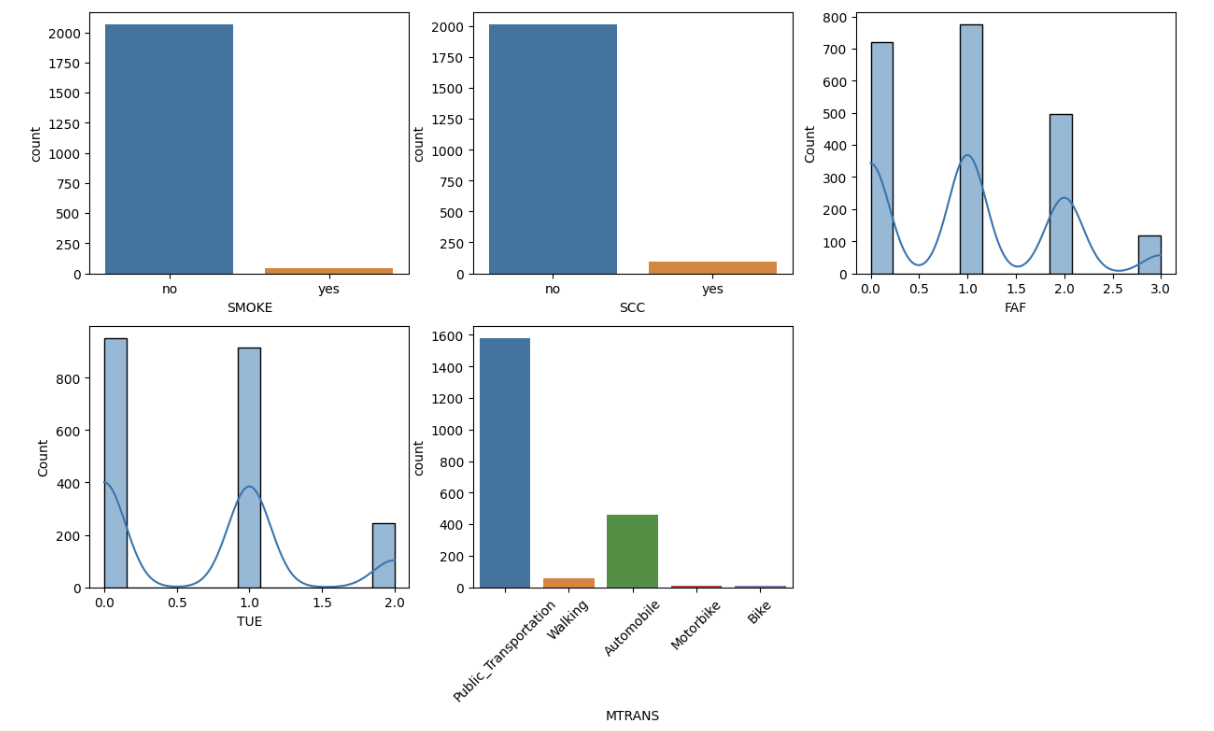


Ta dễ thấy từ biểu đồ của thuộc tính “Gender” (giới tính) được sự cân bằng dường như tuyệt đối giữa tỷ lệ nam và nữ được ghi nhận trong tập dữ liệu. Độ tuổi (thuộc tính “Age”) được khảo sát trải dài từ 14 đến 61 tuổi và có xu hướng tập trung lớn vào nhóm người từ 18 đến 28 tuổi (một khoảng độ tuổi rất quan trọng trong quá trình phát triển của một con người). Chiều cao và cân nặng được phân bố nhiều ở giữa và có xu hướng ít dần về hai phía của biểu đồ (dạng phân phối chuẩn), chiều cao được ghi nhận từ 1m45 đến 1m98, cân nặng được ghi nhận từ 39kg đến 173kg những khoảng thống kê rất bao quát và rất thực tế. Đối với thuộc tính “family\_history\_with\_overweight” (tiền sử bị béo phì trong gia đình) biểu đồ thể hiện sự chênh lệch khi kết quả “yes” được ghi nhận nhiều hơn “no” khá nhiều, có thể là gấp tới 5 lần.

Đối với các thuộc tính ăn uống, ta cùng quan sát biểu đồ về số lần các giá trị được ghi nhận trong dữ liệu như sau:

Các thuộc tính không có sự cân bằng về số lần các giá trị được ghi nhận trong dữ liệu. Số người thường xuyên tiêu thụ thức ăn nhiều calo (“FAVC”) nhiều hơn một cách đáng kể đối với các người ít tiêu thụ thức ăn nhiều calo. Có sự tăng dần về số lượng người ít tiêu thụ rau đến thường xuyên tiêu thụ rau rồi luôn luôn tiêu thụ rau . Số người có 3 bữa ăn chính trong ngày (“NCP”) được ghi nhận nhiều nhất, trong khi các giá trị khác như 1,2 và 4 xuất hiện lít hơn hẵn trong dữ liệu. Đối với mức tiêu thụ thực phẩm giữa các bữa ăn (“CAEC”) thì phần lớn người được khảo sát cho ra kết quả là thỉnh thoảng, cũng có một phần nhỏ là thường xuyên còn lại là không và luôn luôn. Về số lít nước tiêu thụ (“CH2O”) thì đa phần là 2 lít trong một ngày, nhưng các giá trị khác là 1 và 3 cũng chiếm một phần lớn. Dữ liệu cũng thể hiện số người thỉnh thoảng tiêu thụ đồ uống có cồn chiếm phần lớn, tiếp đến là những người không tiêu thụ đồ uống có cồn, một phần khá nhỏ trong dữ liệu là thường xuyên và luôn luôn tiêu thụ đồ uống có cồn.

Cuối cùng là các thuộc tính liên quan đến thói quen sinh hoạt hằng ngày, ta cũng quan sát biểu đồ về số lần các giá trị được ghi nhận trong dữ liệu như sau:



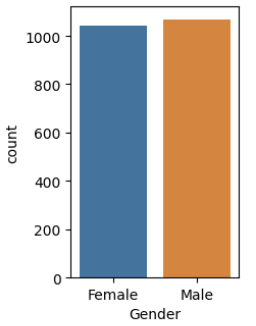
Đối với các thuộc tính “SMOKE” (hút thuốc), “SCC” (theo dõi lượng calo tiêu thụ), “MTRANS” (phương tiện di chuyển) ta thấy được sự không cân bằng về số lần các giá trị được ghi nhận. Số người không hút thuốc chiến phần lớn gần như là toàn bộ trong tập dữ liệu. Số người không theo dõi lượng calo tiêu thụ cũng chiếm rất lớn trong tập dữ liệu. Phương tiện di chuyển được ghi nhận nhiều nhất các phương tiện công cộng, tiếp đến là xe oto còn các phương tiện như xe máy, xe đạp hay đi bộ chiếm một phần khá là nhỏ trong tập dữ liệu. Còn đối với việc hoạt động thể chất và thời gian sử dụng thiết bị công nghệ sự phân bố nhiều ở nhóm người ít vận động và ít sử dụng thiết bị công nghệ, ít dần về phía vận động nhiều và thời gian sử dụng các thiết bị công nghệ nhiều.

# 4 - So sánh chuyên sâu

## 4.1 Các biến liên quan đến nhân khẩu:

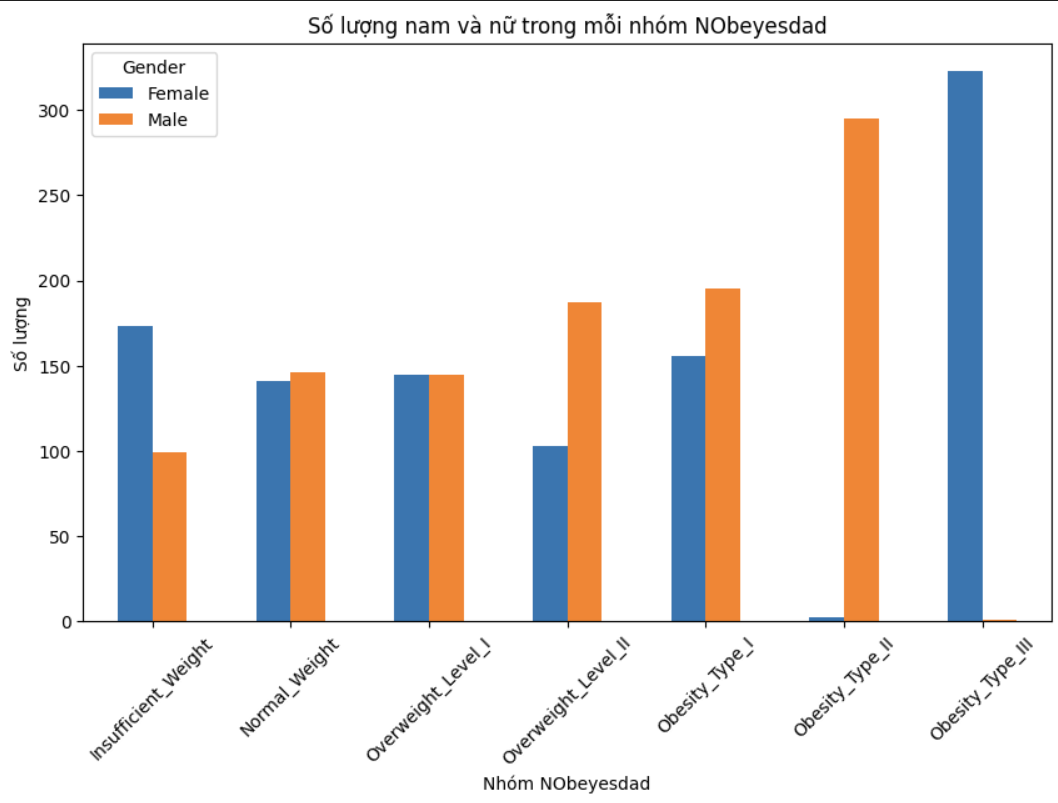
### 1.Gender (n/s) - giới tính: female (nữ), male (nam)

Ta đã thấy sự cân bằng giữa hai giới tính female (nữ) và nam (male) thông qua phần so sánh chung. Ở đây nghiên cứu có trích lại hình ảnh biểu đồ để dễ thấy như sau:



Biểu đồ số lượng các quan sát đối với thuộc tính Gender (Giới tính)

Mặc dù có tỷ lệ cân bằng trong tổng thể như vậy nhưng ta vẫn cần xem trong mỗi nhóm NObeyesdad (mức độ béo phì) thì tỷ lệ có được cân để đóng góp cho mô hình có một kết quả tốt hơn.



