TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA TOÁN - ỨNG DỤNG

BÀI TIỂU LUẬN

ĐỀ BÀI: ÁP DỤNG MÔ HÌNH MÁY HỌC ĐỂ DỰ ĐOÁN KHẢ NĂNG BÉO PHÌ VÀ PHÂN TÍCH CÁC TÁC NHÂN GÂY RA BÉO PHÌ

HỌ VÀ TÊN : NGUYỄN ĐĂNG TIẾN

MSSV: 3123580050

HỌC PHẦN : MÁY HỌC

GIẢNG VIÊN:

TS. VŨ NGỌC THANH SANG

Thành phố Hồ Chí Minh - 2025

LỜI CẢM ƠN

Trước tiên, em xin trân trọng gửi lời cảm ơn sâu sắc đến cán bộ, công nhân viên chức trường đang công tác và làm việc tại trường Đại học Sài Gòn đã tạo điều kiện cung cấp các trang thiết bị cần thiết cũng như đã tạo môi trường cho em để có thể hoàn thành bài tiểu luận lần này.

Em xin bày tỏ lòng biết ơn đặc biệt sâu sắc đến với thầy **Vũ Ngọc Thanh Sang** - giảng viên bộ môn Máy Học - người đã tận tình hướng dẫn, hỗ trợ và đồng hành cùng em trong suốt quá trình thực hiện bài luận.

Do còn hạn chế về kiến thức cũng như kinh nghiệm. Nên bài tiểu luận của em không thể tránh khỏi những điểm sai sót cũng như chưa hoàn thiện. Vì thế mà em rất mong có thể nhận được những ý kiến đóng góp từ các quý thầy cô cùng với quý đọc giả nhằm giúp bài tiểu luận của em có cơ hội được cải thiện cũng như hoàn thiện hơn

Cuối cùng em xin chúc các quý thầy cô khỏe mạnh, năng động tràn đầy niềm tin để có thể vững vàng mang đến những giá trị cao quý của tri thức cho những lớp trẻ tương lai mai sau.

Hồ Chí Minh, tháng 5/2025

Sinh viên

Nguyễn Đăng Tiến

DANH MỤC CÁC KỲ HIỆU VÀ CHỮ VIẾT TẮT

Chữ viết tắt	Nguyên mẫu
AI	Artificial Intelligence
BMI	Body Mass Index)
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
ML	Machine Learning
WHO	World Health Organization,

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH ẢNH

TÓM TẮT

Theo Tổ chức Y tế Thế Giới (World Health Organization -WHO) cứ mỗi năm số người mắc các bệnh béo phì lại không ngừng tăng thậm chí là còn tăng nhanh tăng mạnh. Béo phì là một căn bệnh mãn tính phức tạp. Đặc biệt là trong xã hội lười vận động xuất hiện nhiều đồ ăn chế biến sẵn như ngày nay. Việc ta cần nhìn nhận một cách nghiêm túc cũng như quan tâm sát sao về vấn đề này càng được đề cao và cần phải được dự đoán sớm

Trước công cuộc xây dựng đổi mới đời sống trong xã hội ta cũng đã có cơ hội được tiếp xúc với các mô hình học máy hay học sâu. Chúng giúp phân loại hình ảnh, phát hiện lỗi sai trong bài kiểm tra hay chỉ đơn thuần là dự đoán cổ phiếu, dự đoán một điều gì đó khi ta có dữ liệu được tổ chức thành một tập có sẵn

Trong bài tiểu luận này miêu tả trình bày việc sử dụng các mô hình Học máy để dự đoán phân loại các mức độ béo phì hay thừa cân ở độ tuổi người trưởng thành lấy các yếu tố đầu vào như chế độ ăn uống, tần suất sử dụng mạng điện tử,… Qua đó cũng phân tích đánh giá sâu nhằm mục đích trả lời cho các câu hỏi. Tác nhân nào là quan trọng nhất đến với khả năng mắc béo phì ở người trường thành ?. Độ tuổi nào có khả năng mắc các căn bệnh béo phì nhiều nhất ?. Hay liệu một người bình thường có nguy cơ mắc béo phì hay không ?.

Bài luận này sử dụng tập dữ liệu công khai và thử nghiệm với các mô hình phân loại như **Decision Tree (Cây quyết định), Random Forest và Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN)** mục đích chính là phân loại và dự đoán. Bên cạnh đó bằng cách sử dụng các công thức, phương pháp đánh giá mô hình để thực hiện việc so sánh, từ đó có thể lựa chọn với mô hình nào thì việc dự doán phân loại nào là tốt nhất. Không những vậy bài luận cũng sử dụng các phương pháp chọn lọc đặc trưng nhằm xác định các yếu tố quan trọng ảnh hưởng đến biến mục tiêu (các phân loại béo phì). Các kỹ thuật sử dụng biểu đồ từ thư viện Pandas, Numpy, Matplotlib nhằm phân tích sâu về phân bố với các biến số, so sánh các biến phân loại, mối quan hệ giữa các đặc trưng với nhau

Kết quả cho thấy các mô hình hoạt động tốt nhất với độ chính xác nằm trong khoảng từ 80 đến gần 95% chính là những mô hình như Decision Tree và Random Forest các mô hình được cho là hoạt động kém với độ chính xác nằm trong khoảng từ 70 đến gần 80% chính là những mô hình như Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN). Ngoài ra các yếu tố có thể ảnh hưởng đến mức độ béo phì cao nhất chính là cân nặng tiếp theo sau lần lượt là chiều cao, giới tính và tuổi. Tuy cần phải trải qua nhiều yếu tố xác thực cũng như chọn lọc đánh giá cẩn thận để có thể đưa vào thực tế, nhưng đây vẫn là một kết quả có thể tham khảo nhằm mục đích hỗ trợ việc chẩn đoán và điều chỉnh lối sống trong tương lai

MỤC LỤC

Chương I

Giới thiệu

1. Bối cảnh và ý nghĩa vấn đề

Từ xưa đến nay, dinh dưỡng luôn đóng vai trò thiết yếu trong sự phát triển toàn diện của con người. Sở dĩ dinh dưỡng được xem là yếu tố quan trọng là bởi nó cung cấp năng lượng cho các hoạt động thể chất và tinh thần, giúp cơ thể tỉnh táo, khỏe mạnh và đạt hiệu suất cao trong công việc cũng như học tập. Không thể phủ nhận rằng dinh dưỡng là nền tảng không thể thiếu để duy trì một lối sống lành mạnh và cân bằng.

Chính vì vậy, việc kiểm soát sức khỏe, cân nặng và duy trì vóc dáng cần được quan tâm đúng mức. Những hành động tưởng chừng đơn giản như lựa chọn bữa ăn phù hợp với thể trạng, sắp xếp thời gian nghỉ ngơi hợp lý hay duy trì thói quen tập luyện đều hướng đến một mục tiêu chung: nâng cao chất lượng cuộc sống và hiệu quả hoạt động hàng ngày.

Một bữa ăn đầy đủ dưỡng chất không chỉ giúp tăng cường hệ miễn dịch mà còn làm giảm nguy cơ mắc các bệnh không lây nhiễm như tiểu đường, tim mạch và hỗ trợ kéo dài tuổi thọ. Bên cạnh đó, một thời gian biểu hợp lý, cân bằng giữa nghỉ ngơi và vận động, sẽ giúp lưu thông máu tốt hơn, điều hòa hoạt động các cơ quan nội tạng, từ đó tăng cường sức đề kháng và phòng ngừa bệnh tật.

Từng hành động nhỏ ấy, khi được duy trì đều đặn và có chủ đích, sẽ góp phần xây dựng nên một "công thức sống khỏe" – một nền tảng vững chắc cho một cơ thể khỏe mạnh, bền bỉ và tràn đầy năng lượng.

Tuy nhiên, trong xã hội hiện đại ngày nay, con người ngày càng phụ thuộc vào các thiết bị điện tử, trong khi mức độ vận động lại ngày một suy giảm. Thay vì duy trì lối sống năng động, nhiều người đang dần hình thành những thói quen thiếu lành mạnh: sử dụng thực phẩm chế biến sẵn, ăn uống không kiểm soát, nghỉ ngơi thiếu khoa học và lười vận động. Đáng lo ngại là những xu hướng tiêu cực này đang ngày càng lan rộng, góp phần tạo nên một xã hội ít vận động và có nguy cơ cao mắc các bệnh không lây nhiễm như tim mạch, máu nhiễm mỡ, tiểu đường,… Trong số đó, béo phì chính là hệ quả rõ nét và phổ biến nhất – một "đại dịch âm thầm" nhưng tiềm ẩn nhiều rủi ro nghiêm trọng, đang trở thành vấn đề nhức nhối của toàn cầu.

Béo phì không chỉ ảnh hưởng đến sức khỏe thể chất mà còn làm suy giảm nghiêm trọng chất lượng cuộc sống. Đây là một bệnh lý mãn tính phức tạp, được đặc trưng bởi sự tích tụ mỡ quá mức trong cơ thể, từ đó làm tăng nguy cơ mắc nhiều bệnh lý nguy hiểm như tiểu đường tuýp 2, tim mạch, ung thư và các vấn đề về xương khớp. Ngoài ra, béo phì còn gây khó khăn trong vận động, làm rối loạn giấc ngủ và tác động tiêu cực đến sức khỏe tâm thần (World Health Organization, 2024b).triệu thanh thiếu niên sống chung với béo phì. Béo phì không chỉ là vấn đề y tế mà còn tạo ra gánh nặng kinh tế lớn, dự báo chi phí toàn cầu sẽ lên đến 3 nghìn tỷ USD mỗi năm vào năm 2030 và tăng lên 18 nghìn tỷ USD vào năm 2060 nếu không có biện pháp ngăn chặn kịp thời (World Health Organization, 2024b).

Theo số liệu của Tổ chức Y tế Thế giới (WHO), vào năm 2022 có khoảng 2,5 tỷ người trưởng thành bị thừa cân, trong đó hơn 890 triệu người sống chung với béo phì. Tỷ lệ thừa cân toàn cầu đã tăng từ 25% năm 1990 lên 43% năm 2022, đặc biệt rõ nét tại các quốc gia có thu nhập thấp và trung bình. Không chỉ người lớn, tình trạng thừa cân và béo phì cũng đang lan rộng ở trẻ em và thanh thiếu niên. Năm 2022, khoảng 37 triệu trẻ em dưới 5 tuổi bị thừa cân, trong khi ở độ tuổi 5–19, số lượng thanh thiếu niên béo phì đã tăng từ 2% năm 1990 lên 8% năm 2022, tương đương khoảng 160 triệu người trẻ.

Béo phì không chỉ là một vấn đề sức khỏe mà còn tạo ra gánh nặng kinh tế to lớn. WHO ước tính nếu không có các biện pháp can thiệp hiệu quả, chi phí liên quan đến béo phì có thể lên tới 3.000 tỷ USD mỗi năm vào năm 2030, và có thể tăng đến 18.000 tỷ USD vào năm 2060 (World Health Organization, 2024b). Những con số này cho thấy béo phì đang ngày càng trở thành một thách thức nghiêm trọng đối với sức khỏe cộng đồng toàn cầu – một hệ quả tất yếu của lối sống thiếu lành mạnh và vận động, vốn đang ngày càng phổ biến trong xã hội hiện đại.

Bên cạnh béo phì, suy dinh dưỡng (thiếu cân) cũng đang là một "quả bom nổ chậm" đe dọa đến sức khỏe cộng đồng, đặc biệt là ở các quốc gia có thu nhập thấp và trung bình. Mặc dù thường ít được chú ý hơn so với tình trạng thừa cân, suy dinh dưỡng lại là nguyên nhân sâu xa gây ra nhiều hệ lụy nghiêm trọng như suy giảm miễn dịch, chậm phát triển thể chất và trí tuệ, và tăng nguy cơ tử vong, đặc biệt ở trẻ em.

Theo ước tính của Tổ chức Y tế Thế giới (WHO) và Quỹ Nhi đồng Liên Hợp Quốc (UNICEF), tính đến năm 2022, có khoảng 148 triệu trẻ em dưới 5 tuổi trên toàn cầu bị thấp còi do suy dinh dưỡng mãn tính, và 45 triệu trẻ em bị gầy còm – một chỉ dấu của tình trạng thiếu cân cấp tính. Ngoài ra, hàng triệu người trưởng thành, đặc biệt là phụ nữ mang thai và người cao tuổi, cũng đang sống trong tình trạng thiếu hụt năng lượng và vi chất, dẫn đến các vấn đề sức khỏe lâu dài.

Điều đáng nói là, trong khi một bộ phận dân số phải đối mặt với nguy cơ thừa cân và béo phì, thì một bộ phận khác lại đang thiếu hụt dinh dưỡng nghiêm trọng. Sự tồn tại song song của hai thái cực này – còn gọi là gánh nặng kép về dinh dưỡng – đặt ra một thách thức lớn cho hệ thống y tế và các chính sách chăm sóc sức khỏe toàn cầu.

2. Mục tiêu nghiên cứu

Ngày nay, các công nghệ như máy học (machine learning), trí tuệ nhân tạo (AI) và học sâu (deep learning) đang phát triển mạnh mẽ, đạt được nhiều thành tựu nổi bật trong nhiều lĩnh vực, đặc biệt là trong y học. Nhờ sự tiến bộ này, việc phân loại và dự đoán các loại bệnh lý trở nên ngày càng chính xác, nhanh chóng và phổ biến hơn. Từ thực tế đó, đề tài "Áp dụng mô hình máy học để dự đoán khả năng béo phì và phân tích các tác nhân gây ra béo phì" được lựa chọn nhằm góp phần làm sáng tỏ một vấn đề sức khỏe đang nhức nhối hiện nay.

Thông qua bài tiểu luận này, người đọc sẽ có cái nhìn toàn diện hơn về các yếu tố ảnh hưởng đến nguy cơ béo phì hay thiếu cân, đồng thời tiếp cận được quy trình phân tích dữ liệu và xây dựng mô hình học máy trong thực tiễn. Mục tiêu của bài viết không chỉ dừng lại ở việc áp dụng lý thuyết vào thực hành mà còn nhằm nâng cao

nhận thức cộng đồng về tầm quan trọng của việc kiểm soát cân nặng, từ đó góp phần xây dựng một lối sống khoa học, lành mạnh và bền vững hơn cho xã hội.