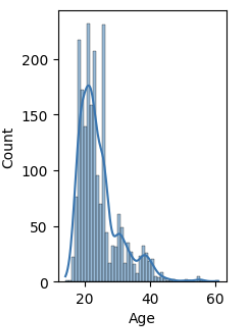
Biểu đồ cho thấy số lương nam và nữ trong mỗi nhóm NObeyesdad, tỷ lệ nam và nữ khá đồng đều trong mỗi nhóm, ngoại trừ hai nhóm “Obesity\_Type\_II” và “Obesity\_Type\_III”. Trong nhóm “Obesity\_Type\_II” số lượng nam chiếm tỉ trọng cao hơn rất nhiều so với số lượng nữ. Còn trong nhóm “Obesity\_Type\_III” thì ngược lại, số lượng nữ chiếm tỷ trọng cao hơn rất nhiều so với số lượng nam. Điều này có thể sẽ rất tệ đối với mô hình nếu có một số thuộc tính khác trong dữ liệu cũng có phân phối như vậy (hiện tượng này còn có thể hiểu như đa cộng tuyến giữa các thuộc tính). Ở đây ta xem tần suất phân bố các giá trị của thuộc tính “Gender” vào từng nhóm “NObeyesdad” như vậy liệu có ý nghĩa đối với mô hình hay nói cách khác là có mối quan hệ giữa giới tính và mức độ béo phì, bằng kiểm định Chi-square (kiểm tra sự khác biệt giữa các nhóm dựa trên tần số quan sát) với giả thuyết H0 là “Không có mối quan hệ giữa giới tính và mức độ béo phì” thì ta có kết quả như sau:

Chi-square statistic: 657.746227342968

p-value: 8.088897293329991e-139

Thống kê Chi-Square cao (657.746): Giá trị Chi-Square này rất cao, cho thấy có sự khác biệt rất lớn giữa thuộc tính “Gender” và nhãn “NObeyesdad”. Giá trị p (8.088e-139): Giá trị p cực kỳ nhỏ, nhỏ hơn 0,05 (mức ý nghĩa thống kê thông thường). Điều này cho thấy khả năng xảy ra ngẫu nhiên kết quả quan sát được là rất thấp (gần như bằng 0) và ta sẽ bác bỏ giả thuyết H0 ban đầu. Từ đầy ta có thể đánh giá được thuộc tính “Gender” có ảnh hướng đến kết quả của mô hình mặc dù có sự phân bố không đều về tỷ lệ nam nữ trong hai nhóm “Obesity\_Type\_II” và “Obesity\_Type\_III”. Điều này cũng thể hiện là béo phì có thể xảy ra ở cả nam và nữ chứ không tập trung vào giới tính nào cả. Và ta có thể quyết định giữ nguyên sự phân bố như vậy mà không cần sử dụng các phương pháp xử lý dữ liệu khác. (Một số phương pháp thể sử dụng để cân bằng lại phân phối của tỷ lệ nam nữ trong hai nhóm này có thể kể đến như mô phỏng Monte Carlo để tạo thêm những giá trị hoặc ta có thể gộp chung hai nhóm béo phì này lại một nhưng đối với phương pháp nào đi nữa thì cần phải cân nhắc kỹ vì điều này có thể dẫn đến việc mất cân bằng của các thuộc tính khác mô hình chung dẫn đến mô hình không có độ chính xác như ban đầu).

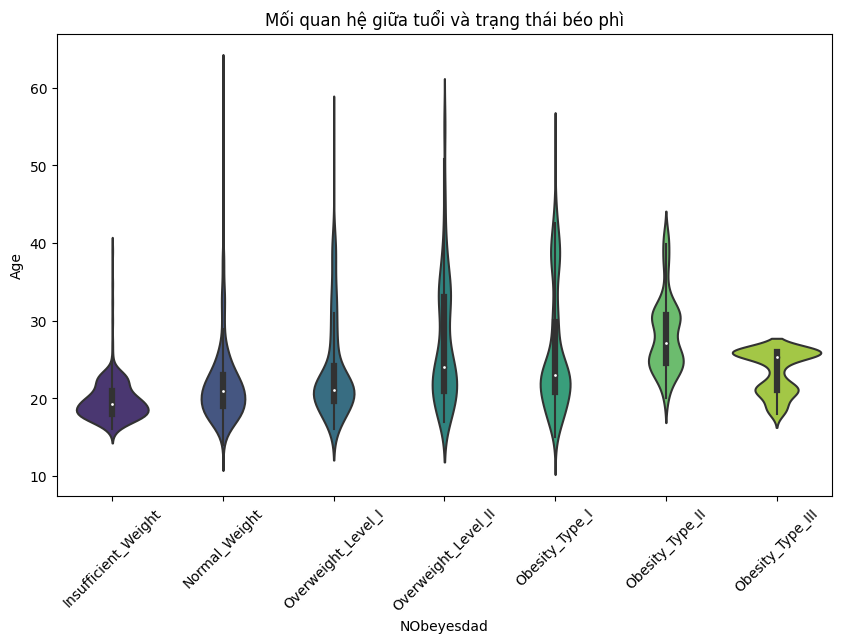
### 2.Age (s) - tuổi: giá trị số



Biểu đồ thể hiện sự phân bố tuổi của những quan sát

Ta thấy được độ tuổi được khảo sát phân bổ từ 14 đến 61 tuổi nhưng phần lớn là nhóm người trẻ. Điều này cũng dễ hiểu vì trong thực tế nhóm người trẻ sẽ có nhu cầu chẩn đoán béo phì cao hơn, họ muốn biết trước để thay đổi một số thói quen hằng ngày trước khi mắc phải một số bệnh như tiểu đường, cao huyết áp, đái tháo đường,… và thậm chí là ung thư.

Để xem xét mức độ tin cậy và sự ảnh hưởng của thuộc tính “Age” đối với nhãn “NObeyesdad”, ta quan sát biều đồ sau:



Biểu đồ violin thể hiện mối quan hệ giữa tuổi và nhóm béo phì

Ta thấy các biểu đồ phân bổ dài trên nhiều độ tuổi khác nhau, không có một nhóm người nào ổn định trong một độ tuổi nhất định.

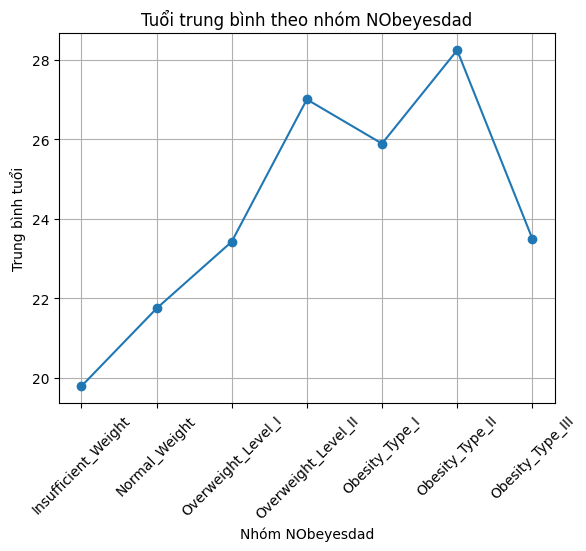
Hình bầu ở đáy nhóm “Insufficient\_Weight” lớn cho thấy những người thiếu cân ở độ tuổi dưới 20 chiến số lượng lớn.

Hình bầu ở nhóm “Obesity\_Type\_III” lớn ở đỉnh cho thấy nhóm người bị béo phì ở mức độ này có độ tuổi lớn.

Hình bầu ở các nhóm khác nhỏ hơn nhưng vẫn thể hiện ý nghĩa tương tự.

Biều đồ dài cho thấy béo phì có thể xảy ra ở bất kỳ độ tuổi nào.

Nhìn vào dấu chấm (màu trắng) ở tâm mỗi violin đây là giá trị trung bình độ tuổi ứng với mỗi nhóm người béo phì. Ta có thể nhìn biểu đồ đường sau để dễ hình dung hơn:

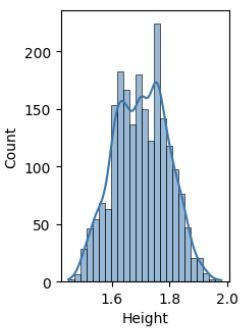


Ta thấy được nhóm người bị béo phì sẽ có độ tuổi trung bình cao hơn so với nhóm người có cân nặng bình thường và thiếu cân. Dễ thấy nó hoàn toàn giống với thực tế, những người thiếu cân thường rơi vào nhóm thanh niên đang phát triển dẫn đến chiều cao tăng nhanh và những người dễ bị thiếu cân. Còn đối với nhóm người béo phì có độ tuổi trung bình cao hơn là có thể do nhiều lý do như:

* Thay đổi về trao đổi chất
* Giảm hoạt động thể chất
* Thay đổi về nội tiết tố
* Các bệnh lý mãn tính

Qua đó ta thấy được mức độ phân tán của thuộc tính “Age” đối với nhãn “NObeyesdad” là đồng đều ở mỗi nhóm và còn thể hiện đúng với thực tế nên ta cần phải sử dụng thuộc tính cho mô hình dự đoán để mô hình tốt hơn cả về mặt toán học và thực tế.

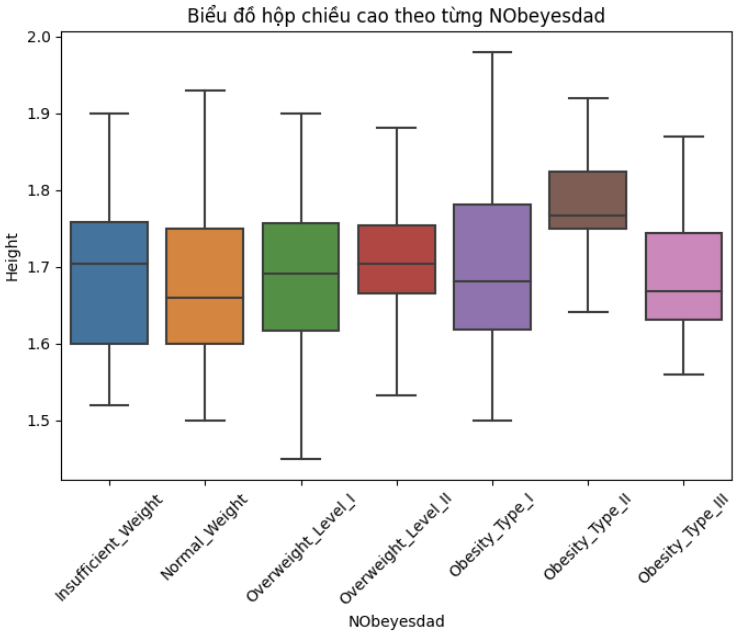
### 3 .Height (s) - chiều cao: giá trị số tính bằng mét (m)



Biểu đồ tần suất các giá trị chiều cao

Dựa vào biểu đồ tần suất các giá trị chiều cao được thu thập trong dữ liệu ta thấy được chiều cao phân bổ nhiều ở giá trị từ 1m48 đến 1m98, quá hợp lý đối với độ tuổi được ghi nhận nhiều ở khoảng 14-61 tuổi. Ở khoảng độ tuổi 14-18 (độ tuổi dậy thì) có nhiều người được khảo sát dường như chưa đến tuổi dậy thì nên chiều cao có thể rất nhỏ đóng góp vào phần bên trái biểu đồ, còn ở khoảng độ tuổi lớn hơn thì chiều cao đã phát triển và ổn định thì thống kế sẽ cho ra nhiều giá trị chiều cao rơi vào khoảng từ 1m60 đến 1m80.

Ngoài ra ta còn có biểu đồ hộp chiều cao của từng nhóm NObeyesdad như sau:



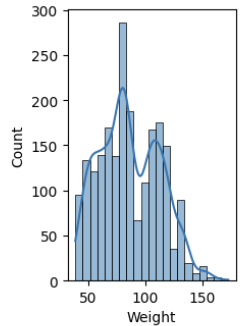
Qua biểu đồ ta thấy được chiều cao trung bình phân bố khá đều ở từng nhóm béo phì, điều này cho thấy việc đánh giá một người thuộc nhóm béo phì nào dựa vào chiều cao dừng như là không thể mà cần phải nhiều yếu tố khác. Và liệu rằng sự phân bố chiều cao như vậy có mang lại ý nghĩa gì đối từng nhóm béo phì thì ta sử dụng kiểm định ANOVA (kiểm định dùng để kiểm tra sự khác biệt trung bình giữa các nhóm), kết qua cho ra như sau:

F-statistic: 38.43231255660025

p-value: 1.6858535844061656e-44

Giá trị p-value rất nhỏ nên ta có thể công nhận một điều là sự khác biệt trung bình giữa các nhóm chiều cao là ý nghĩa đối với kết quả, và ta nên dùng thuộc tính “Height” (chiều cao) vào trong mô hình dự đoán nhằm đạt được độ chính xác cao và đúng với thực tế.

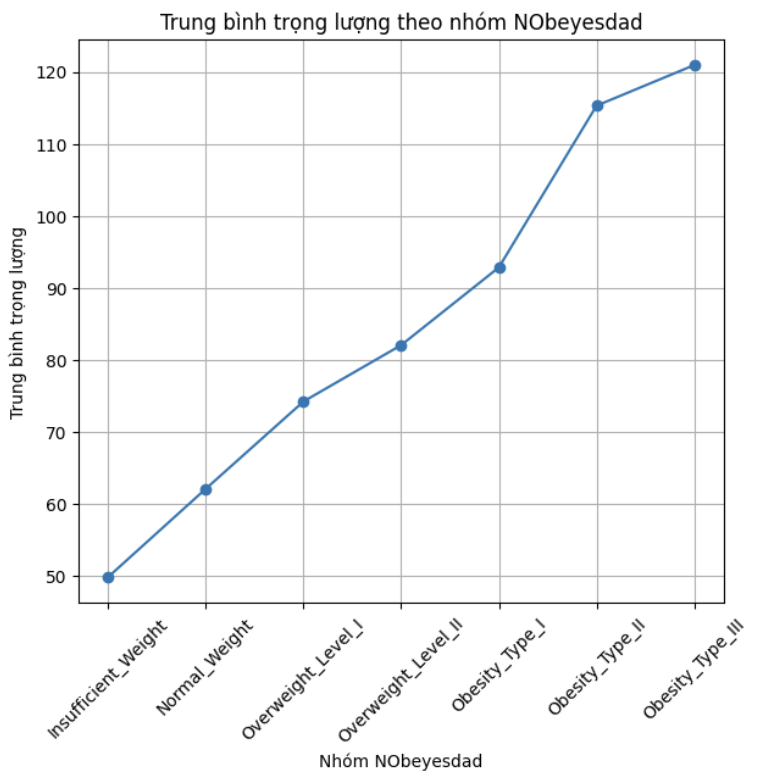
### 4. Weight (s) - cân nặng: giá trị số tính bằng kilôgam (kg)



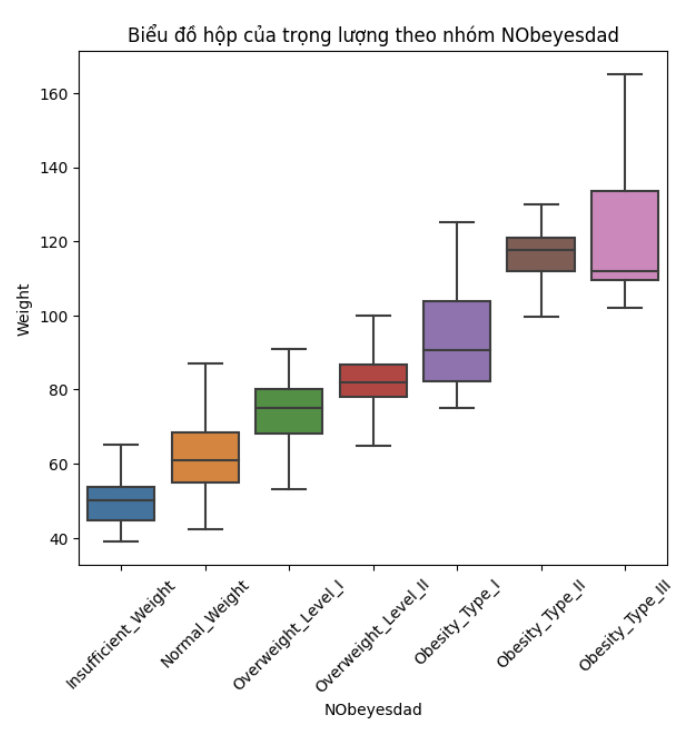
Biểu đồ phân tán trọng lượng

Từ biểu đồ phân tán của thuộc tính Weight (Trọng lượng) trong tập dữ liệu ta thấy dữ liệu được thu thập trải đều trên nhiều khoảng cân nặng khác nhau.

Ta thấy được mức độ ảnh hưởng của “Weight” đối với nhãn “Nobeyesdad” là rất lớn, thông qua biểu đồ trung bình trọng lượng theo nhóm NObeyesdad như sau:

Từ biểu đồ ta thấy được trọng lượng trung bình từng nhóm tăng dần theo từng mức độ béo phì từ “Insufficient\_Weight” cho đến “Obesity\_Type\_III” (Béo phì cấp độ 3 - Tình trạng béo phì nặng nhất). Trên thực tế thì ta cũng dễ hình dung được điều này vì một người bị béo phì dường như chắc chắn có số cân nặng lớn hơn người bình thường, và người bình thường có số cân nặng lớn những người bị thiếu cân.

Nhưng ta sẽ không thể chỉ dựa vào cân nặng để đưa ra được kết luận rằng người đó thuộc nhóm người “Insufficient\_Weight”, “Normal\_Weight” hay bị béo phì được. Để kiểm chứng điều đó ta cùng xem biểu đồ hộp của trọng lượng theo nhóm NObeyesdad:



Ở từng nhóm người, ta thấy được sự phân bố không đồng đều về cân nặng. Ví dụ ở nhóm người “Normal\_Weight” (cân nặng bình thường) vẫn có cơ số người có cân nặng lớn hơn các nhóm bị đánh giá là thừa cân và thậm chí là béo phì cấp độ một, ngược lại vẫn có nhiều người có cân nặng bé hơn nhóm người “Insufficient\_Weight” (thiếu cân). Điều này cho thấy rất khó để chỉ với cân nặng ta có thể cho ra được kết quả dự đoán mà cần phải có thêm những thuộc tính khác trong dữ liệu.

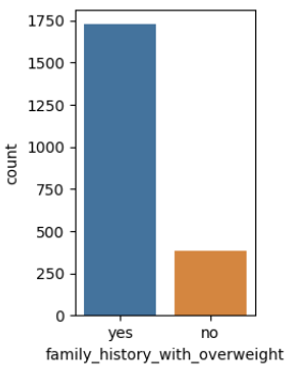
Và còn dựa vào kiểm định ANOVA (kiểm định dùng để kiểm tra sự khác biệt trung bình giữa các nhóm) ta có thể chắc chắn thuộc tính “Weight” có ý nghĩa rất lớn đối với mô hình khi có kết quả như sau:

F-statistic: 2299.8077480715037

p-value: 0.0

Giá trị F-statistic (tính toán tỷ lệ giữa phương sai giữa các nhóm và phương sai trong nhóm. Một giá trị F-statistic lớn đề cập đến việc có ít sự tương đồng giữa các nhóm hơn so với sự đồng nhất bên trong các nhóm) rất lớn cho thấy có sự khác biệt rất lớn về trọng lượng đối với các nhóm béo phì. Giá trị p-value bé giúp ta đưa ra kết luận rằng sự khác biệt trung bình giữa các nhóm trọng lượng là ý nghĩa đối với kết quả. Từ những lý do đã nêu ở trên ta nên đưa thuộc tính “Weight” vào mô hình nhằm nâng cao hiệu suất dự đoán của mô hình.