3. Phạm vi nghiên cứu

Phạm vi nghiên cứu của bài tiểu luận này tập trung vào việc ứng dụng các thuật toán máy học để dự đoán nguy cơ béo phì dựa trên các đặc trưng liên quan đến lối sống, nhân khẩu học và thói quen ăn uống của từng cá nhân. Dữ liệu sử dụng trong nghiên cứu là một tập dữ liệu đã được thu thập và xử lý sẵn, bao gồm các biến số như: tuổi, giới tính, tần suất vận động thể chất, chế độ ăn uống, thời gian sử dụng thiết bị điện tử, tình trạng hút thuốc, và các thói quen sinh hoạt khác.

Nghiên cứu không đi sâu vào phân tích lâm sàng hay can thiệp y tế mà chủ yếu khai thác khả năng phân loại và dự đoán bằng thuật toán, nhằm mục tiêu xây dựng một mô hình có thể đưa ra nhận định ban đầu về nguy cơ béo phì. Ngoài ra, phạm vi cũng giới hạn ở việc sử dụng một số thuật toán máy học phổ biến như: Decision Tree, Random Forest, KNN, Logistic Regression và SVM, từ đó so sánh hiệu quả dự đoán giữa các mô hình cũng như sử dụng các phương pháp chọn lọc đặc trưng nhằm xem xét đánh giá xem nguyên nhân nào tác động mạnh đến với nguy cơ béo phì hay thiếu cân.

Phạm vi nghiên cứu không bao gồm việc triển khai mô hình trong môi trường thực tế hay các hệ thống hỗ trợ y tế. Thay vào đó, trọng tâm chính là phân tích dữ liệu, tiền xử lý, huấn luyện và đánh giá mô hình trên bộ dữ liệu mẫu, nhằm kiểm chứng tính khả thi và tiềm năng ứng dụng của học máy trong lĩnh vực y tế, cụ thể là dự đoán béo phì

Chương II

Cơ sở lý thuyết

1. Khái niệm về béo phì

Trong thực tế, béo phì được xem là một căn bệnh mãn tính nguy hiểm, thường xuất phát từ lối sống ít vận động, chế độ ăn uống không hợp lý và thiếu kiểm soát cân nặng. Khi năng lượng nạp vào vượt quá năng lượng tiêu hao trong thời gian dài, mỡ sẽ tích tụ dần trong cơ thể và dẫn đến tình trạng béo phì. Đây là yếu tố làm gia tăng nguy cơ mắc các bệnh lý nghiêm trọng như tim mạch, tiểu đường tuýp 2, rối loạn chuyển hóa, và nhiều bệnh không lây nhiễm khác. Ngoài ra, béo phì còn ảnh hưởng đến sức khỏe tâm thần, làm suy giảm chất lượng cuộc sống và gây ra gánh nặng lớn cho nền kinh tế và hệ thống y tế.

Theo định nghĩa của Tổ chức Y tế Thế giới (WHO), béo phì được đánh giá dựa trên chỉ số khối cơ thể (BMI - Body Mass Index). Đây là một chỉ số được tính toán từ cân nặng và chiều cao của một người theo công thức:

$$BMI = \frac{Cân\ nặng\ (kg)}{(Chiều\ cao\ (cm))^2}$$

Dựa trên giá trị BMI, WHO phân loại như sau:

Phân loại	BMI (kg/m²)
Cân nặng bình thường	18,5 – 24,9
Thừa cân (tiền béo phì)	25,0 – 29,9
Béo phì độ I (nhẹ)	30,0 – 34,9
Béo phì độ II (vừa)	35,0 – 39,9
Béo phì độ III (nặng/bệnh lý)	≥ 40,0

Bảng 2.1 Phân chia béo phì theo BMI

Giải thích các mức độ:

- **Thừa cân (BMI 25,0 – 29,9):** Là giai đoạn tiền béo phì, cảnh báo nguy cơ chuyển sang béo phì thực sự nếu không kiểm soát kịp thời.
- **Béo phì độ I (30,0 – 34,9):** Mức độ nhẹ, nhưng đã làm tăng đáng kể nguy cơ mắc các bệnh không lây nhiễm.
- **Béo phì độ II (35,0 – 39,9):** Mức độ trung bình, liên quan chặt chẽ đến các biến chứng tim mạch, nội tiết và xương khớp.
- **Béo phì độ III (≥ 40):** Béo phì bệnh lý, rất nguy hiểm, thường đi kèm với nhiều bệnh nền nghiêm trọng và giảm đáng kể tuổi thọ.

Lưu ý:

Chỉ số BMI chỉ mang tính chất tham khảo đối với người trưởng thành. Đối với trẻ em và thanh thiếu niên, việc đánh giá tình trạng cân nặng cần dựa trên đồ thị tăng trưởng theo độ tuổi và giới tính.

Ở một số dân tộc (ví dụ người châu Á), ngưỡng BMI đánh giá nguy cơ sức khỏe có thể thấp hơn do sự khác biệt về tỷ lệ mỡ cơ thể.

Béo phì không chỉ là vấn đề về ngoại hình mà còn là một rối loạn chuyển hóa nghiêm trọng, liên quan đến nhiều biến chứng y khoa như đột quỵ, ung thư, thoái hóa khớp, các bệnh hô hấp và nội tiết. Ngoài ra, tình trạng này còn tác động tiêu cực đến giấc ngủ, khả năng vận động, và sức khỏe tâm lý, đặc biệt làm tăng cảm giác tự ti và nguy cơ trầm cảm.

Trong những năm gần đây, béo phì đã trở thành một vấn đề y tế công cộng toàn cầu, với tỷ lệ gia tăng nhanh chóng ở cả người lớn và trẻ em. Nguyên nhân chủ yếu đến từ sự thay đổi lối sống hiện đại: con người ngày càng phụ thuộc vào thiết bị công nghệ, giảm vận động thể chất, tiêu thụ thực phẩm chế biến sẵn nhiều đường, muối và chất béo bão hòa. Ngoài ra, yếu tố di truyền, nội tiết, và tâm lý cũng góp phần làm tăng nguy cơ béo phì ở nhiều nhóm dân số.

2. Giới thiệu học máy

2.1 Khái niệm học máy

Học máy (Machine Learning) là một nhánh của trí tuệ nhân tạo (AI), cho phép máy tính tự học hỏi từ dữ liệu mà không cần lập trình chi tiết từng bước. Thay vì viết mã cụ thể để xử lý từng tình huống, học máy sử dụng các thuật toán nhằm phát hiện các mô hình (patterns) hoặc quy luật ẩn trong dữ liệu, từ đó đưa ra các dự đoán hoặc quyết định dựa trên dữ liệu mới.

"Lĩnh vực nghiên cứu giúp máy tính có khả năng học mà không cần được lập trình rõ ràng"

Một định nghĩa khác hiện đại hơn định nghĩ trên:

"Một chương trình máy tính được gọi là học từ kinh nghiệm E đối với nhiệm vụ T và thước đo hiệu suất P, nếu hiệu suất của nó đối với T, được đo bằng P, cải thiện qua kinh nghiệm E."

-Mitchell, T. M. (1997). Machine Learning. McGraw Hill.-

Sự ra đời và phát triển của học máy được xem như một trong những bước tiến quan trọng của cuộc cách mạng công nghiệp lần thứ tư, mang lại giá trị to lớn trong lĩnh vực trí tuệ nhân tạo. Hiện nay, học máy đang len lỏi vào nhiều khía cạnh của cuộc sống hàng ngày mà chúng ta đôi khi không hề nhận ra. Chẳng hạn như trong các lĩnh vự như:

- **Y tế:** Giúp chẩn đoán bệnh, dự đoán nguy cơ mắc bệnh, phân tích hình ảnh y khoa.
- **Tài chính:** Dự đoán thị trường chứng khoán, phát hiện gian lận thẻ tín dụng.
- **Giao thông:** Dự đoán lưu lượng, tối ưu hóa lộ trình, phát triển xe tự lái.
- **Thương mại điện tử:** Đề xuất sản phẩm, cá nhân hóa trải nghiệm khách hàng.
- **Xử lý ngôn ngữ tự nhiên:** Nhận diện giọng nói, dịch máy, chatbot.
- **An ninh mạng:** Phát hiện tấn công, lọc spam và các hành vi bất thường.
- **Ứng dụng trong game:** Máy chơi cờ hoặc game điện tử tự học để thắng đối thủ.

2.2 Phân loại học máy

Các thuật toán học máy được chia thành ba nhóm chính dựa vào cách thức học và dữ liệu mà chúng sử dụng

2.2.1 Học có giám sát (Supervised Learning)

Thuật toán được huấn luyện trên dữ liệu có nhãn (label), nghĩa là dữ liệu đầu vào đi kèm với kết quả đúng. Mục tiêu là học được mô hình để dự đoán nhãn cho dữ liệu mới chưa biết.

Trong loại học này, máy tính được cung cấp dữ liệu đầu vào kèm theo nhãn (label), ví dụ như ảnh được gán nhãn "chó" hoặc "mèo". Thuật toán sẽ học cách phân biệt và dự đoán nhãn chính xác cho ảnh mới.

Ví dụ: Mô hình phân loại email thành "spam" hoặc "không spam". Mô hình học từ tập dữ liệu email đã được dán nhãn trước đó và áp dụng cho email mới.

2.2.2 Học không giám sát (Unsupervised Learning)

Thuật toán làm việc trên dữ liệu không có nhãn, cố gắng tìm ra cấu trúc hoặc mẫu ẩn bên trong dữ liệu.

Loại học này không có dữ liệu nhãn, máy sẽ tự tìm ra các mẫu hoặc nhóm trong dữ liệu. Đây là phương pháp hữu ích khi muốn khám phá cấu trúc dữ liệu mà chưa có thông tin trước.

Ví dụ: Phân nhóm khách hàng dựa trên hành vi mua sắm để đề xuất sản phẩm phù hợp. Máy học phân chia khách hàng thành các nhóm có đặc điểm tương đồng mà không cần biết trước nhóm nào là gì.

2.2.3 Học tăng cường (Reinforcement Learning):

Thuật toán học cách đưa ra quyết định tối ưu qua tương tác với môi trường và nhận phản hồi dưới dạng phần thưởng hoặc hình phạt.

Thuật toán học thông qua việc tương tác với môi trường, nhận phần thưởng hoặc hình phạt để cải thiện hành vi của mình theo thời gian.

Ví dụ: Một con robot tự học cách đi lại bằng cách thử nghiệm các hành động khác nhau và điều chỉnh dựa trên kết quả nhận được.

Bảng ưu điểm và nhược điểm khi áp dụng thuật toán học máy cho bài toán đặt ra

Loại thuật toán	Ưu điểm	Nhược điểm
Học có giám sát	- Dễ dàng đánh giá hiệu quả dựa trên dữ liệu có nhãn- Mô hình thường có độ chính xác cao khi dữ liệu đầy đủ và chất lượng tốt	- Cần dữ liệu đã được gán nhãn đầy đủ và chính xác- Khó áp dụng với dữ liệu phức tạp hoặc không có nhãn
Học không giám sát	- Không cần dữ liệu có nhãn- Phát hiện các mẫu hoặc nhóm dữ liệu mới	- Khó đánh giá độ chính xác- Có thể không tìm được kết quả rõ ràng
Học tăng cường	- Thích hợp với các bài toán cần tương tác liên tục- Máy học tự động cải thiện qua thời gian	- Phức tạp, đòi hỏi nhiều tài nguyên tính toán- Ít phổ biến trong bài toán dự đoán y tế truyền thống

Bảng 2.2 Ưu và nhược điểm khi huấn luyện thuật toán học máy

3. Các mô hình học máy phổ biến

Với tính đa dạng cũng như phong phú của những thuật toán trong từng phân loại của học máy ta có thể lựa chọn những mô hình thuật toán phù hợp với mục tiêu bài toán mà ta muốn hướng đến. Với bài toán dự đoán khả năng béo phì dựa trên các đặc trưng y tế, hành vi và dinh dưỡng dưới đây, học có giám sát là lựa chọn tối ưu nhờ vào khả năng sử dụng dữ liệu đã được gán nhãn

3.1. Hồi quy Logistic (Logistic Regression)

Khái quát: Hồi quy Logistic là một thuật toán học máy thuộc nhóm phân loại dùng để dự đoán xác suất của một biến nhị phân (chỉ có hại trạng thái "có" hoặc "không")

Công thức và ý tưởng: Hồi quy Logistic sử dụng hàm **sigmoid** để chuyển đổi dầu ra tuyến tính thành xác suất, mô hình được huấn luyện bằng cách sử dụng các thuật toán tối ưu như Gradient Descent để cập nhật trọng số zero sao cho hàm mấy mát đạt giá trị nhỏ nhất

Công thức hàm dự đoán

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Trong đó:

$\sigma(z)$*: ký hiệu hàm Sigmoid của z*

Z là hàm tuyến tính

Sau khi đã tính được xác suất, ta gán nhãn $\sigma(X) = P(Y = 1|X)$

Nếu P(Y = 1|X) ≥ 0.5 thì quan sát được phân loại vào lớp dương

Nếu P(Y = 1|X) < 0.5 thì quan sát được phân loại vào lớp âm

Hàm mất mát (Loss function): Để tối ưu mô hình, Logistic Regression sử dụng hàm log-loss (binary cross-entropy)

Mục tiêu tối thiểu hoá hàm mất mát đeer mô hình học được mối quan hệ giữa dữ liệu và nhãn, có thể nói hàm mất mát là chìa khóa cho mô hình Hồi qui Logistic hoạt động hiệu quả

Công thức:

$$L(y', y) = -[ylog(y') + (1 - y)log(1 - y')]$$

Trong đó:

Y là nhãn thực tế của quan sát

Y' là xác suất dự đoán của mô hình được tính qua việc đưa qua hàm sigmoid

- Nếu mô hình dự đoán đúng thì hàm mất mát tiệm cận về 0
- Ngược lại nếu mô hình dự đoán sai hoàn toàn thì hàm mất mát tăng rất lớn

Ưu điểm nhược điểm và khi huấn luyện với tập dữ liệu:

- **Ưu điểm:** Dễ triển khai, dễ hiểu, giải thích tối, phù hợp với bài toán phân loại nhị phân, không yêu cầu tài nguyên tính toán quá lớn

- **Nhược điểm:** Hiệu quả giảm nếu như dữ liệu có quan hệ phi tuyến phức tạp, dễ bị ảnh hưởng bởi dữ liệu mất cân bằng, trong thực tế không phải lúc nào đặc trưng cung độc lập trong thực tế

Hồi quy Logistic là một thuật toán nền tảng, luôn là thuật toán đầu tiên được sử dụng trong yêu cầu phân loại. Nhờ tính dễ hiểu và khả năng dự đoán tốt đối với dữ liệu đơn giản. Song thuật toán này cũng nên được cân nhắc dự đoán nguy cơ béo phì

3.2 Cây quyết định (Decision Tree)

Khái quát: Cây quyết định là một trong những mô hình học máy có cấu trúc đơn giản nhưng cực kì hữu dụng đối với các bài toán phân loại lẫn hồi qui đây là mô hình học có giám sát và được sử dụng nhiều trong học máy bởi tính dễ hiểu và dễ giải thích của chúng .