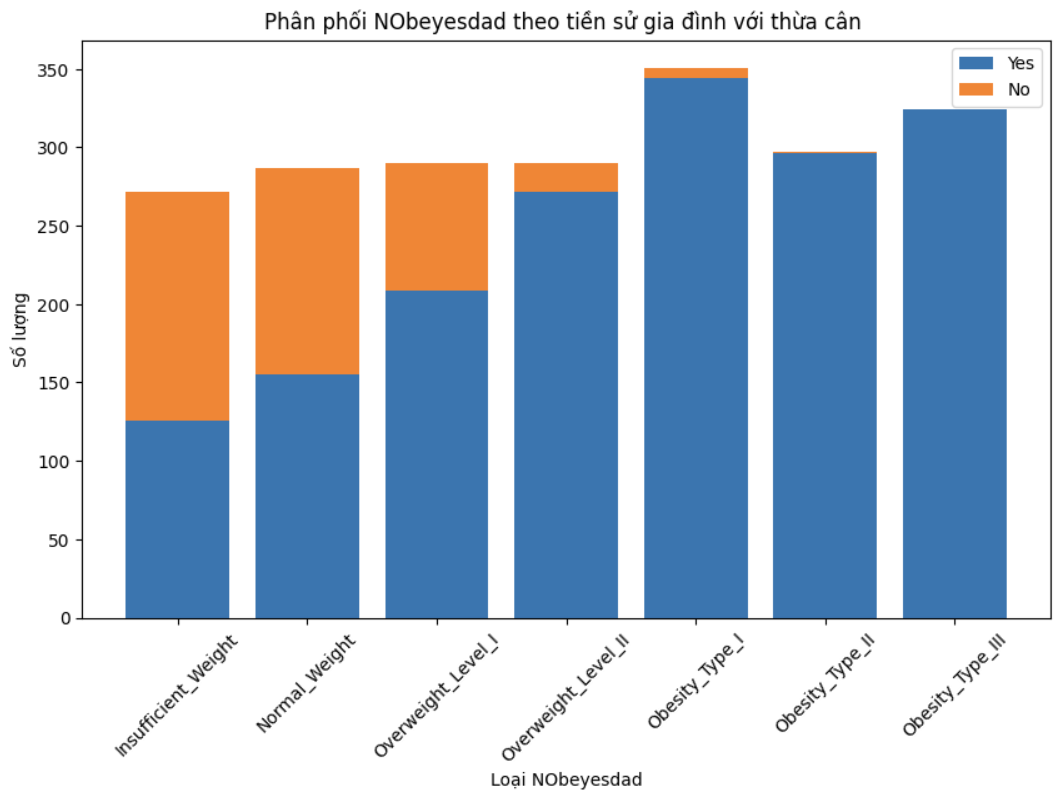
### 5 .family\_history\_with\_overweight (n/s) - đã từng có thành viên trong gia đình bị thừa cân: yes (có), no (không)



Biểu đồ tần suất các giá trị của family\_history\_with\_overwweight

Tỷ lệ người được khảo sát có người thân bị béo phì nhiều hơn gấp 5 lần so với những người được khảo sát là không có người thân bị béo phì. Điều này cũng dễ hiểu vì béo phì có yếu tố di truyền, tức là nếu một người trong gia đình bị béo phì thì khả năng người đó cũng bị béo phì lớn hơn so với người khác. Và trong tập dữ liệu có tới 5 nhóm thuộc nhóm béo phì (“Overweight\_Level\_I”, “Overweight\_Level\_II”, “Obesity\_Type\_I”, “Obesity\_Type\_II”, “Obesity\_Type\_III”) chiếm 74% trong khi nhóm không béo phì có 2 nhóm (“Insufficient\_Weight”, “Normal\_Weight”) chiếm khoảng 26%. Từ những lập luận đó ta dễ hiểu cho việc phân bố không đều giữa hai nhóm trong thuộc tính “family\_history\_with\_overweight” của dữ liệu.

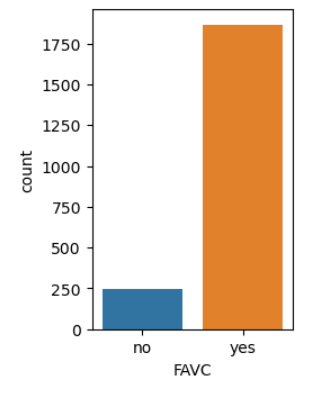
Ta còn thấy được yếu tố di truyền của bệnh béo phì thể hiện rất rõ thông qua biểu đồ phân phối NObeyesdad theo tiền sử gia đình với thừa cân:



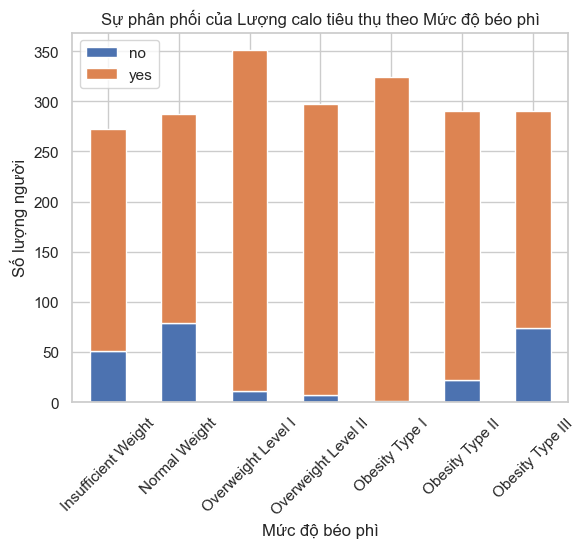
Biểu đồ thể hiện tỷ lệ một người bị béo phì cấp độ càng nặng thì khả năng người thân họ cũng bị béo phì càng lớn. Từ yếu tố này và thực tế chứng mình ta có thể đưa ra quyết định thuộc tính “family\_history\_with\_overweight” sẽ là một yếu tố quan trọng để đưa vào mô hình.

## 4.2 - Các thuộc tính liên quan đến ăn uống:

### 6. FAVC - Thường xuyên tiêu thụ thực phẩm nhiều calo: yes (có), no (không)



Dựa vào biểu đồ ta thấy có sự phân bố không đồng đều giữa nhóm có và không thường xuyên tiêu thụ nhiều thực phẩm nhiều calo, nhóm không thường xuyên tiêu thụ thấp hơn rất nhiều so với nhóm thường xuyên tiêu thụ. Thực tế việc này phụ thuộc vào nhận thức về sức khỏe của từng cá nhân, chỉ một phần nhỏ dân số ý thực và quan tâm tới việc kiểm soát chế độ ăn uống hợp lý hằng ngày. Thêm nữa, yếu tố văn hóa cũng có ảnh hưởng khá lớn. Ví dụ, các món ăn của Mexico, Peru và Colombia thường có các thành phần như bắp, đậu, thịt đỏ và các loại dầu mỡ, có thể tăng hàm lượng calo. Vậy nên sự phân bố không đồng đều của 2 cột yes và no là dễ hiểu.



Ngoài ra ta cũng có thể quan sát rõ ràng rằng việc không tiêu thụ thực phẩm nhiều calo thì có khả năng bị béo phì ít hơn nhưng số người này ở cột béo phì loại 3 cao hơn bất ngờ. Điều này có thể là do những người thuộc nhóm này buộc phải giảm lượng calo hấp thụ vào cơ thể khi phát hiện mình bị béo phì loại 3 bằng cách tránh tiêu thụ thực phẩm nhiều calo.

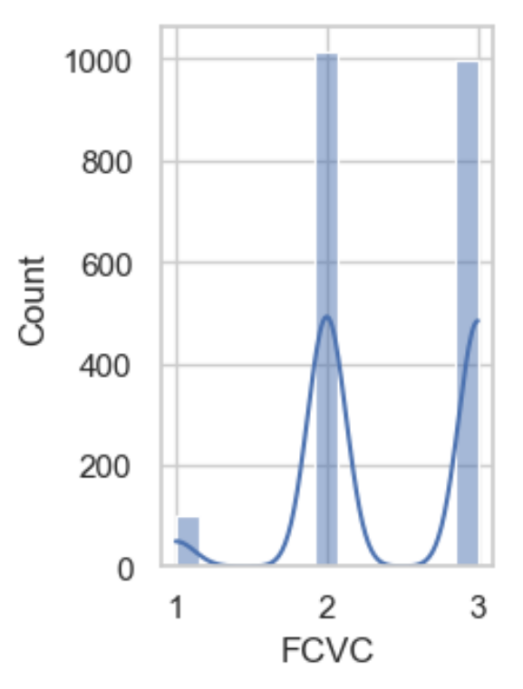
Bằng việc thực hiện kiểm định t-test cho hai nhóm có và không thường xuyên tiêu thụ nhiều thực phẩm nhiều calo ta có được giá trị sau:  
*Giá trị t-statistic: 2.0494289482131802*

*p-value: 0.040543648116553574*

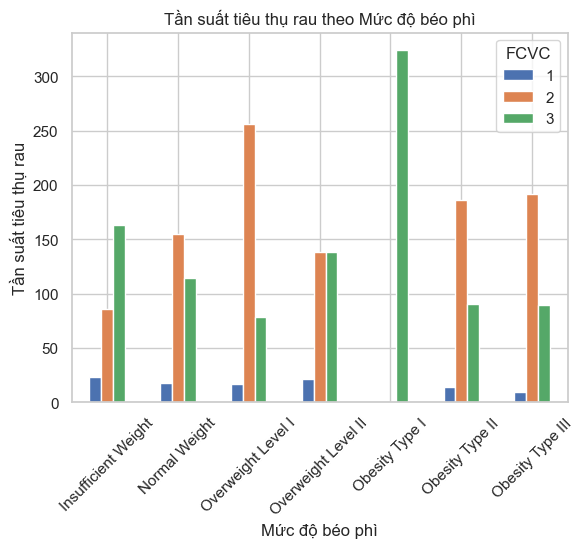
Kết quả của kiểm định t-test là: t-statistic = 2.049 và p-value = 0.041. Với giá trị p-value nhỏ hơn ngưỡng ý nghĩa thống kê alpha (thường là 0.05), chúng ta có đủ bằng chứng để bác bỏ giả thuyết H0, tức là có sự khác biệt đáng kể về mức độ béo phì giữa nhóm tiêu thụ thực phẩm nhiều calo thường xuyên và nhóm không thường xuyên.

Từ kết quả này, chúng ta có thể suy ra rằng thói quen tiêu thụ thực phẩm nhiều calo thường xuyên có thể gắn liền với mức độ béo phì cao hơn so với nhóm không thường xuyên. Điều này có thể phản ánh thực tế rằng việc tiêu thụ nhiều calo thường xuyên có thể dẫn đến tích tụ năng lượng dư thừa trong cơ thể và góp phần vào mức độ béo phì. Vậy thuộc tính FACV có thể có ảnh hưởng nhất định đến tập dữ liệu.