Tưởng tượng mô hình là một cái cây mỗi nút là một điều kiện kiểm tra, mỗi nhánh là kết quả của điều kiện, và mỗi nút lá là kết quả của dự đoán

Ý tưởng mô hình: Thuật toán xây dựng cây quyết định sẽ tìm cách chia các tập dữ liệu đầu vào thành các nhóm nhỏ hơn dựa trên các đặc trưng sao cho mỗi nhóm con càng đồng nhất càng tốt. Quá trình này được lặp lại cho đến khi đạt được các điều kiện dừng như: dữ liệu trong nhóm đã thuần nhất, độ sâu tối đa của cây đã đạt đến, hoặc số lượng mẫu trong một nst đã nhỏ hơn ngưỡng tối thiểu cho phép

Trong từng bài toán khác nhau thì sẽ có những tiêu chí khác nhau chẳng hạn:

- *Đối với bài toán phân loại:*

Gini Impurity: đo mức độ hỗn tạp trong một nút, càng thấp càng tốt.

Information Gain (Entropy): đo mức giảm thông tin sau khi phân chi

- *Đối với bài toán hồi quy, tiêu chí thường dùng để đánh giá mô hình là:*

Mean Squared Error (MSE) hoặc Mean Absolute Error (MAE) để đo sai số trong dự đoán.

Ưu điểm nhược điểm và khi huấn luyện với tập dữ liệu

- Ưu điểm: Dễ hiểu và trực quan, thích hợp để có thể trình bày và giải thích cho người không chuyên môn, không yêu cầu chuẩn hoá dữ liệu, có thể xử lý dữ liệu hỗn hợp cả định tính và định lượng

- **Nhược điểm:** Dễ dẫn đến dữ liệu bị quá khớp (overfitting) nếu cây đạt độ sâu quá sâu cũng như có nhiều nhánh, nhạy cảm khi có dữ liệu nhỏ bị thay đổi

3.3 Rừng ngẫu nhiên (Random Forest)

Khái quát: Trong khi cây quyết định dễ bị quá khớp với dữ liệu thì mô hình rừng ngẫu nhiên lại có thể thoát được điều ấy, với cây quyết định thì chỉ xây dựng dựa trên một

cái cây nhưng đối với rừng ngẫu nhiên thì sẽ là một rừng cây. Đây cũng là một thuật toán học có giám sát được sử dụng cho cả bài toán hồi quy và phân loại

Ý tưởng mô hình: Thuật toán xây dựng một rừng gồm nhiều cây quyết định như ở trên và sau đó kết hợp các kết quả từ những cây này nhằm đưa ra dự đoán

Mô hình sẽ chia tập dữ liệu gốc gồm nhiều mẫu thành k tập dữ liệu con với số lượng cây quyết định tương đương, mỗi tập con khi được chia sẻ được theo phương pháp chọn mẫu ngẫu nhiên và có hoàn lại từ tập dữ liệu ban đầu, một mẫu có thể tồn tại hoặc không trong từng tập dữ liệu con (tạo ra tính đa dạng của việc chọn mẫu tránh mô hình quá khớp)

Việc dự đoán sẽ được thực hiện với từng cây quyết định mỗi cây sẽ có một tập dữ liệu con khác nhau được tách từ tập dữ liệu gốc

- Với bài toán phân loại mô hình sẽ thực hiện lấy đa số phiếu bầu để đưa ra lớp cuối cùng
- Với bài toán hồi quy mô hình sẽ trả về trung bình cộng của tất cả các dự đoán

Ưu điểm nhược điểm của từng ngẫu nhiên

- **Ưu điểm:** Giảm thiểu dữ liệu quá khớp so với một cây đơn lẻ, hiệu quả trên cả dữ liệu lớn và dữ liệu có nhiễu, tự động đánh giá tầm quan trọng của các đặc trưng, không yêu cầu chuẩn hoá dữ liệu

- **Nhược điểm:** Khó giải thích, Tốn nhiều thời gian và tài nguyên tính toán do số lượng cây lớn, hoạt động chậm khi yêu cầu dự đoán trên thời gian thực

3.4 K- Láng giềng gần nhất (K-Nearest Neighbors – KNN)

Khái quát: K-Nearest Neighbors (KNN) là một trong những thuật toán học máy đơn giản nhưng hiệu quả, thường được sử dụng trong các bài toán phân loại và hồi quy. KNN hoạt động dựa trên nguyên tắc "gần giống thì gần nhau", nghĩa là một điểm dữ liệu mới sẽ được phân loại dựa trên k điểm dữ liệu gần nó nhất trong tập huấn luyện.

Ý tưởng mô hình: Thuật toán này hoạt động dựa trên giả định khả năng những điểm dữ liệu có đặc điểm giống nhau thường nằm gần nhau trong không gian đặc trưng. Do đó, để dự đoán nhãn của một điểm dữ liệu mới, K-Nearest Neighbors sẽ đưa sẽ dựa vào các điểm "láng giềng gần nhất" đã biết nhãn trong tập huấn luyện.

Ý tưởng mô hình thực hiện các bước chọn làng giềng k, tính khoảng cách giữa các điểm cần phân loại và tất cả các điểm trong tập dữ liệu với nhau thường là theo công thức tính khoảng cách Euclid.

Điểm dữ liệu cần dự đoán: $X = (x_1, x_2, x_3, \ldots, x_n)$

Các điểm dữ liệu huấn luyện: $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, mỗi điểm $A^{(i)} = (a_1^{(i)}, a_2^{(i)}, \ldots, a_n^{(i)})$

$$d(X, A^{(i)}) = \sqrt{\sum_{j=1}^{n} (x_j - a_j{}^{(i)})^2}$$

Ưu điểm và nhược điểm khi huấn luyện mô hình

- **Ưu điểm:** Để hiểu và dễ triển khai, không cần huấn luyện mô hình rõ ràng – mô hình chỉ "học" khi dự đoán, hoạt động tốt với dữ liệu nhỏ ít nhiễu

- **Nhược điểm:** Chậm khi áp dụng với tập dữ liệu lớn (vì phải tính khoảng cách với toàn bộ tập huấn luyện), nhạy cảm với dữ liệu nhiễu và các đặc trưng không liên quan, hiệu quả giảm khi có quá nhiều đặc trưng (curse of dimensionality).

3.5 Support Vector Machine (SVM)

Khái quát: SVM là một thuật toán học máy mạnh mẽ thường được sử dụng trong các bài toán phân loại và hồi quy. SVM hoạt động bằng cách tìm ra một siêu phẳng tối ưu để phân chia dữ liệu thành các lớp khác nhau sao cho khoảng cách từ siêu phẳng đó đến các điểm dữ liệu gần nhất của mỗi lớp là lớn nhất có thể - đây gọi là lề (margin)

Ý tưởng mô hình:

Siêu phẳng phân chia:

Với một bài toán phân loại nhị phân, SVM cố gắng tìm một siêu phẳng trong không gian đặc trưng sao cho dữ liệu được chia thành hai lớp rõ ràng. Siêu phẳng có dạng:

$$\omega^T x + b = 0$$

Trong đó:

ω: là vector trọng số, xác định hướng của siêu phẳng

x: là đầu vào, vecto dữ liệu

b: là hệ số điều chỉnh.

$\omega^T x$: là tích vô hướng giữa vectơ trọng số và dữ liệu đầu vào

Hỗ trợ điểm và lề:

Các support vectors là những điểm dữ liệu gần siêu phẳng nhất – chúng đóng vai trò quyết định trong việc xác định vị trí và hướng của siêu phẳng. Lề (margin) là khoảng cách từ siêu phẳng đến các điểm này, và mục tiêu của SVM là tối đa hóa lề để đạt được sự phân chia tốt nhất.

Dữ liệu không tuyến tính – Kernel Trick:

Trong trường hợp dữ liệu không thể phân chia tuyến tính, SVM sử dụng một kỹ thuật gọi là kernel trick, giúp ánh xạ dữ liệu sang không gian chiều cao hơn nơi dữ liệu có thể được phân tách tuyến tính. Các hàm kernel phổ biến:

- **Linear kernel:** $K(x, x`) = x^T x`$
- **Polynomial kernel:** $K(x, x`) = (x^T x` + c)^d$
- **Radial Basis Function (RBF) kernel:** $K(x, x`) = exp(-\gamma \|x - x`\|^2)$

Dự đoán:

Sau khi huấn luyện, SVM sẽ sử dụng siêu phẳng tìm được để dự đoán nhãn của dữ liệu mới dựa trên vị trí của điểm đó so với siêu phẳng.

Ưu điểm và nhược điểm của mô hình

- Ưu điểm: Hoạt động tốt với không gian đặc trưng cao (nhiều chiều), phân loại mạnh với biên lớn, hiệu quả với dữ liệu có biên phân tách rõ ràng, có thể xử lý quan hệ phi tuyến qua kernel trick.

- **Nhược điểm:** Không hiệu quả với dữ liệu lớn (thời gian huấn luyện chậm), khó chọn kernel và các siêu tham số, kém hiệu quả với dữ liệu có nhiều nhiễu hoặc lớp chồng lấn.

Tóm lại ta có bảng so sánh sau

Thuật toán	Ưu điểm	Nhược điểm
Logistic Regression	- Đơn giản, dễ triển khai và giải thích- Hiệu quả với phân loại nhị phân- Tính toán nhanh	- Không phù hợp với dữ liệu phi tuyến- Dễ bị ảnh hưởng bởi dữ liệu mất cân bằng và đa cộng tuyến
Decision Tree	- Dễ hiểu và trực quan- Không cần chuẩn hóa dữ liệu- Xử lý được cả dữ liệu định lượng và định tính	- Dễ bị quá khớp nếu không cắt tỉa- Nhạy cảm với dữ liệu nhiễu và thay đổi nhỏ
Random Forest	- Giảm overfitting so với cây đơn- Độ chính xác cao- Hỗ trợ đánh giá độ quan trọng của đặc trưng	- Tốn tài nguyên tính toán- Khó giải thích mô hình- Thời gian huấn luyện lâu hơn cây đơn
K-Nearest Neighbors (KNN)	- Không cần huấn luyện mô hình- Dễ hiểu, dễ cài đặt- Phù hợp với dữ liệu nhỏ và biên phân lớp rõ ràng	- Tốc độ chậm với dữ liệu lớn- Dễ bị nhiễu- Nhạy cảm với tỉ lệ đặc trưng và khoảng cách
Support Vector Machine (SVM)	- Hiệu quả với không gian nhiều chiều- Hoạt động tốt với biển phân cách rõ ràng- Có thể áp dụng kernel để xử lý quan hệ phi tuyến	- Chậm với dữ liệu lớn- Cần chọn kernel và siêu tham số phù hợp- Không phù hợp với dữ liệu nhiều nhiễu hoặc phân lớp chồng lấn

Bảng 2.3 So sánh ưu và nhược điểm giữa các mô hình học máy

4. Tổng quan nghiên cứu trước

Trong những năm gần đây, việc áp dụng học máy vào lĩnh vực y tế, đặc biệt là chẩn đoán và dự đoán các bệnh không lây nhiễm như béo phì, đang trở thành một

hướng nghiên cứu nổi bật. Nhiều nghiên cứu đã chứng minh hiệu quả của các mô hình học máy trong việc phân tích dữ liệu y tế và hỗ trợ ra quyết định lâm sàng.

Một trong những nghiên cứu điển hình là của **Pereira et al. (2020)**, trong đó tác giả sử dụng tập dữ liệu gồm các yếu tố như chỉ số BMI, thói quen ăn uống, mức độ hoạt động thể chất và tiền sử bệnh để dự đoán khả năng béo phì bằng các thuật toán như Decision Tree, Random Forest, K-Nearest Neighbors và Logistic Regression. Kết quả cho thấy mô hình Random Forest đạt độ chính xác cao nhất (~95%) trong việc phân loại người có nguy cơ béo phì.

Nghiên cứu của **Kavakiotis et al. (2017)** tổng hợp lại nhiều ứng dụng học máy trong y học và nhấn mạnh rằng các mô hình như SVM và Neural Networks rất phù hợp để phân tích các bệnh mãn tính, bao gồm béo phì. Tuy nhiên, các mô hình này đòi hỏi dữ liệu lớn và chất lượng cao để đạt hiệu quả tối ưu.

Ngoài ra, **Rahman et al. (2021)** sử dụng Logistic Regression và Support Vector Machines để dự đoán béo phì trên một bộ dữ liệu đa quốc gia. Kết quả nghiên cứu nhấn mạnh vai trò quan trọng của các yếu tố như tuổi, giới tính, thời lượng ngủ, thói quen ăn uống và hoạt động thể chất.

Tại Việt Nam, các nghiên cứu về học máy trong lĩnh vực sức khỏe vẫn còn khá hạn chế, nhưng đã có một số bước tiến nhất định. Nhiều nghiên cứu bắt đầu áp dụng mô hình máy học để phân tích dữ liệu từ các cuộc khảo sát y tế, bệnh án điện tử nhằm phát hiện sớm nguy cơ béo phì ở người trẻ tuổi.

Từ những nền tảng trên, đề tài tiểu luận này hướng đến việc kế thừa và mở rộng các phương pháp đã được nghiên cứu, đồng thời kiểm tra hiệu quả của các mô hình học máy trong việc dự đoán béo phì trên một tập dữ liệu cụ thể. Mục tiêu là lựa chọn được mô hình phù hợp, đồng thời phân tích các yếu tố đóng vai trò quan trọng nhất trong việc ảnh hưởng đến nguy cơ béo phì.

Chương III

Phương pháp nghiên cứu

1. Mô tả và nguồn dữ liệu

Bộ dữ liệu được sử dụng trong nghiên cứu có tên **"***Obesity DataSet Raw and Data Sinthetic***"**, được công khai trên nền tảng *Kaggle* – một trong những trang web chia sẻ dữ liệu và tổ chức thi đấu học máy uy tín. Bộ dữ liệu này bao gồm **2.111 mẫu dữ liệu** với **17 đặc trưng đầu vào**, phản ánh các yếu tố ảnh hưởng đến nguy cơ béo phì hoặc thiếu cân của một cá nhân. Mục tiêu của bộ dữ liệu là hỗ trợ các nghiên cứu trong việc phân loại mức độ béo phì dựa trên thói quen ăn uống, lối sống và tiền sử gia đình.

Theo mô tả từ *Kaggle, **77% dữ liệu*** trong bộ này được tạo ra bằng công cụ ***Weka*** và kỹ thuật ***SMOTE*** (Synthetic Minority Over-sampling Technique) để cân bằng phân bố lớp, trong khi ***23% còn lại*** là dữ liệu thực được thu thập từ các quốc gia

có tỷ lệ béo phì cao như **Mexico**, **Peru** và **Colombia**, từ các cá nhân ở độ tuổi **14 đến 61**.

Các đặc trưng cụ thể trong bộ dữ liệu bao gồm:

Đặc trưng	Giải thích
gender	Giới tính
age	Tuổi
height	Chiều cao (mét)
weight	Cân nặng (kg)
family_history_with_overweight	Có người thân trong gia đình từng bị thừa cân
freq_high_calories_food	Có thường xuyên tiêu thụ thực phẩm giàu calo
freq_of_vegetable	Tần suất ăn rau trong bữa ăn
number_of_main_meal	Số bữa chính trong một ngày
food_between_meal	Có ăn vặt giữa các bữa ăn hay không
smoking	Có hút thuốc hay không
daily_water	Lượng nước uống trung bình mỗi ngày
monitors_calories_daily	Có kiểm soát lượng calo tiêu thụ hàng ngày
freq_of_physical_activity	Tần suất vận động thể chất
time_using_technology	Thời gian sử dụng thiết bị công nghệ mỗi ngày
freq_of_alcohol	Tần suất sử dụng đồ uống có cồn
transportation	Phương tiện đi lại chính
obesity_level	Nhãn mục tiêu – mức độ béo phì

Bảng 3.1 Đặc trưng và ý nghĩa của chúng trong tập dữ liệu

Biến mục tiêu obesity level gồm **7 mức phân loại**:

Tên lớp trong tập dữ liệu	Giải thích
insufficient_weight	Underweight (Thiếu cân)
normal	Normal weight (Cân nặng bình thường)
overweight_I	Overweight Level I (Thừa cân cấp độ I)

verweight_II	Overweight Level II (Thừa cân cấp độ II)
obesity_I	Obesity Type I (Béo phì cấp độ I)
obesity_II	Obesity Type II (Béo phì cấp độ II)
obesity_III	Obesity Type III (Béo phì cấp độ III)

Bảng 3.2 Các lớp trong biên mục tiêu

Các đặc trưng trong tập dữ liệu chủ yếu ở dạng định tính hoặc định lượng đơn giản, dễ dàng xử lý bằng các thuật toán học máy. Trước khi huấn luyện mô hình, dữ liệu đã được **chuẩn hóa** và **mã hóa lại** bằng các kỹ thuật tiền xử lý như **One-Hot Encoding** và **Min-Max Scaling**, giúp tăng hiệu quả và độ chính xác của mô hình.

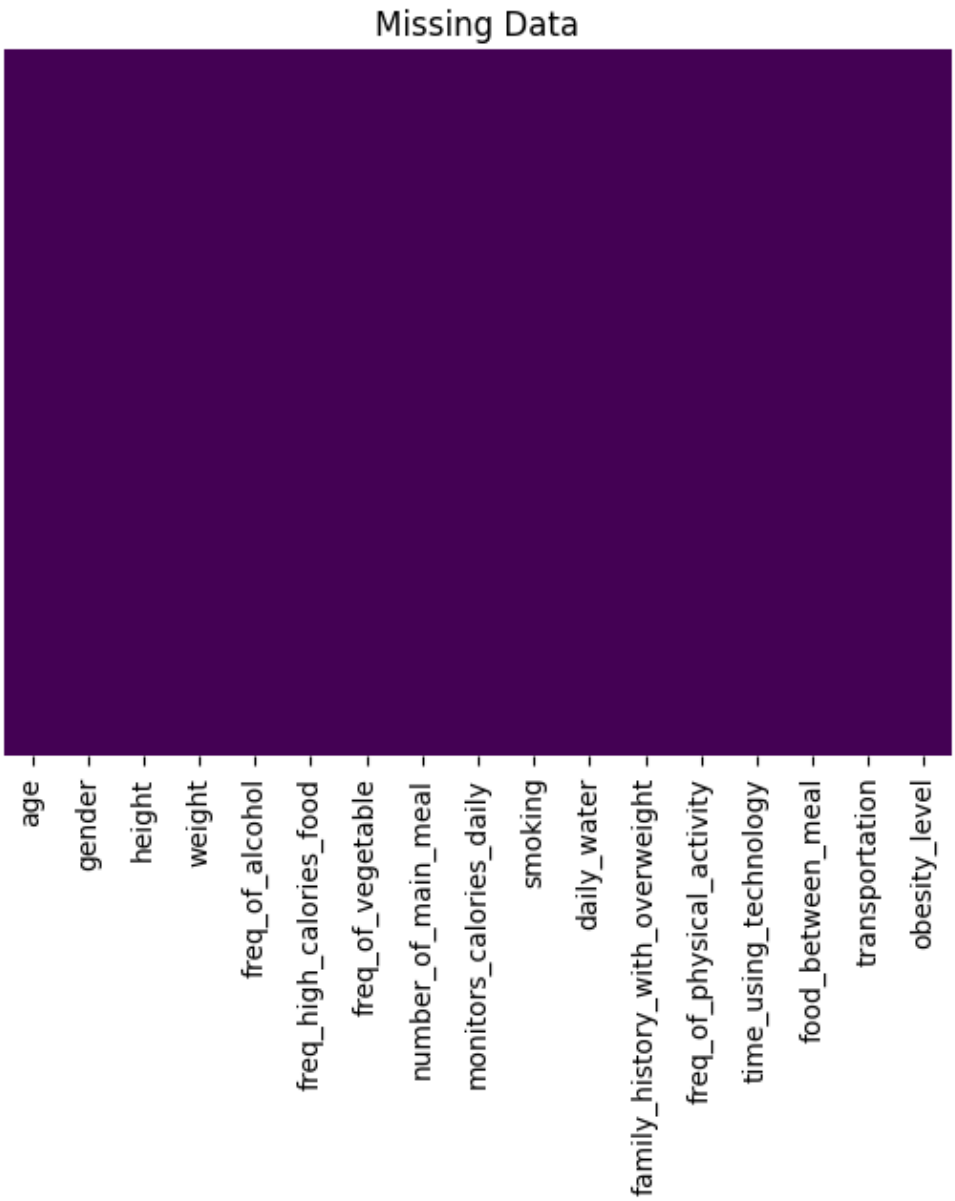
Nhìn chung, đây là một bộ dữ liệu có độ tin cậy cao, được nhiều nghiên cứu sử dụng, mang tính ứng dụng thực tiễn và có khả năng mở rộng trong các bối cảnh khác nhau.

2. Tiền xử lý dữ liệu

Trước khi có thể đưa dữ liệu vào mô hình học máy nhằm huấn luyện hay đưa ra dự đoán, quá trình tiền xử lý đóng via trò vô cùng quan trọng nhằm đảm bảo chất lượng và độ chính xác của các kết quả của dự đoán. Trong bài luận này các bước xử lý dữ liệu sẽ được thực hiện như sau

2.1 Kiểm tra dữ liệu bị thiếu (null values) và dữ liệu bị trùng

Bộ dữ liệu "*Obesity DataSet Raw and Data Sinthetic*" không chứa giá trị bị thiếu (null), vì vậy không cần thực hiện loại bỏ hoặc thay thế dữ liệu thiếu. Điều này giúp quá trình xử lý nhanh chóng và đảm bảo độ đầy đủ của mẫu.



Hình 1: Biểu đồ biểu thị giá trị bị thiếu theo các đặc trưng

Thêm vào đó trong tệp dữ liệu này các cột có kiểu dữ liệu số như:

Phân loại	**Kiểu dữ liệu**	**Tên đặc trưng**
Kiểu dữ liệu số (Numerical)	Số nguyên (Integer)	age
	Số thực (Float)	height,weight,freq_of_vegetable,number_of_main_meal, daily_water, freq_of_physical_activity, time_using_technology
	Danh tính (Nominal)	gender, smoking,family_history_with_overweight, transportation

Kiểu dữ liệu phân loại (Categorical)	Thứ tự (Ordinal)	freq_of_alcohol, freq_high_calories_food, monitors_calories_daily, food_between_meal, obesity_level

Bảng 3.2.1 Kiểu dữ liệu của từng đặc trưng

Bên cạnh đó tập dữ liệu cũng có 24 mẫu dữ liệu bị trùng và sau khi thực hiện việc loại bỏ tập dữ liệu chỉ còn 2087 mẫu

Đối với từng đặc trong phân loại trong tập dữ liệu gồm có các đặc trưng như sau:

Đặc trưng	**Thành phần**
gender	Female (*Nữ*), Male (*Nam*)
freq_of_alcohol	No (*Không*) , Sometimes (*Thông thường*) , Frequently (*Có tần suất*), Always (*Luôn luôn*)
freq_high_calories_food	No (*Không*) , Yes (*Có*)
monitors_calories_daily	No (*Không*) , Yes (*Có*)
smoking	No (*Không*) , Yes (*Có*)
family_history_with_overweight	No (*Không*) , Yes (*Có*)
food_between_meal	No (*Không*) , Sometimes (*Thông thường*) , Frequently (*Có tần suất*), Always (*Luôn luôn*)
transportation	Public_transportation (*Phương tiện công cộng*), Walking (*Đi bộ*), Automobile (*Phương tiện di động*), Motorbike (*Xe máy*), Bike (*Xe đạp*)

Bảng 3.2.2 Các lớp với từng đặc trưng phân loại

2.2 Xử lý chuẩn hoá và phân chia dữ liệu

Mã hoá biến phân loại (Categorical Encoding)

Một số đặc trưng trong tập dữ liệu mang giá trị định tính như (giới tính, tần suất sử dụng rượu, tiền xử gia đình có mắc các bệnh béo phì hay không, hút thuốc,…) cần phải được mã hoá dưới dạng số nhằm giúp cho mô hình có thể hiểu và thực hiện học. Phương pháp Label Encoding được sử dụng để mã hoá các biến phân loại này, giúp tránh việc mô hình hiểu nhầm mối quan hệ thứ bậc giữa các giá trị danh mục cũng như có thể sử dụng các biểu đồ nhằm đánh giá mối quan hệ với biến mục tiêu

Chuẩn hóa dữ liệu (Normalization/Standardization)

Để đảm bảo tất cả các đặc trưng đầu vào có cùng đơn vị đo và ảnh hưởng công bằng đến mô hình, các đặc trưng mang tính định lượng như chiều cao, cân nặng, tuổi tác, tần suất vận động, … được chuẩn hóa về khoảng [-1,1]. Việc này giúp gia tăng độ

chính xác và hiệu quả khi huấn luyện mô hình, đặc biệt với các thuật toán nhạy cảm như KNN hay SVM

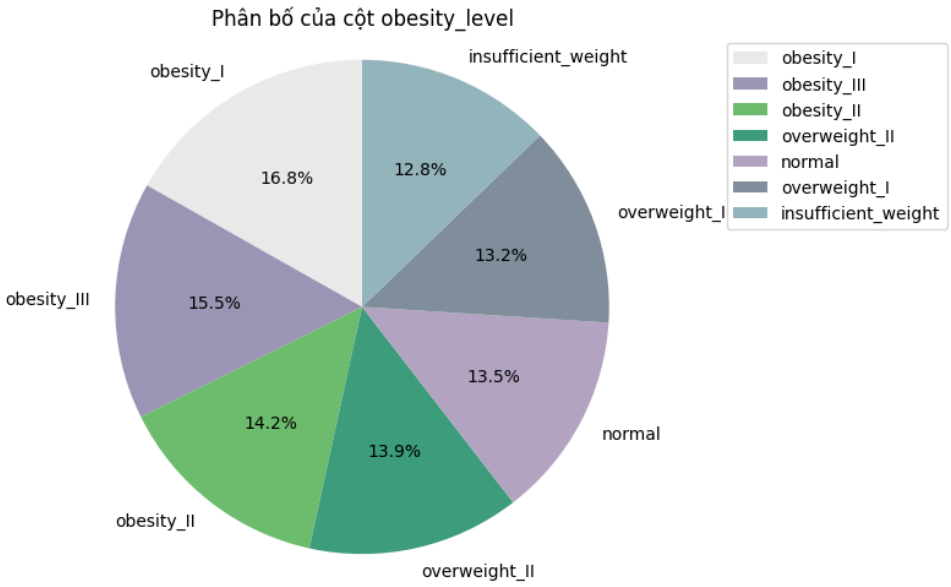
Xử lý dữ liệu mất cân bằng (Imbalanced Data)