### 7. FCVC - Frequency of consumption of vegetables (n/s) - Tần suất tiêu thụ rau: 1 (không bao giờ), 2 (đôi khi), 3 (luôn luôn)

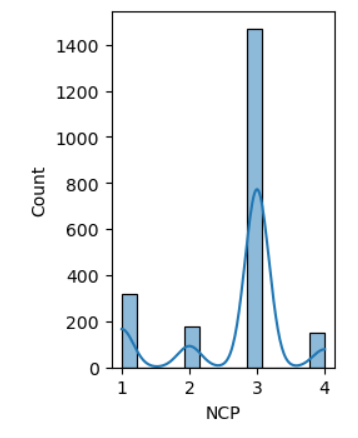


Biểu đồ cho thấy rằng rất ít người không bao giờ tiêu thụ rau, trong khi đa số thống kê có giá trị là đôi khi và luôn luôn. Điều này có thể lý giải bằng việc rau là một phần không thể thiếu trong bữa ăn của con người.



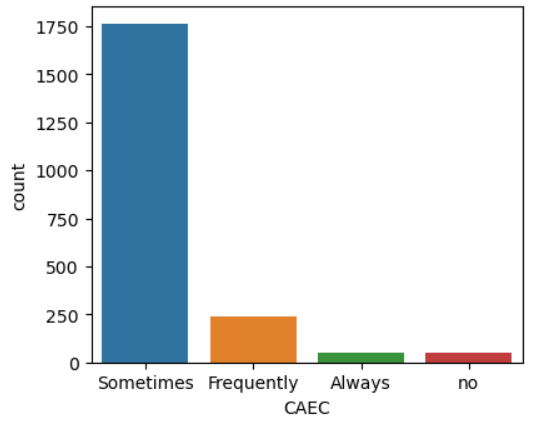
Biểu đồ trên thể hiện sự phân bố giữa các nhóm tiêu thụ rau cho thấy rằng số lượng người không bao giờ ăn rau phân bố đều ở các nhóm béo phì. Trong khi đấy ở nhóm béo phì loại 1 không có các giá trị của việc đôi khi và luôn luôn tiêu thụ rau. Các yếu tố trên có thể có ảnh hưởng không tốt lên tập dữ liệu. Khi liên hệ với thực tế tuy tần suất tiêu thụ rau có thể là một yếu tố quan trọng, nhưng không phải là yếu tố quyết định về cân nặng và béo phì. Các loại thực phẩm khác như đạm, tinh bột và chất béo, cũng đóng vai trò quan trọng trong việc duy trì cân nặng, sức khỏe và ảnh hưởng tới tình trạng béo phì.

### 8. NCP - Number of main meals (n/s) - Số bữa ăn chính hàng ngày: 1,2,3,4



Biểu đồ thể hiện rằng 3 bữa ăn chính hằng ngày chiếm nhiều nhất, trong khi các giá trị khác chỉ chiếm một phần rất nhỏ. Điều này có thể hiểu được vì đa số mọi nơi trên thế giới đều ăn 3 bữa chính một phần vì văn hóa từ thời xưa, phần vì cơ chế tiêu hóa của con người.Trong thực tế, số bữa ăn chính có thể ảnh hưởng tới cân nặng và tình trạng béo phì, tuy nhiên cần phải xem xét đến các yếu tố dinh dưỡng khác của bữa ăn như lượng calo cũng như đạm và chất béo được hấp thụ. Do đó, số lượng bữa chính trong ngày mặc dù có thể ít nhiều ảnh hưởng tới tình trạng béo phì nhưng nó không thể là nguyên nhân quan trọng trong tập dữ liệu này.

### 9. CAEC - Consumption of food between meals (n/s) - Mức tiêu thụ thực phẩm giữa các bữa ăn: no (không), Sometimes (đôi khi), Frequently (thường xuyên), Always (luôn luôn)

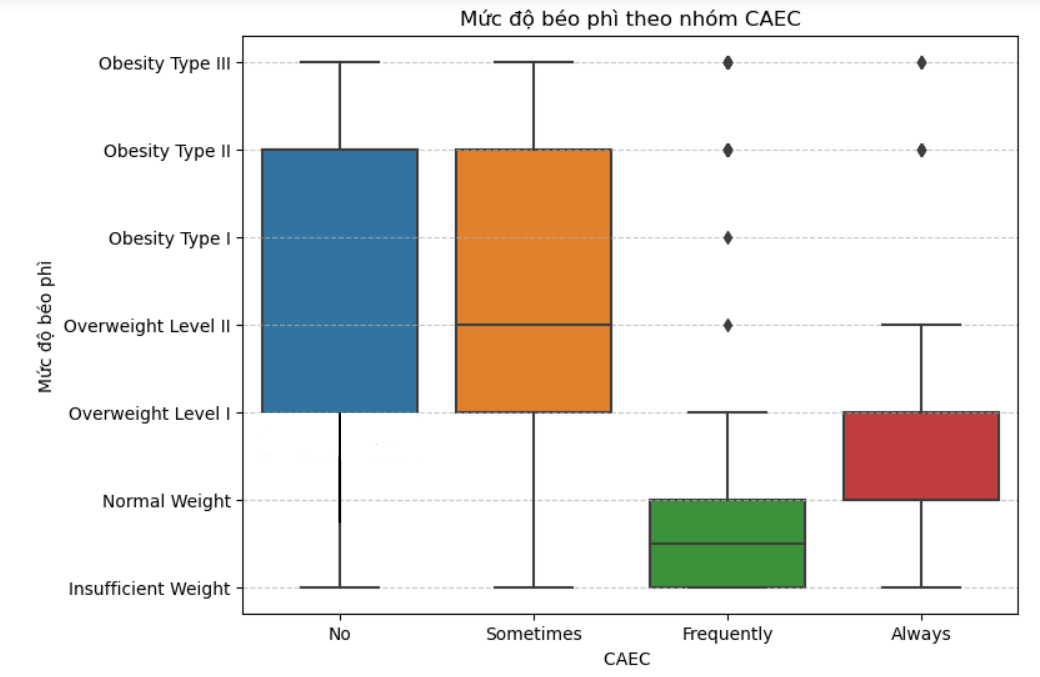


Dựa trên số liệu thống kê, có thể nhận thấy rằng có sự phân bố không đồng đều giữa các nhóm 'Sometimes', 'Frequently', 'Always' và 'No'. Cụ thể, nhóm 'Sometimes' có số lượng thống kê cao nhất, tiếp theo là 'Frequently', sau đó là 'Always' và 'No' đều có số lượng thống kê thấp không đáng kể, chỉ nhỏ hơn 50. Giá trị "sometimes" có thể chiếm nhiều hơn so với "frequently", "always", và "no" vì nó phản ánh một mẫu sống phổ biến trong xã hội hiện đại, nhiều người có thể có thói quen ăn uống linh hoạt, không đặt ra một lịch trình ăn cố định. Họ có thể ăn giữa các bữa ăn chính chỉ khi cảm thấy đói hoặc có cơ hội bởi cuộc sống hiện nay khá là bận rộn và làm như thế thể tiết kiệm được kha khá thời gian. Hơn nữa các mẫu mã đồ ăn vặt thường được quảng cáo và trờ nên phổ biến rộng rãi

Sau khi thực hiện kiểm định ANOVA cho 4 nhóm mức tiêu thụ ta có được giá trị sau:  
*Giá trị F-statistic: 114.29735393628926*

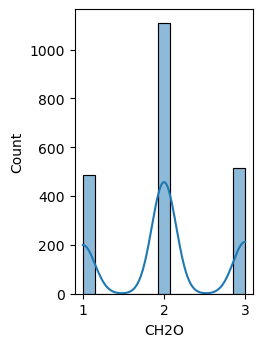
*p-value: 1.4205872553744423e-68*

Kết quả kiểm định ANOVA cho thấy có sự khác biệt đáng kể giữa ít nhất một cặp nhóm dựa trên thuộc tính CAEC. Giá trị F-statistic là 114.30 và p-value là khoảng 1.42e-68, rất nhỏ hơn ngưỡng ý nghĩa thống kê alpha = 0.05. Do đó, chúng ta có đủ bằng chứng để bác bỏ giả thuyết H0, tức là có ít nhất một cặp nhóm có sự khác biệt đáng kể về mức độ béo phì dựa trên mức độ tiêu thụ thực phẩm giữa các bữa ăn.

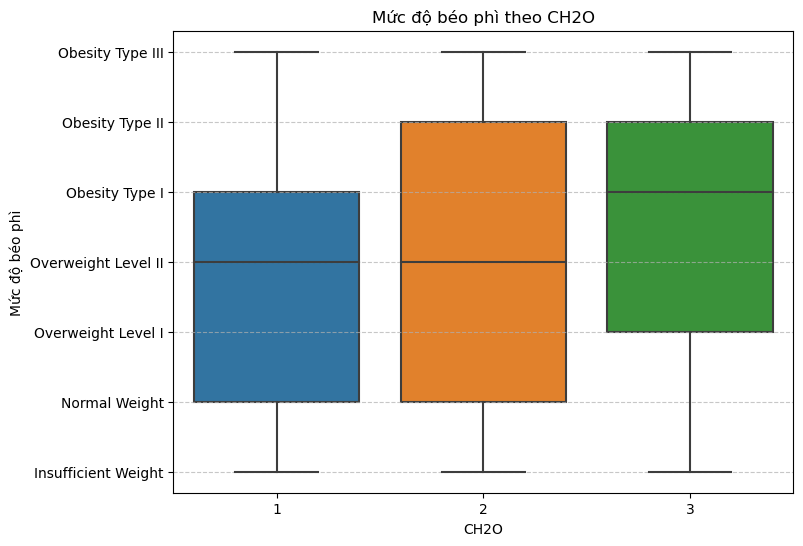


Qua biểu đồ hộp trên ta thấy được sự phân bố của các nhóm có sự khác biệt rõ rệt của ít nhất một cặp nhóm, điều này đúng với kiểm định đã thực hiện ở trên. Kết quả này cho thấy rằng mức độ tiêu thụ thực phẩm giữa các bữa ăn có thể ảnh hưởng đến mức độ béo phì của một cá nhân. Việc này có thể phản ánh thói quen ăn uống hàng ngày và cách tổ chức chế độ dinh dưỡng có thể ảnh hưởng đến trạng thái sức khỏe và cân nặng của mỗi người. Từ kết quả này, chúng ta có thể suy ra rằng thói quen tiêu thụ thực phẩm nhiều calo thường xuyên có thể gắn liền với mức độ béo phì cao hơn so với nhóm không thường xuyên. Điều này có thể phản ánh thực tế rằng việc tiêu thụ nhiều calo thường xuyên có thể dẫn đến tích tụ năng lượng dư thừa trong cơ thể và góp phần vào mức độ béo phì. Vậy thuộc tính CAEC có thể có ảnh hưởng nhất định đến tập dữ liệu.