Việc xử lý mất cân bằng dữ liệu cũng có thể đóng vai trò cực kì quan trọng trong việc huấn luyện mô hình học máy. Bởi lẽ việc cân bằng có thể thiên lệch trong dự đoán ưu tiên các lớp phổ biến hơn và bỏ qua các lớp ít xuất hiện

Nếu xảy ra mất cân bằng dữ liệu có thể ảnh hưởng tới:

- Giảm độ chính xác của mô hình với các lớp hiếm
- F1-score hoặc recall thấp với các nhóm bị mất cân bằng
- Làm sai lệch đánh giá hiệu năng của mô hình nếu chỉ dựa vào độ chính xác tổng thể

Tuy nhiên tệp dữ liệu đã khá cân bằng nên việc áp dụng các phương pháp cân bằng dữ liệu không được sử dụng trong bài tiểu luận này



Hình 2: Biểu đồ tròn thể hiện phân tán của các lớp trong biến mục tiêu

Phân chia dữ liệu huấn luyện và kiểm tra

Sau khi đã hoàn thành các bước tiền xử lý dữ liệu, bước tiếp theo trong quá trình xây dựng mô hình học máy chính là phân chia dữ liệu thành hai tập chính: tập huấn luyện và tập kiểm tra

Mục tiêu của việc phân chia

- Tập huấn luyện (training set) được sử dụng để mô hình học các đặc trưng từ dữ liệu và xây dựng mối quan hệ giữa các biến. Đồng thời cũng sẽ là tập dữ liệu được sử dụng để có thể xây dựng các biểu đồ nhằm đánh giá quan sát mối quan hệ giữa các đặc trưng và biến mục tiêu

- Tập kiểm tra (test set) dùng để đánh giá khả năng tổng quát hoá của mô hình, tức là kiểm tra xem mô hình có thể dự đoán tốt trên dữ liệu mới chưa thấy hay không

- Dữ liệu được chia theo tỷ lệ phổ biến là 80% huấn luyện và 20% kiểm tra. Tỷ lệ này đảm bảo rằng mô hình có đủ lực để học, đồng thời vẫn giữ lại được một phần lớn để đánh giá hiệu quả mô hình

- Việc phân chia dữ liệu một cách hợp lý sẽ tránh được hiện tượng quá khớp (overfitting), đồng thời cho phép đánh giá công bằng và khách quan năng lực của các mô hình học máy khi triển khai trên dữ liệu thực tế

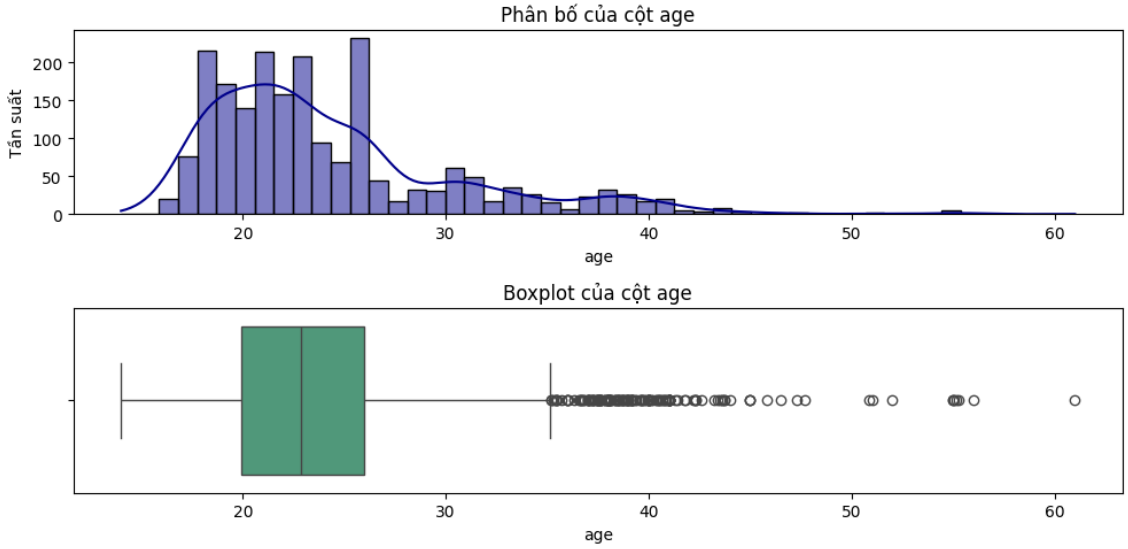
3. Khai phá dữ liệu

Sau bước tiền xử lý, quá trình **khai phá dữ liệu (Exploratory Data Analysis - EDA)** được thực hiện nhằm hiểu rõ hơn về cấu trúc, phân bố và mối quan hệ giữa các đặc trưng trong dữ liệu. Mục tiêu của bước này là khám phá thông tin tiềm ẩn và hỗ trợ lựa chọn mô hình phù hợp cho bài toán phân loại béo phì.

3.1 Đánh giá với các đặc trưng mang kiểu dữ liệu số

Tuổi (Age): Phân bố tuổi chủ yếu dao động từ 14 đến 61 tuổi. Trung bình độ tuổi nằm quanh mức 25, với phần lớn mẫu dữ liệu tập trung vào độ tuổi trưởng thành trong khoảng 19 đến 27 tuổi, xu hướng giảm dần về phía bên phải

Biểu đồ Boxplot: cũng cho thấy đặc trưng này có khá nhiều giá trị ngoại lai (outlier) với giá trị độ tuổi thấp nhất là dưới 20, dữ liệu ngoại lai xuất hiện trong khoảng trên 30 đến hơn 60 tuổi, giá trị trung bình trong khoảng độ tuổi này là giữa 20 và 30 tuổi

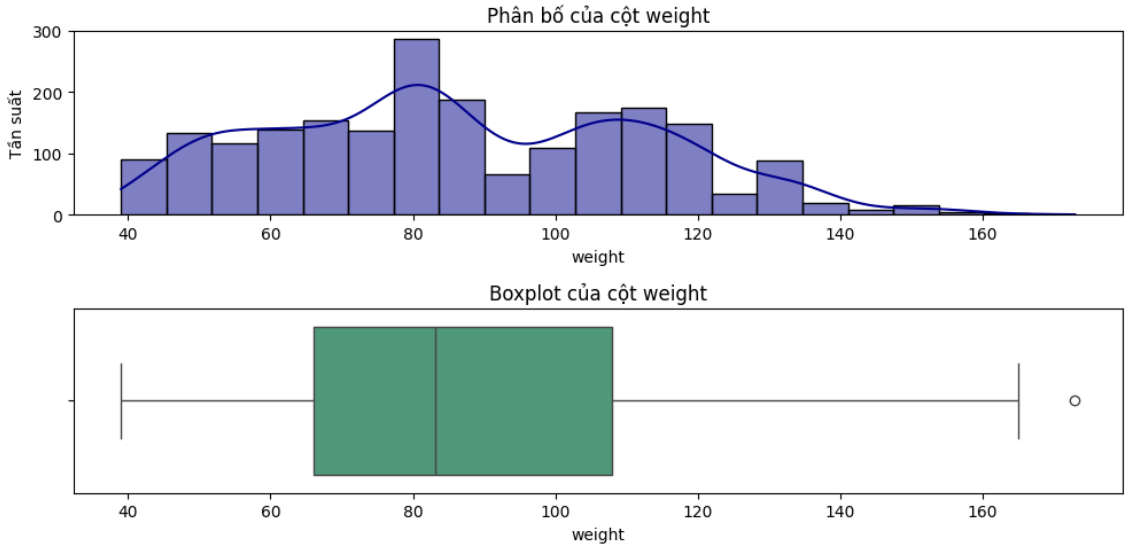


Hình 3.1: Biểu đồ thể hiện sự phân bố của đặc trưng "Age"

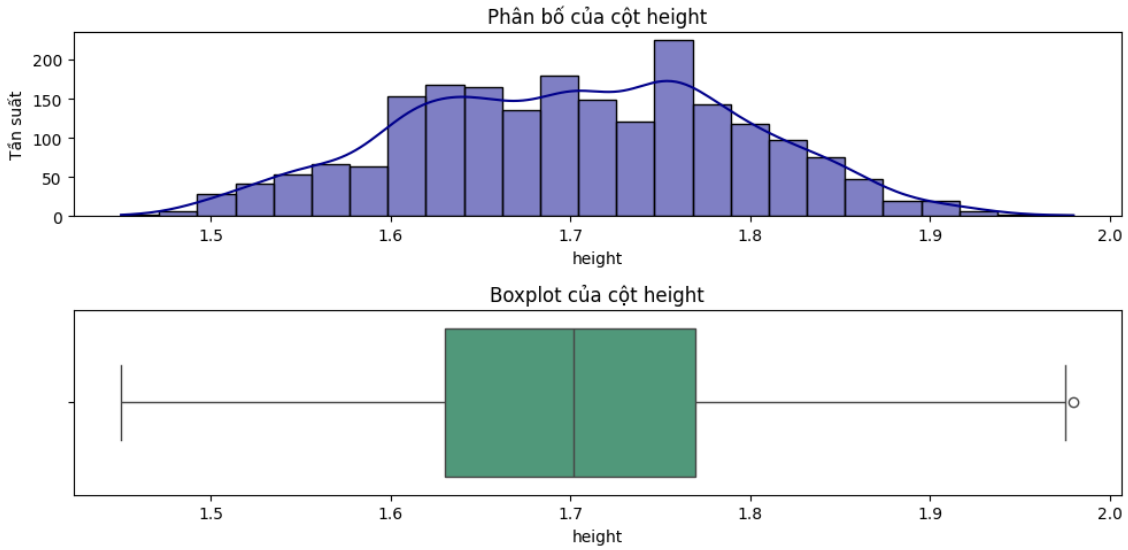
Cân nặng (Weight) và Chiều cao (Height): Cân nặng phân bố không đều, cho thấy có sự hiện diện của các cá nhân béo phì và thiếu cân. Chiều cao có xu hướng gần chuẩn với độ lệch nhỏ.

Biểu đồ Boxplot: Đối với đặc trưng cân nặng đây là một đặc trưng tương đối sạch có giá trị thấp nhất là khoảng 40kg và cao nhất là trên 160kg giá trị trung bình là khoảng 80-100kg lượng phân bố đều của đặc trưng cân nặng này là khoảng trên 60 đến hơn 100kg. Bên cạnh cân nặng thì chiều cao cũng là một đặc trưng "sạch" không kém

giá trị cao nhất dưới 2m và thấp nhất là dưới 1,5m giá trị trung bình là 1,7m phân bố đều nằm trong khoảng trên 1,6m và dưới 1,8m



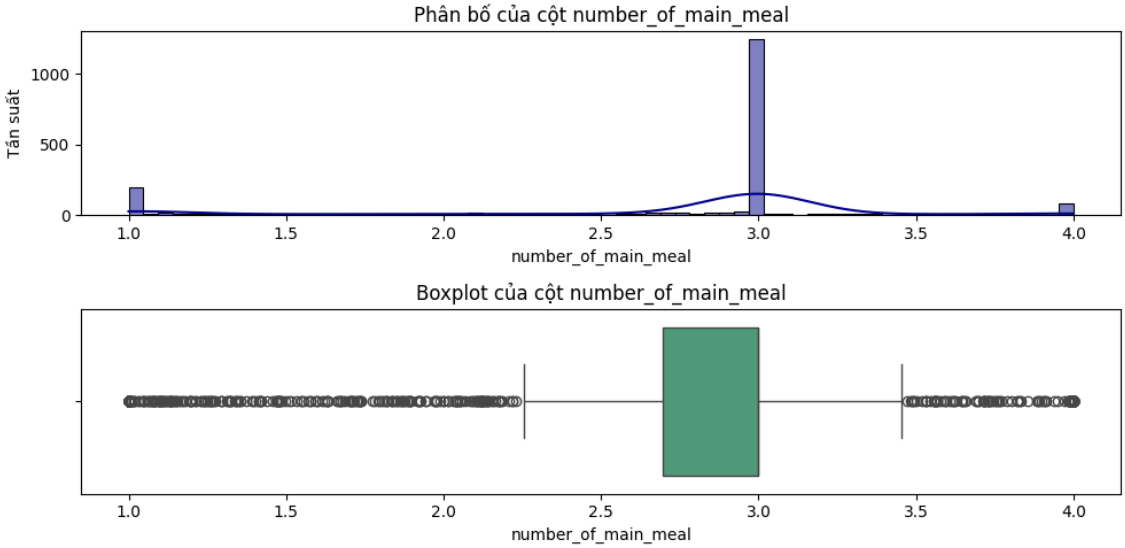
Hình 3.2: Biểu đồ thể hiện sự phân bố của đặc trưng "Weight"



Hình 3.3: Biểu đồ thể hiện sự phân bố của đặc trưng "Height"

Số lượng bữa chính (Number of main meal): Có thể thấy số lượng bữa chính phân bố nhiều nhất vẫn chính là 3 bữa một ngày có cả 4 bữa một ngày và cả 1 bữa trong một ngày. Tuy nhiên trung bình nhất vẫn là trong khoảng từ 2.5 đến 3 bữa.

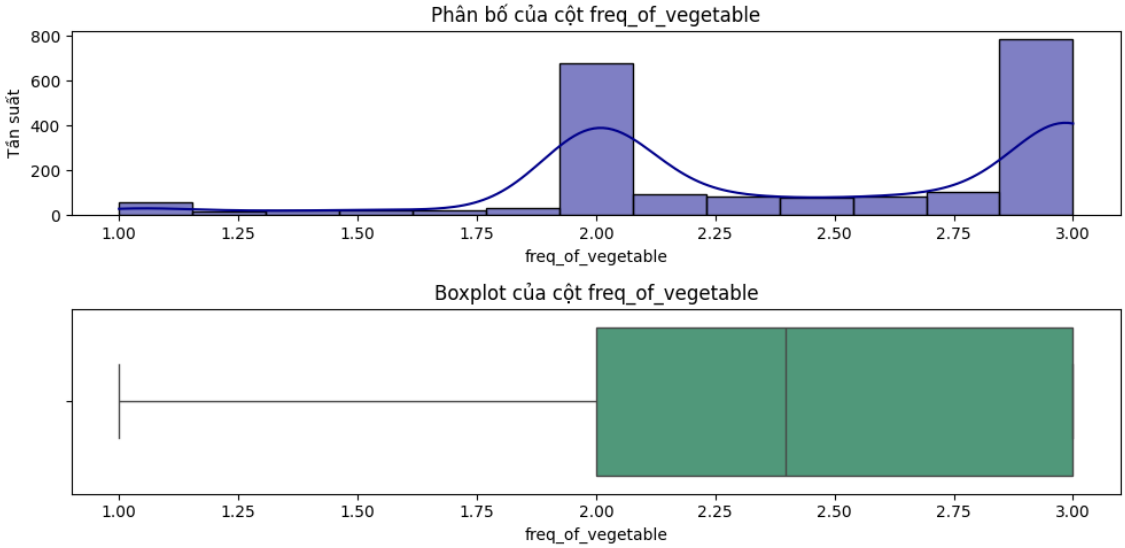
Biểu đồ Boxplot: Đặc trưng này là nhiều giá trị ngoại lai nhất cụ thể thường nằm trong khoảng dưới 1.0 đến 2.0 và 3.5 đến 4.0



Hình 3.4: Biểu đồ thể hiện sự phân bố của đặc trưng "Number of main meal"

Tần suất ăn rau (Freq of vegetables): Tần suất sử dụng rau củ đa số năm ở mức 2 bữa và nhiều nhất là ở mức 3, có sự phân bố không đồng đều. Đáng chú í ở mức một tần suất tiêu thụ rau vẫn khá cao vào chỉ trải dải từ mức 2 đến 3

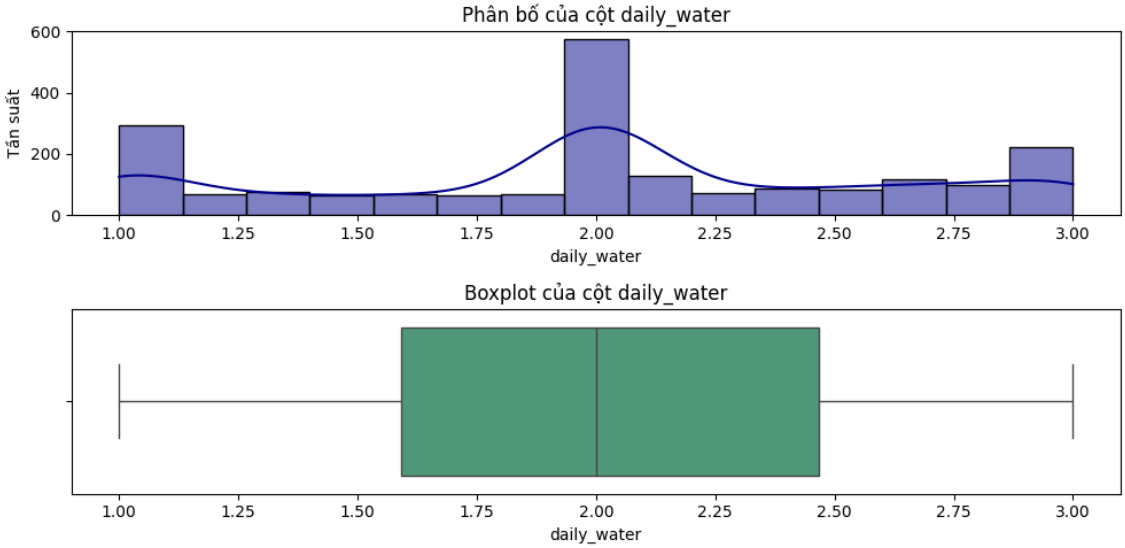
Biểu đồ Boxplot: Đặc trưng này không chứa dữ liệu ngoại lại phân bố đều trong khoảng từ 2.0 đến 3.0 với giá trị trung bình nằm giữa khoảng 2.2 đến 2.5 giá trị lớn nhất là 3 và thấp nhất là 1



Hình 3.5: Biểu đồ thể hiện sự phân bố của đặc trưng "Freq of vegetables"

Tần suất lượng nước được nạp vào cơ thể (Daily water): Lượng nước được nạp vào cơ thể phân bố nhiều nhất ở lượng 2 lít mỗi ngày thấp dần 1 lít và thật bất ngờ lượng nước 3 lit một ngày cũng xuất hiện

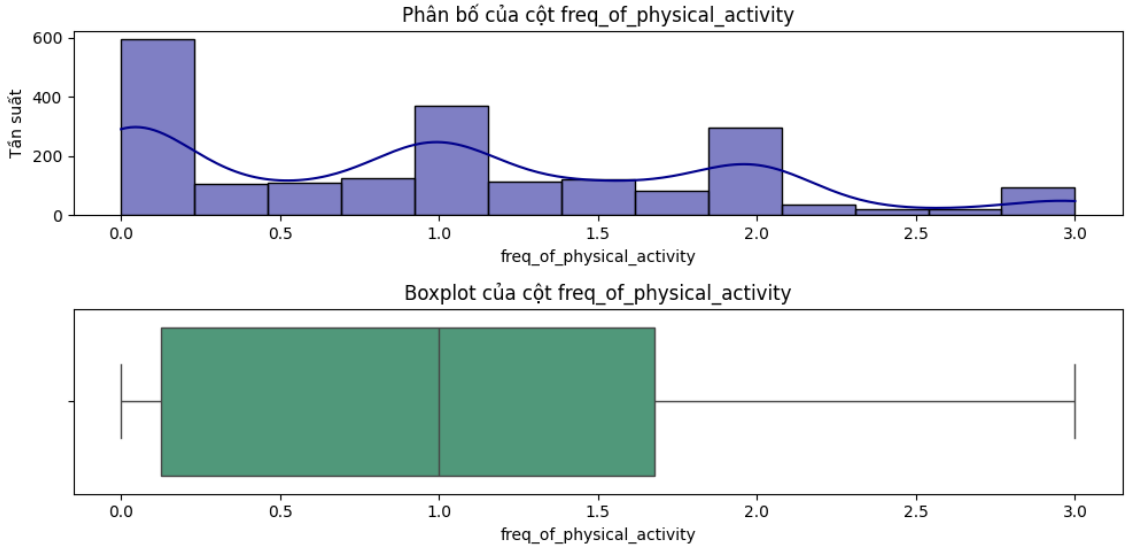
Biểu đồ Boxplot: Đặc trưng "sạch" với giá trị phân bố đều được nằm trong khoảng trên 1.5 đến 2.5l một ngày giá trị trung bình 2l giá trị cao nhất là 3l và giá trị thập nhất là 1l



Hình 3.6: Biểu đồ thể hiện sự phân bố của đặc trưng "Daily water"

Tần suất luyện tập thể dục (Freq of physical activity): Phân bố nhiều nhất chính là thường không luyện tập thể dục điều này cũng phản ánh được rằng đa số các mẫu trong tập dữ liệu các mẫu đa số đều bị mắc các bệnh béo phì hay thừa cân

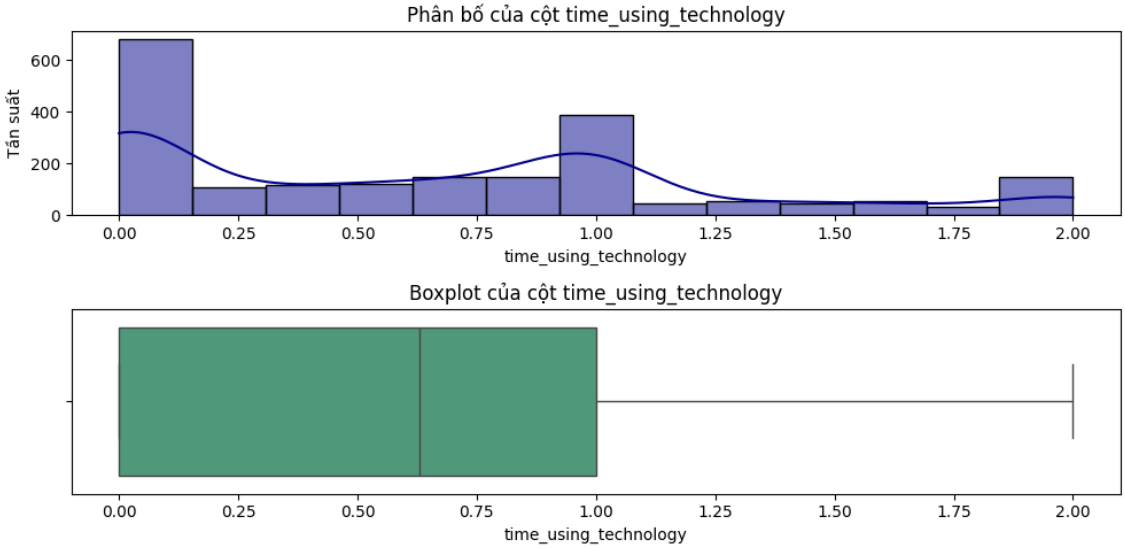
Biểu đồ Boxplot: Đặc trưng phân bố đều nằm trong khoảng trên 0 đến hơn 1,5 với giá trị trung bình là 1 ta có giá trị lớn nhất là 3.0 và nhỏ nhất là 0.0



Hình 3.7: Biểu đồ thể hiện sự phân bố của đặc trưng "Freq of physical activity"

Thời gian sử dụng công nghệ (Time using technology) Phân bố cao nhất là không điều này khá nghịch lý đối với đa số các mẫu trong tập dữ liệu bởi lẽ ngày nay việc mắc các bệnh béo phì càng lớn thì tần suất sử dụng công nghệ càng cao nhưng trong tập dữ liệu thời gian sử dụng công nghệ cao nhất ở mức không

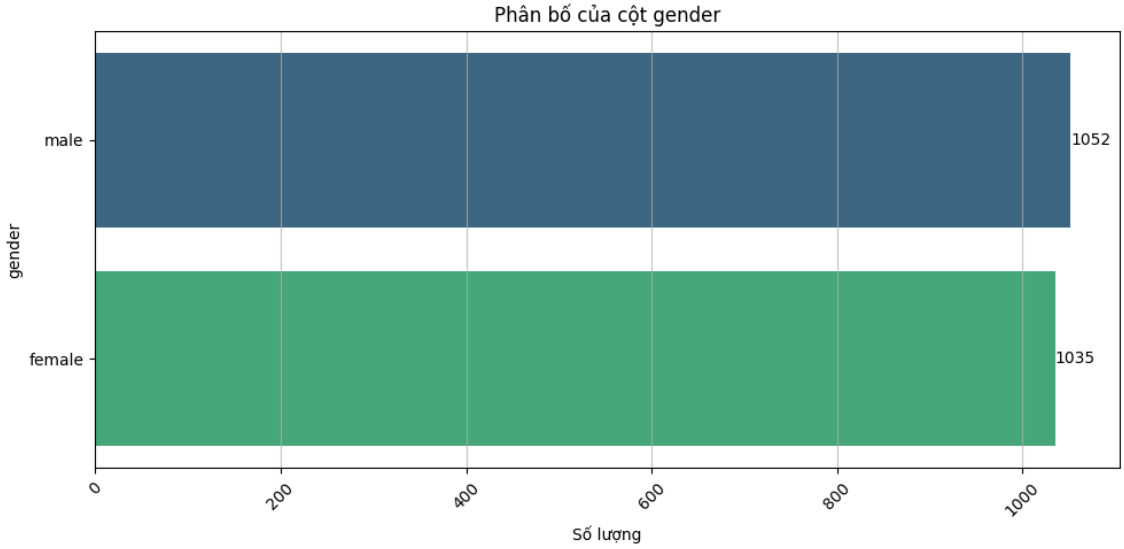
Biểu đồ Boxplot: Đây là một đặc trưng "sạch nhất" với độ phân bố nằm trong khoảng từ 0 đến 1l nước giá trị thấp nhất là 0 và nhiều nhất là 3



Hình 3.8: Biểu đồ thể hiện sự phân bố của đặc trưng "Time using technology"

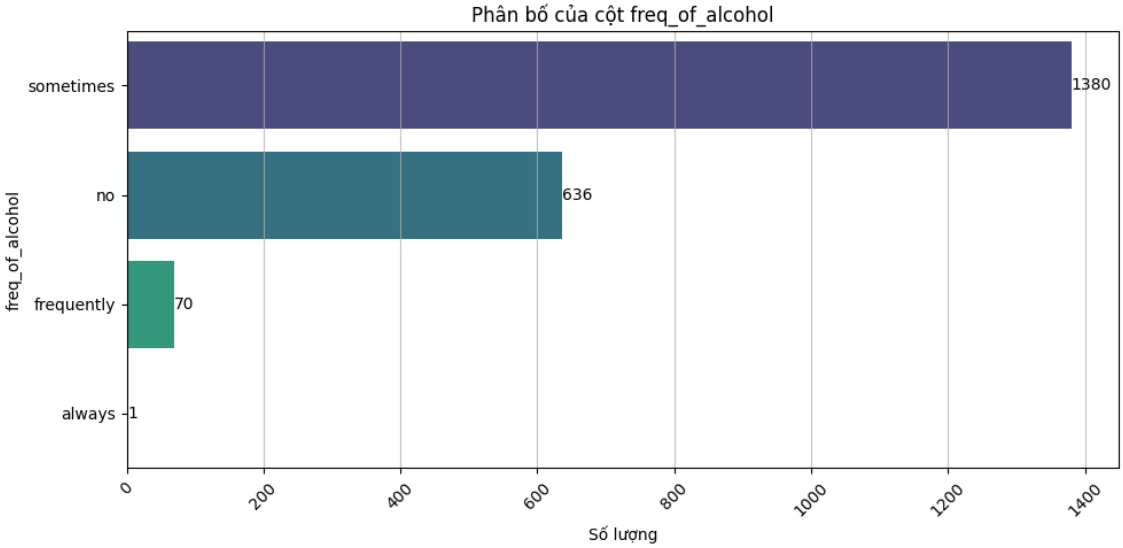
3.2 Đánh giá với các đặc trưng mang kiểu dữ liệu phân loại

Giới tính (Gender): Dữ liệu phân bổ khá đồng đều giữa nam và nữ, đảm bảo tính cân bằng trong phân tích theo giới.