### 10. CH20 - Consumption of water daily (n/s) - Tiêu thụ nước hàng ngày(lít): 1, 2, 3

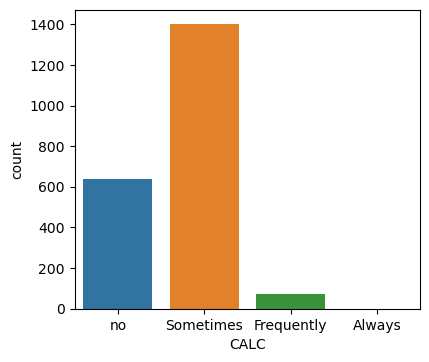


Dựa trên số liệu, có thể thấy được rằng số người uống 2 lít nước mỗi ngày nhiều hơn gấp đôi so với nhóm người uống 1 và 3 lít mỗi ngày. Điều này có thể hiểu được là do 2 lít nước mỗi ngày là giá trị cần thiết để cơ thể vận động bình thường và cảm thấy thoải mái.

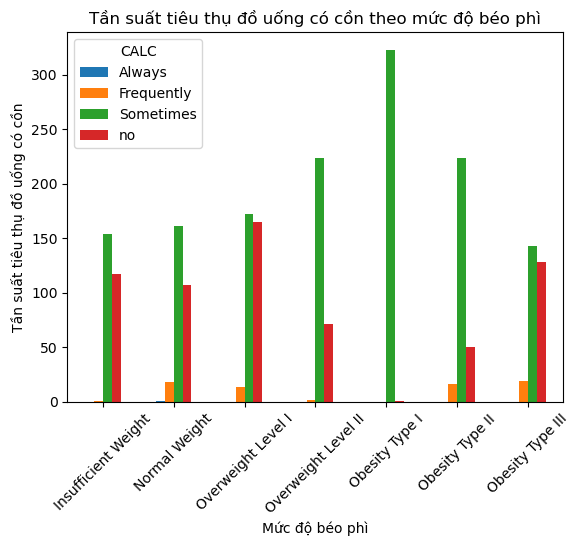


Biểu đồ mức độ béo phì theo thuộc tính CH2O cho ta thấy được sự phân bố giữa các nhóm không có sự khác biệt đáng kể, cả ba nhóm đều tập trung chủ yếu từ OverWeight loại 1 tới Obesity loại 2. Ngoài ra, biểu đồ còn cho thấy rằng người có xu hướng uống nhiều nước hơn sẽ dễ có các mức độ béo phì nghiêm trọng hơn, điều này phản ánh sai với hiện thực rằng uống nước có ảnh hưởng tới việc trao đổi chất của cơ thể cũng như giảm cảm giác đói. Vậy nên thuộc tính CH2O có thể có ảnh hưởng không tốt đến tập dữ liệu.

### 11. CALC - Consumption of alcohol (n/s) - Tiêu thụ đồ uống có cồn: no (tôi không uống), sometimes (đôi khi), frequently (thường xuyên), always (luôn luôn)



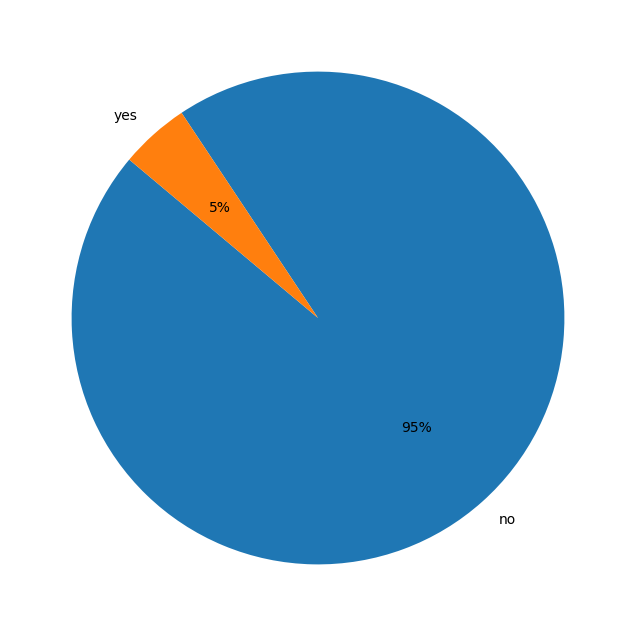
Dựa vào biểu đồ trên, ta có thể thấy nhóm người đôi khi sử dụng đồ uống có cồn chiếm phần lớn thống kê, điều này có thể liên quan đến việc rất nhiều người chỉ chọn uống trong các dịp đặc biệt hoặc để thư giãn sau một ngày làm việc. Trong khi đó, không có thống kê cho số người luôn luôn sử dụng đồ uống có cồn cho thấy rằng trong mẫu dữ liệu này, không có ai thực sự uống rượu liên tục và không bao giờ ngừng.



Từ biểu đồ về tần suất tiêu thụ đồ uống có cồn theo mức độ béo phì ta có thể thấy thấy được giá trị Sometimes có xu hướng phổ biến hơn trong các cột mức độ béo phì nghiêm trọng hơn và giá trị no phổ biến với các cột mức độ béo phì nhẹ hơn. Liên hệ đến thực tế, mặc dù đồ uống có cồn có thể liên quan đến béo phì qua các tác động của cồn đến quá trình hấp thụ cũng như chuyển hóa calo và dinh dưỡng nhưng nó không phải là nguyên nhân chính dẫn đến béo phì bởi cần xem xét thêm các thuộc tính khác như di truyền hay tần suất vận động thể chất.

## 4.3 - Các thuộc tính liên quan đến tình trạng thể chất:

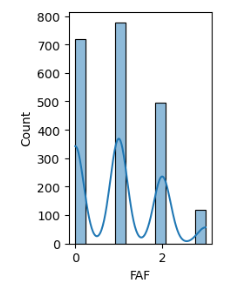
### 12. SCC - Calories consumption monitoring (n/s) - theo dõi lượng calo tiêu thụ: yes (có), no(không)



*Biểu đồ tròn phân phối thuộc tính SCC - theo dõi lượng calo tiêu thụ*

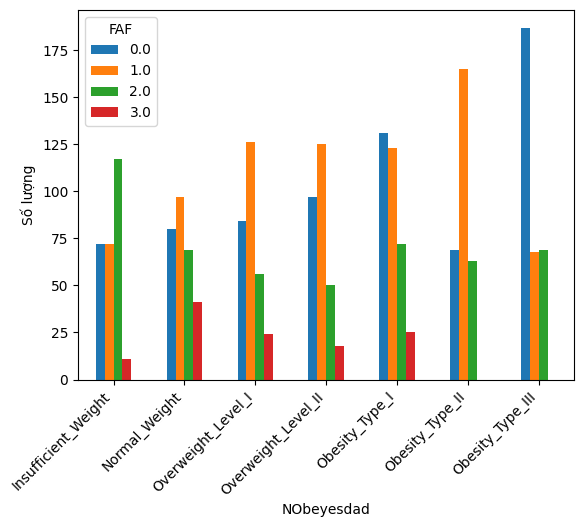
Sự phân bổ không đều giữa nhóm có và không theo dõi lượng calo tiêu thụ, nhóm theo dõi calo tiêu thụ thấp hơn rất nhiều so với nhóm không theo dõi. Thực tế, việc này phụ thuộc vào nhiều yếu tố như nhận thức về sức khỏe, giáo dục, văn hóa và tình trạng sức khỏe cá nhân, nhưng nó thường không phải là một hành vi phổ biến trong cộng đồng. Chỉ một phần nhỏ của dân số thực sự quan tâm và đặc biệt theo dõi lượng calo tiêu thụ hàng ngày. Các nghiên cứu và khảo sát thường chỉ ra rằng tỉ lệ người thực hiện việc giám sát calo chủ yếu là những người đang có mục tiêu cụ thể về giảm cân hoặc duy trì cân nặng. Việc giám sát calo có thể giúp một số người duy trì hoặc giảm cân bằng cách làm cho họ nhận thức rõ ràng hơn về lượng calo họ tiêu thụ và tiêu hao. Điều này có thể dẫn đến việc thay đổi lối sống và chế độ ăn uống để đạt được mục tiêu cân nặng. Tuy nhiên, mỗi cá nhân có những yếu tố riêng biệt ảnh hưởng đến nguy cơ béo phì, bao gồm di truyền, thói quen ăn uống, hoạt động thể chất và môi trường sống. Do đó, một phần lớn nguy cơ béo phì có thể được quyết định bởi những yếu tố này hơn là việc giám sát calo. Vậy việc giám sát calo tiêu thụ là một công cụ hữu ích để quản lý cân nặng nhưng không thể xem đây là nguyên nhân chính gây ra béo phì.

### 13. FAF - Physical activity frequency (n/s) - tần suất hoạt động thể chất: I do not have (tôi không có), 1 or 2 days (1 hoặc 2 ngày), 2 or 4 days (2 hoặc 4 ngày), 4 or 5 days (4 hoặc 5 ngày)



*Biểu đồ phân phối tần suất hoạt động thể chất*

Sự phân bố số lượng của nhóm không hoạt động thể chất thể hiện ở cột đầu tiên, cột thứ hai ba và bốn là sự phân bổ số lượng của nhóm hoạt động thể chất ở các mức độ một hai ba tương ứng với 1 -2 ngày /tuần, 2-4 ngày/tuần , 4-5 ngày/tuần. Số lượng người hoạt động thể chất tương đối cao nhưng chất lượng hoạt động thể chất chưa cao cụ thể tần suất hoạt động ở mức độ vừa phải (2- 4 ngày/tuần) hoặc ở mức độ thường xuyên(4-5 ngày/tuần) còn thấp, điều này có thể chỉ ra rằng mặc dù có một số người tham gia hoạt động thể chất, nhưng họ thường không hoạt động đủ để đạt được lợi ích sức khỏe lâu dài. Vậy việc không hoạt động thể chất và có hoạt động thể chất có ý đối quan trọng với béo phì hay không?