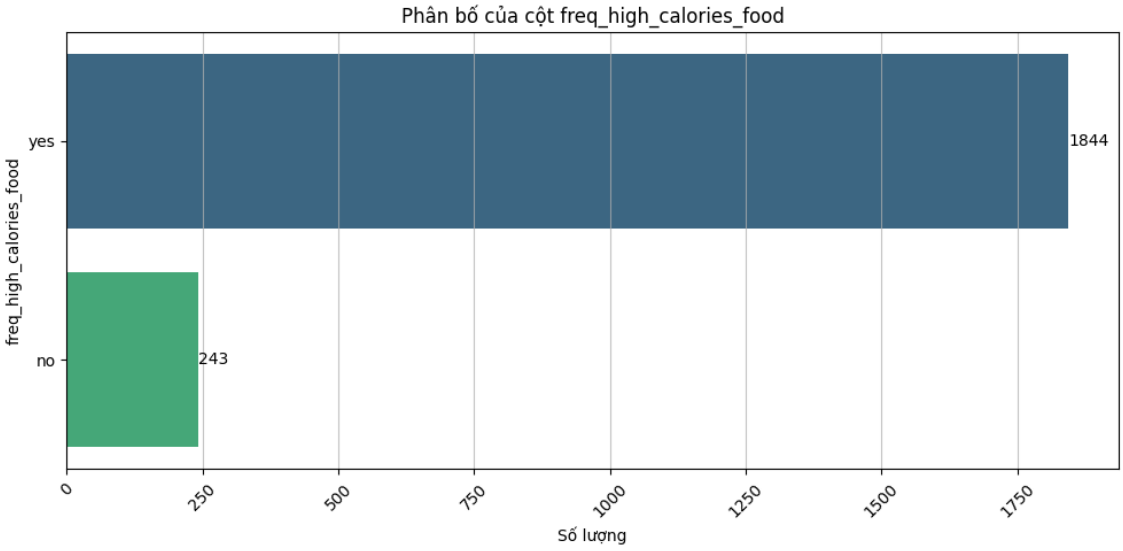
Hình 4.1: Biểu đồ thể hiện sự phân bố của đặc trưng "Gender"

Cột tần suất sử dụng cồn (freq of alcohol) tần suất chiếm nhiều nhất thường là sometimes xếp vị trí thứ 2 là no và đáng mừng là tần suất always chỉ xuất hiện rất ít hầu như được xem là không có



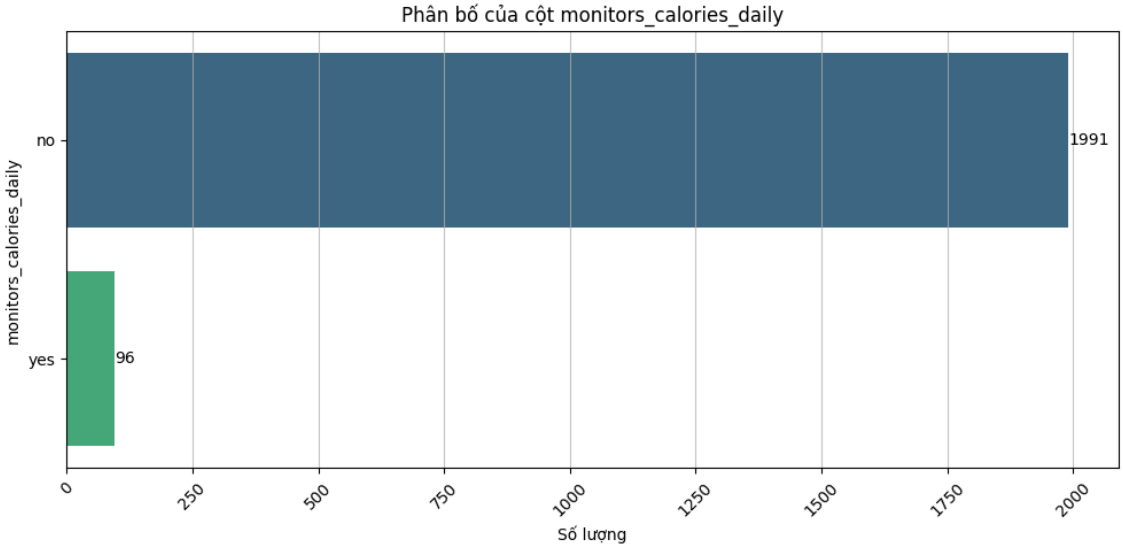
Hình 4.2: Biểu đồ thể hiện sự phân bố của đặc trưng "Freq of alcohol"

Phân bố của cột tần suất tiêu thụ món ăn nhiều calories đạt được nhiều nhất vẫn là yes chiếm cũng khá nhiều so với tổng thể



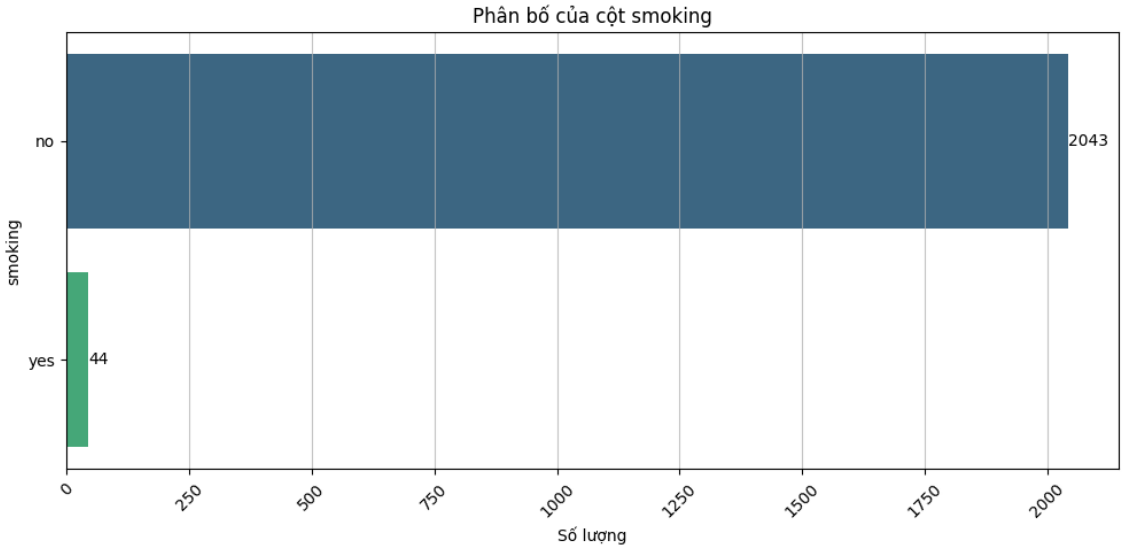
Hình 4.3: Biểu đồ thể hiện sự phân bố của đặc trưng "Gender"

Việc quan sát lượng calories nạp vào cơ thể (monitors_calories_daily) nhiều nhất chiếm đa số là không cũng dễ hiểu khi đa số các mẫu trong tập dữ liệu đều thuộc nhóm thừa cân hoặc béo phì



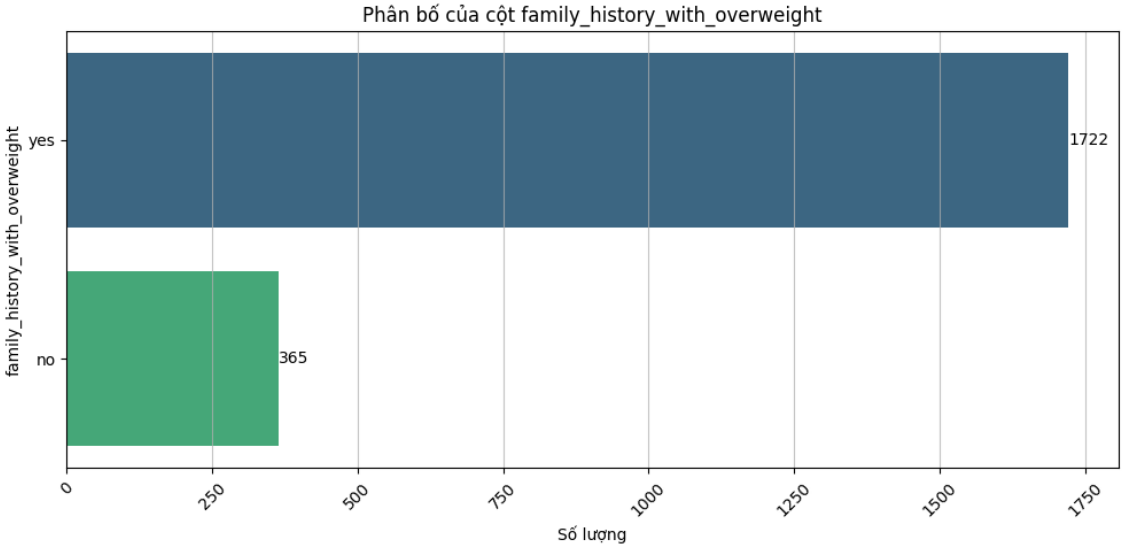
Hình 4.4: Biểu đồ thể hiện sự phân bố của đặc trưng "Monitor calories daily"

Phân bố của cột hút thuốc (smoking) chiếm nhiều nhất là lớp No cho thấy đặc trưng hút thuốc không có ảnh hưởng gì nhiều đến với khả năng thừa cân hay béo phì



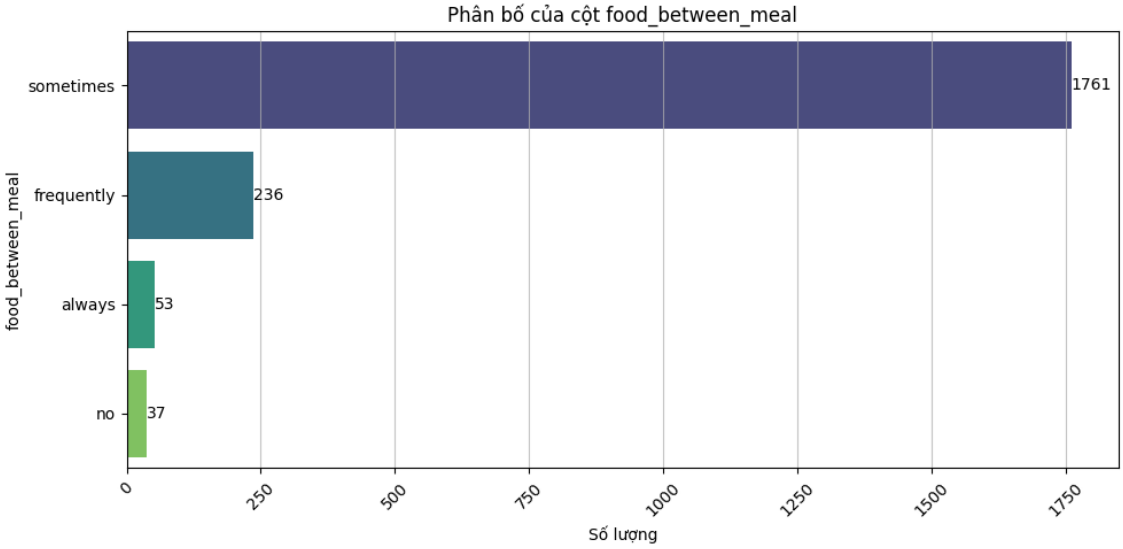
Hình 4.5: Biểu đồ thể hiện sự phân bố của đặc trưng "Smoking"

Phân bố của tiền xử việc gia đình bị thừa cân (family_history_with_overnight) Chiếm đa số vẫn là yes diều này cho thấy đặc trưng này cũng phần nào ảnh hương nhỏ ít nhiều đến với biến mục tiêu

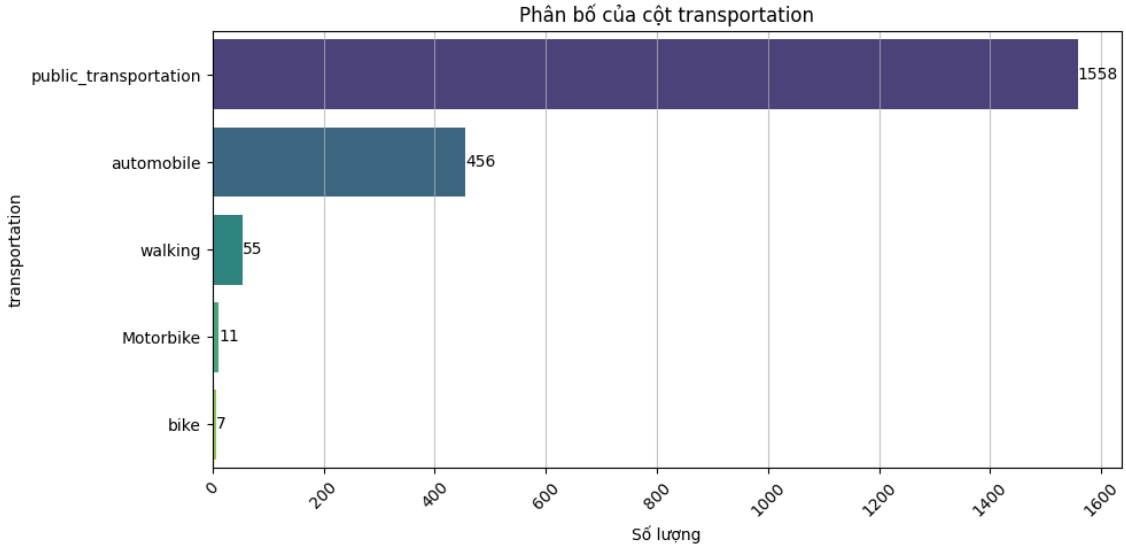


Hình 4.6: Biểu đồ thể hiện sự phân bố của đặc trưng "Family history with overweight"