*Biểu đồ phân phối tần suất hoạt động thể chất theo từng nhóm*

Ta có thể quan sát thấy rằng mức độ không hoạt động thể chất ở nhóm người béo phì mức độ 1 cao hơn gấp rưỡi lần và nhóm béo phì mức độ 3 cũng hơn gấp hai lần so với nhóm người có cân nặng bình thường.

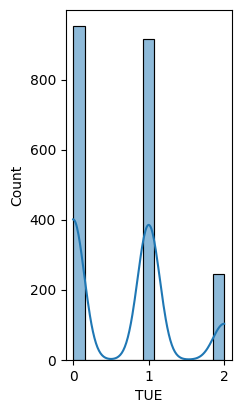
Với mức độ hoạt động 1 khoảng từ 1 đến 2 ngày trên một tuần nhóm béo phì loại 1 và 2 có tần suất cao hơn nhóm người bình thường, nhóm béo phì 3 lại thấp hơn nhóm bình thường. Mức độ hoạt động 2 khoảng từ 2 đến 4 ngày ở các nhóm béo phì 1 2 3 tương đối bằng với tần suất hoạt động của người bình thường. Nhưng ở mức độ hoạt động 3 tức 4 hoặc 5 ngày trên tuần số lượng ở nhóm béo phì 1 thấp hơn so với người bình thường đặc biệt ở nhóm béo phì 2 và 3 không hoạt động ở mức 3( 4 hoặc 5 ngày trên tuần).

Tổ chức WHO và các các cơ quan y tế khác thường khuyến nghị người lớn tối thiểu 150 đến 300 phút hoạt động thể chất vừa phải mỗi tuần hoặc 75 đến 150 phút hoạt động thể chất mạnh mẽ hàng tuần (tương đương 4 hoặc 5 ngày trên một tuần) là điều cần thiết để ngăn ngừa tăng cân và cải thiện thể lực.

Tuy nhiên, đối với những người muốn giảm cân, ít nhất 200 đến 300 phút hoạt động thể chất vừa phải đến mạnh mẽ mỗi tuần được khuyến nghị để giảm cân.

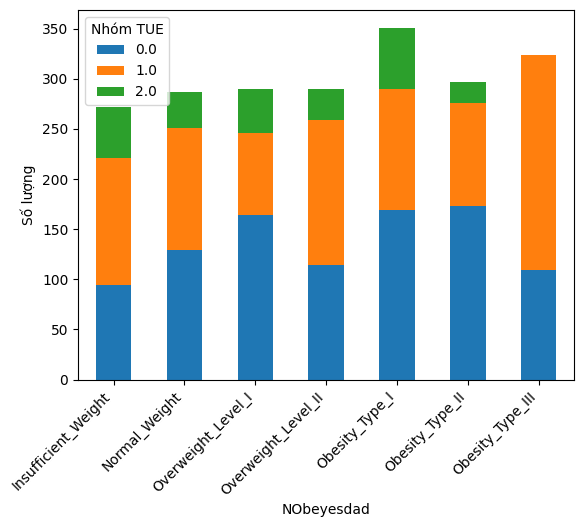
Rõ ràng, béo phì là kết quả của việc tiêu thụ năng lượng quá mức đã duy trì trong một thời gian dài. Hoạt động thể chất không phải là nguyên nhân chính gây béo phì, nhưng có vai trò trong việc quản lý cân nặng và cân bằng năng lượng. Trong một nghiên cứuvới 20 năm theo dõi, Hankison cho thấy việc duy trì mức hoạt động thể chất cao giúp giảm thiểu đáng kể việc tăng cân, đặc biệt là ở phụ nữ. Trong nghiên cứu đó, những người hoạt động đã tăng cân ít hơn trong thời gian tiếp theo so với những người thường xuyên không hoạt động. Những phát hiện tương tự đã được đưa ra từ báo cáo nhóm sinh đôi Phần Lan, có thể được coi là một phương pháp thử nghiệm tự nhiên để điều tra vai trò của hoạt động thể chất và gen hoặc các yếu tố di di truyền khác đối với tỷ lệ mắc bệnh trong tương lai. Cặp song sinh trong suốt hơn 30 năm liên tiếp có sự khác biệt về hoạt động thể chất, khác biệt đáng kể với nhau về trọng lượng cơ thể và cấu trúc cơ thể; cặp song sinh hoạt động thể chất có trọng lượng cơ thể, BMI và phần trăm chất béo thấp hơn đáng kể, và họ có ít hơn (50%) mỡ nội tạng và gan so với cặp song sinh không hoạt động của họ. Bằng chứng này cho thấy rõ ràng rằng mức hoạt động thể chất cao hơn có liên quan đến việc giảm tỷ lệ tăng cân ngay cả khi xét đến yếu tố di truyền. Tăng hoạt động thể chất chắc chắn có thể tạo ra thâm hụt năng lượng thông qua tăng tiêu hao năng lượng. Vì lý do này, hoạt động thể chất có tiềm năng như là một phần của giải pháp cho dịch bệnh béo phì đang diễn ra.

### 14. TUE - Time using technology devices (n/s) - thời gian sử dụng các thiết bị công nghệ: 0–2 hours (0 đến 2 giờ), 3–5 hours (3 đến 5 giờ), more than 5 hours (hơn 5 giờ)



*Biểu đồ cột phân phối TUE - thời gian sử dụng thiết bị bị công nghệ*

Trong biểu đồ này, thời gian sử dụng được chia thành các khoảng: 0 - 2 giờ, 3-5 giờ và trên 5 giờ. Cột thứ nhất đại diện cho khoảng thời gian từ 0 đến 2 giờ, cột thứ hai đại diện cho khoảng thời gian từ 3 đến 5 giờ, và các cột còn lại đại diện cho thời gian sử dụng trên 5 giờ. Biểu đồ này nhấn mạnh rằng phần lớn thời gian sử dụng thiết bị công nghệ nằm trong khoảng từ 0 đến 5 giờ, trong khi thời gian sử dụng trên 5 giờ tương đối ít.



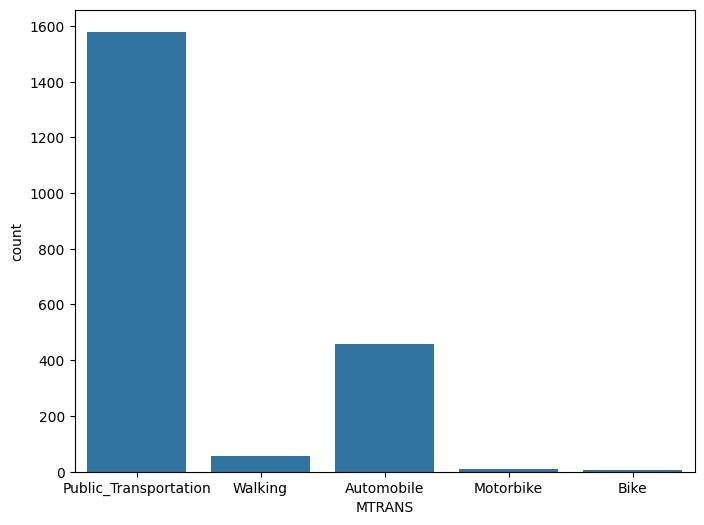
*Biểu đồ cột phân phối thời gian sử dụng thiết bị công nghệ theo từng nhóm béo phì*

Biểu đồ phân bố thời gian sử dụng thiết bị công nghệ theo từng nhóm có thể hiển thị sự khác biệt rõ ràng giữa nhóm người béo phì và nhóm người có cân nặng bình thường. Trong trường hợp này, thời gian sử dụng ở ba nhóm người béo phì cao hơn so với nhóm người có cân nặng bình thường. Điều này có thể cho thấy rằng các nhóm người béo phì có xu hướng sử dụng thiết bị công nghệ nhiều hơn so với nhóm người có cân nặng bình thường.

Theo một nghiên cứu mới từ Trường Y tế Công cộng Harvard TH Chan. Thanh thiếu niên dành nhiều giờ mỗi ngày cho thiết bị công nghệ có thể dễ bị béo phì hơn những người không dành nhiều thời gian cho các thiết bị điện tử này,. Trong số gần 25.000 thanh thiếu niên Hoa Kỳ được theo dõi trong giai đoạn 2013-2015, 20% dành hơn năm giờ mỗi ngày trước màn hình.Thực chất, những thanh thiếu niên này có khả năng uống đồ uống có đường mỗi ngày cao gấp đôi và không ngủ đủ giấc hoặc tập thể dục - và có khả năng bị béo phì cao hơn 43% so với thanh thiếu niên dành ít thời gian hơn để sử dụng các thiết bị công nghệ.

Nghiên cứu dịch tễ học cho thấy rằng thanh thiếu niên sử dụng nhiều thời gian cho thiết bị công nghệ tiêu thụ ít rau củ và nhiều loại snack, đồ uống nhiều calo và thức ăn nhanh, do đó có tổng calo tiêu thụ cao hơn. Bên cạnh đó, người ta cho rằng việc sử dụng thiết bị thông minh có thể làm giảm thời gian dành cho hoạt động thể chất, đặc biệt là khi người dùng dành nhiều thời gian trước màn hình thay vì tham gia các hoạt động ngoài trời hoặc tập thể dục. Việc này có thể dẫn đến thiếu hoạt động thể chất và góp phần vào nguy cơ béo phì. Do đó yếu tố sử dụng nhiều thời gian cho thiết bị công nghệ này có thể được xem là một trong những yếu tố có tác động đến nguy cơ béo phì

### 15. MTRANS - Transportation used (n/s) - phương tiện di chuyển được sử dụng: automobile(ô tô), motorbike (xe máy), bike (xe đạp), public transportation (phương tiện công cộng),walking (đi bộ)

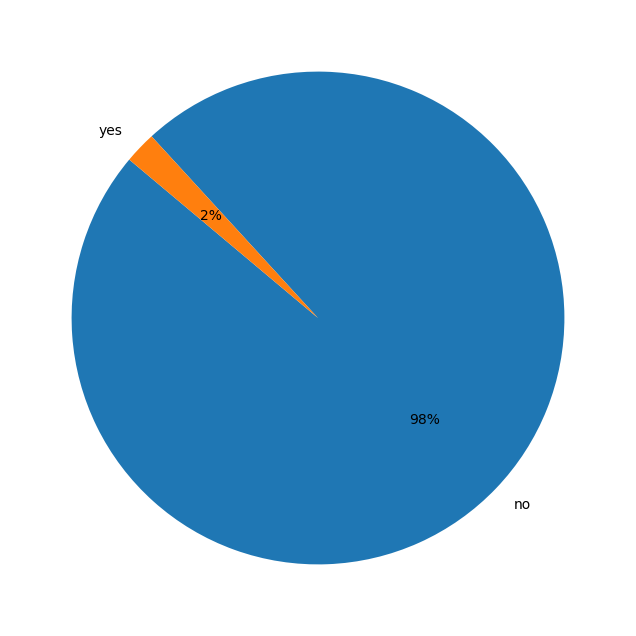


*Biểu đồ phân phối thuộc tính phương tiện di chuyển*

Phương tiện di chuyển được sử dụng nhiều nhất ở đây là phương tiện công cộng, một số ít là sử dụng phương tiện cá nhân là ô tô còn lại là đi bộ, xe đạp tương đối ít. Việc sử dụng phương tiện công cộng dường như phổ biến với tất cả các nhóm người từ thiếu cân tới béo phì.

Nghiên cứu của Ellen Flint và các đồng nghiệp về việc sử dụng phương tiện giao thông công cộng, xe đạp hoặc đi bộ so với ô tô để đi làm, được công bố trên The Lancet Public Health, là một tiến bộ lớn trong việc hiểu sự đóng góp của tất cả các phương thức hoạt động thể chất đối với béo phì và các kết quả sức khỏe khác. Nghiên cứu này định nghĩa đi lại tích cực là đi bộ và đi xe đạp, được nhóm lại với giao thông công cộng và so sánh với đi lại bằng ô tô. Như đã xem xét trong bài báo, một số nghiên cứu trước đây, bao gồm hai nghiên cứu tiền cứu, đã chỉ ra mối liên hệ giữa di chuyển tích cực và tình trạng cân nặng. Một thế mạnh quan trọng của nghiên cứu của Flint và các đồng nghiệp là việc sử dụng chỉ số khối cơ thể (BMI) được đo lường khách quan, được ưa thích hơn BMI tự báo cáo (self-reported) - một thước đo phổ biến trong các tài liệu hiện có. Với một mẫu gần 6000 người trưởng thành trong độ tuổi 40-69, các tác giả đã có thể kiểm tra mối quan hệ giữa những thay đổi trong chế độ đi lại và thay đổi BMI trong 4 năm. Những phát hiện chính là những người thay đổi từ đi lại tích cực hoặc công cộng sang đi lại bằng ô tô có chỉ số BMI tăng tương đối là 0.3 kg/m2 và những người thay đổi từ đi lại bằng ô tô sang đi lại tích cực hoặc công cộng có chỉ số BMI giảm tương đối là 0.3 kg /m2. Những kết quả này gần giống với một nghiên cứu trước đây sử dụng BMI tự báo cáo, củng cố niềm tin vào kết quả. Những thay đổi trong 4 năm về BMI, có vẻ nhỏ, nghiên cứu của Flint và các đồng nghiệp không phản ánh đầy đủ tác động tiềm tàng của hành vi di chuyển đối với BMI, do đó không thể kết luận mối liên hệ của yếu tố này với béo phì. Trong phạm vi dự án, yếu tố phương tiện di chuyển này không có ý nghĩa nhiều đối với dự án, có thể bỏ qua yếu tố này trong quá trình đánh giá các yếu tố ảnh hưởng béo phì.

### SMOKE (n/s) - hút thuốc: yes (có), no (không)



*Biểu đồ tròn phân phối thuộc tính SMOKE - hút thuốc*

Phân bố giữa 2 nhóm hút thuộc và không hút thuốc có sự chênh lệch lớn, thiên về nhóm không hút thuốc với 98% và 2% còn lại là nhóm hút thuốc.

Thực tế, trên toàn cầu, hơn 1,1 tỷ người từ 15 tuổi trở lên là người sử dụng thuốc lá. Từ tuổi trưởng thành trẻ, những người hút thuốc có xu hướng có trọng lượng cơ thể thấp hơn những người không hút thuốc. Những người trưởng thành hiện đang hút thuốc lá có xu hướng có chỉ số BMI thấp hơn so với những người chưa bao giờ hoặc những người đã từng hút thuốc. Tuy nhiên, trong số những người hút thuốc lá, 26,3% cũng bị béo phì. Do đó, trong khi trung bình, những người hút thuốc trưởng thành có xu hướng có trọng lượng cơ thể thấp hơn so với những người không hút thuốc, một số lượng đáng kể những người hút thuốc bị béo phì. Trong các nghiên cứu tiền lâm sàng, việc tiếp xúc với nicotin làm giảm lượng thức ăn ăn vào điều này có thể dẫn đến một lượng calo tiêu thụ ít hơn, giảm kích thước bữa ăn do tác động của nó đến hệ thống thần kinh và sự cảm nhận về cảm giác no do đó dẫn đến khoảng cách giữa các bữa ăn dài hơn. Hơn nữa, nicotin đã được chứng minh là có tác dụng gây chán ăn và làm thay đổi cách ăn uống. Mặc dù người ta thường thừa nhận rằng hút thuốc lá là một chất ức chế sự thèm ăn, nhưng tác động của việc hút thuốc đối với sự thèm ăn ở người không phải lúc nào cũng được quan sát thấy.

Để đánh giá tác động của thuộc tính hút thuốc đối với béo phì, cần thu thập thông tin chi tiết hơn về hút thuốc, bao gồm số lượng hút thuốc mỗi ngày, thời gian hút thuốc, thời gian đã hút thuốc và thời gian kể từ khi bỏ hút thuốc. Hiện tại, thông tin về tình trạng hút thuốc trong tập dữ liệu không đủ để đánh giá mối quan hệ một cách toàn diện. Việc bổ sung thông tin này sẽ giúp xây dựng giả thuyết tốt nhất và đưa ra kết luận chính xác hơn về mối liên hệ giữa hút thuốc và béo phì.

# 5. Kết luận

Từ việc phân tích, so sánh thực tế và kết hợp sử dụng các kiểm định như đã nêu ở trên ta có thể chọn ra những thuộc tính từ tập dữ liệu mà chúng ảnh hưởng lớn tới kết quả của mô hình hay nói theo một cách thực tế là các yếu tố và thói quen có thể ảnh hưởng đến mức độ béo phì của một người theo danh sách như sau:

1. 'Gender' : Giới tính
2. 'Age': Độ tuổi
3. 'Height': Chiều cao
4. 'Weight': Cân nặng
5. 'family\_history\_with\_overweight': tiền sử bị béo phì trong gia đình
6. 'FAVC': Tần suất tiêu thụ thức ăn nhiều calo
7. 'CAEC': Tần suất tiêu thụ thực ăn giữa các bữa ăn
8. 'FAF': Tần suất vận động thể chất
9. 'TUE': Thời gian sử dụng các thiết bị công nghệ

Ta thấy có tổng cộng 9 yếu tố khác nhau ảnh hưởng lớn tới việc đưa ra kết luận rằng một người đang ở mức độ béo phì nào. Ngoài chiều cao và cân nặng để tính chỉ số BMI như thông thường chúng ta thường dùng thì kết luận có thêm các yếu tố liên quan đến thói quen sinh hoạt, ăn uống hằng ngày cùng các yếu tố liên quan đến nhân khẩu như giới tính, độ tuổi, tiền sử bị béo phì các thành viên trong gia đình. Dựa vào đó, nghiên cứu có thể đưa ra lời khuyên cho mọi người để hạn chế và cải thiện được tình trạng béo phì là cần phải cân đối các thói quen như tần suất thụ thức ăn nhiều calo, tiêu thụ thức ăn giữa các bữa ăn, việc hoạt động thể chất và thời gian sử dụng các thiết bị công nghệ. Cụ thể hơn là ta nên hạn chế tiêu thụ thức ăn nhiều calo, có những bữa ăn phụ phù hợp với tần suất vận động của cơ thể, đối với những người bị thừa cân hay béo phí hãy giảm bớt thời gian sử dụng các thiết bị công nghệ giảm bớt các bữa ăn phụ và lượng calo tiêu thụ và tăng tần suất hoạt động thể chất, đối với những người thiếu cân thì hay làm ngược lại.

Từ tập dữ liệu đã được chọn lại như vậy thì mô hình học máy Gradient Boosting sẽ là một lựa chọn tối ưu để xây dựng mô hình dự đoán. Vì số lượng nhãn cần dự đoán lên tới 7 loại khác nhau và mô hình này cũng sẽ đạt được hiệu suất cao hơn nhờ việc sử dụng kết hợp nhiều cây quyết định theo cách tối ưu trong quá trình huấn luyện

Ngoài các thuộc tính có ảnh hưởng đáng kể đến mức độ béo phì đã phân tích ở trên, ta có thể nghiên cứu thêm về một số yếu tố khác như chất lượng giấc ngủ hay tình trạng tâm lý để có thêm cái nhìn toàn diện hơn. Bởi giấc ngủ ảnh hưởng đến sự thèm ăn cũng như sự trao đổi chất trong cơ thể trong khi tình trạng tâm lý tác động tới thói quen ăn uống và rèn luyện cơ thể. Để duy trì trạng thái cơ thể không bị béo phì, ta nên kiểm soát chất lượng giấc ngủ cũng như quản lý tốt tâm trạng.

Kết quả nghiên cứu này có thể cung cấp thông tin hữu ích cho các chương trình can thiệp và chính sách công cộng nhằm giảm nguy cơ béo phì và cải thiện sức khỏe cộng đồng. Các biện pháp can thiệp có thể tập trung vào việc thúc đẩy lối sống lành mạnh, tăng cường hoạt động vận động thể chất, và tăng ý thức về tác động của các thói quen hàng ngày đến sức khỏe.

# Tài liệu tham khảo

1. Wiklund, P. (no date) The role of physical activity and exercise in obesity and weight management: Time for critical appraisal, Journal of sport and health science. U.S. National Library of Medicine. Available at: [Link](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6188737/) (Accessed: 24 April 2024).
2. New evidence for the role of transportation in health. Available at: [Link](https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(16)30008-1/fulltext) (Accessed: 24 April 2024).
3. Robinson, T. N., Banda, J. A., Hale, L., Lu, A. S., Fleming-Milici, F., Calvert, S. L. and Wartella, E. (no date) Screen Media Exposure and Obesity in Children and Adolescents, Pediatrics. U.S. National Library of Medicine. Available at: [Link](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5769928/) (Accessed: 25 April 2024).