Phân bố của hai đặc trưng cuối cùng thức ăn giữa các buổi (food_between_meal) và phương tiện di chuyển (transportation) cũng không thể nói gì nhiều đến khả năng ảnh hưởng đến biến mục tiêu đặc trưng món ăn giữa các bữa ăn thì tần suất thường phân bố cáo và phương tiện công cộng cũng có phân bố cao nhất trong đặc trưng "phương tiện di chuyển" và rất ít người đi bộ sử dụng xe đạp

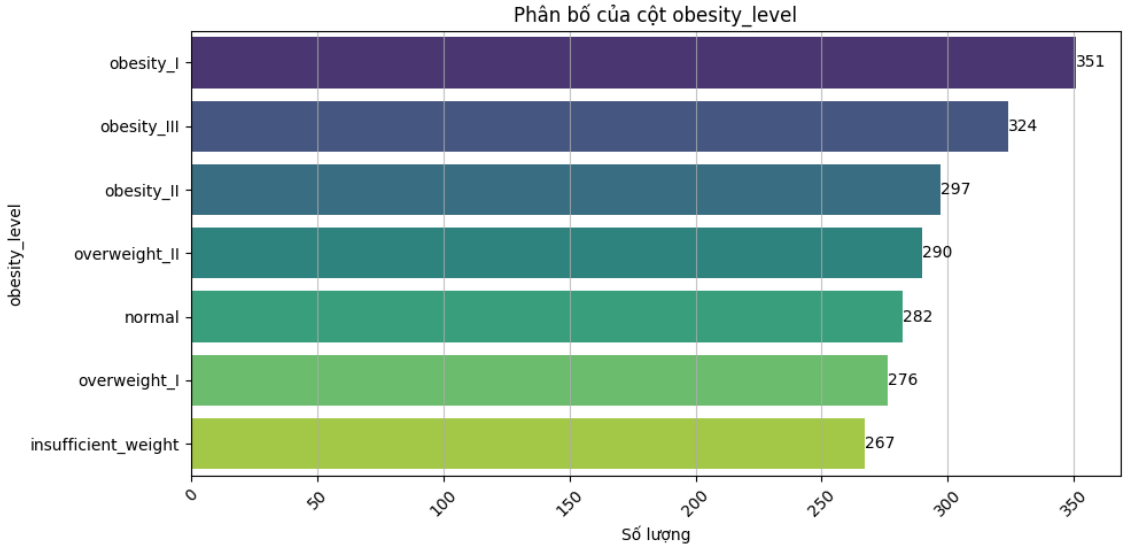


Hình 4.7: Biểu đồ thể hiện sự phân bố của đặc trưng "Food between meal"



Hình 4.8: Biểu đồ thể hiện sự phân bố của đặc trưng "Transportation"

Biến mục tiêu "obesity level"gồm 7 lớp từ thiếu cân đến béo phì độ III. Qua biểu đồ tần suất, có thể quan sát. Lượng mẫu "Béo phì loại 1" nhiều hơn cả thấy và thấp nhất chính là nhóm bị thiếu cân kết hợp cũng biểu đồ tròn đã quan sát ở trên ta có thể không cần phải sử dụng các phương pháp làm giảm sự mất cân bằng dữ liệu độ chênh lệch giữa các lớp với nhau có thể được xem như không đáng kể



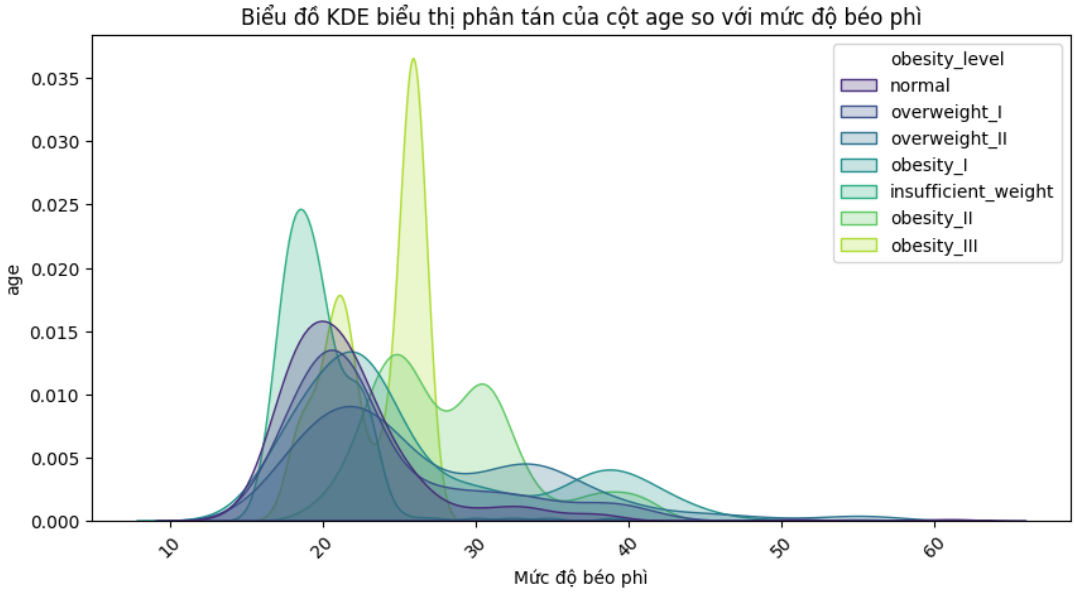
Hình 4.9: Biểu đồ thể hiện sự phân bố của đặc trưng "Obesity level"

3.3 Mối quan hệ giữa các đặc trưng và biến mục tiêu

Đối với từng lớp khác nhau trong biến mục tiêu chung có những mối quan hệ với những đặc trưng cụ thể như sau:

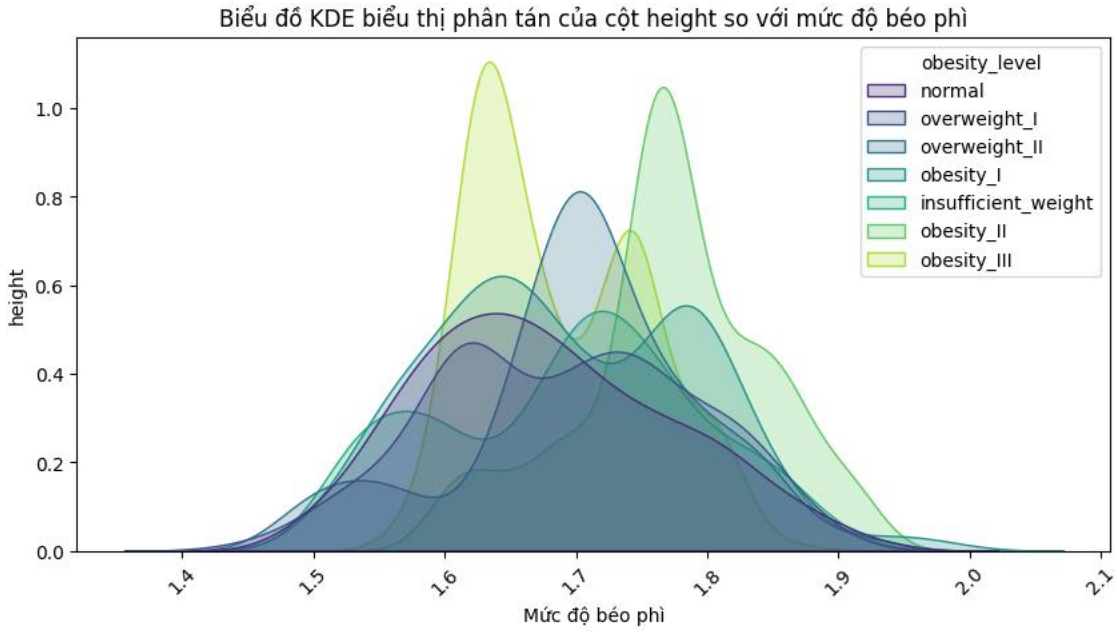
Với tuổi tác, cân nặng, chiều cao

- Với tuổi tác: Như đã nói ở trên độ tuổi thường nằm trong khoảng từ hơn 20 đến gần 30 tuổi đặc biệt với béo phì loại 3 có tần suất cũng như nhiều mẫu nhất nằm trong độ tuổi từ giữa độ tuổi từ 20 đến 30 tuổi và đa số những lớp còn lại thì đều nằm trong khoảng tuổi thanh thiếu niên



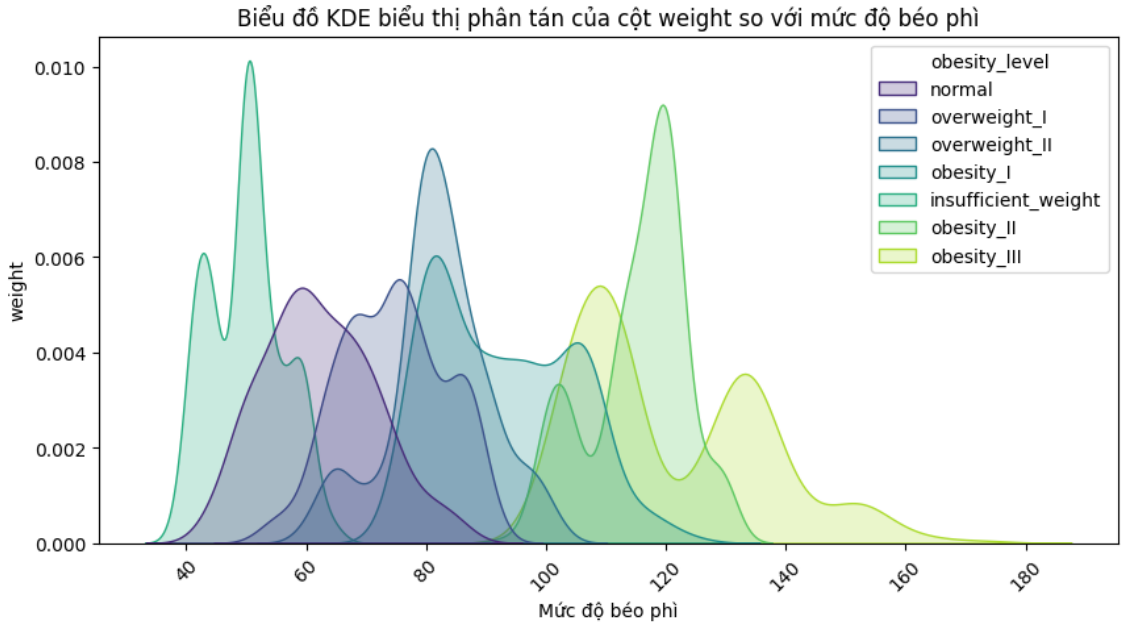
Hình 5.1: Biểu đồ thể thị sự phân tán của cột "Age" với mức độ béo phì

- Với chiều cao: Phân bố nhiều nhất đối với chiều cao 1m7 tuy nhiên đối với nhóm mắc béo phì loại 3 chiều cao sẽ giao động trong chiều cao 1m6 với các nhóm khác thì nằm trong khoảng trung bình 1m7. Đặc biệt với những lớp bị thiếu cân lại có chiều cao nằm trong khoảng 1m8 đến 1m9



Hình 5.2: Biểu đồ thể thị sự phân tán của cột "Height" với mức độ béo phì

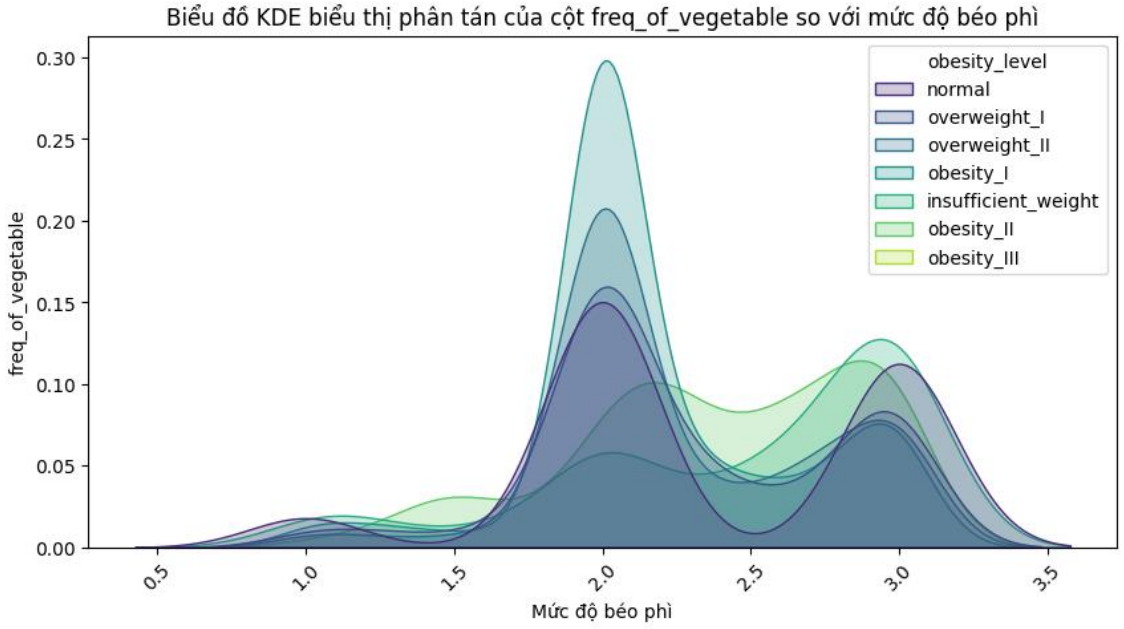
- Với cân nặng: Quan sát biểu đồ ta có thể thấy việc từng lớp béo phì phân chia theo từng nhóm dựa theo cân nặng. Đồng thời ta cũng có thể dự đoán được trong tập dữ liệu có nhiều nhất có thể là những người thuộc lớp thiếu cân và béo phì loại 2



Hình 5.3: Biểu đồ thể thị sự phân tán của cột "Weight" với mức độ béo phì

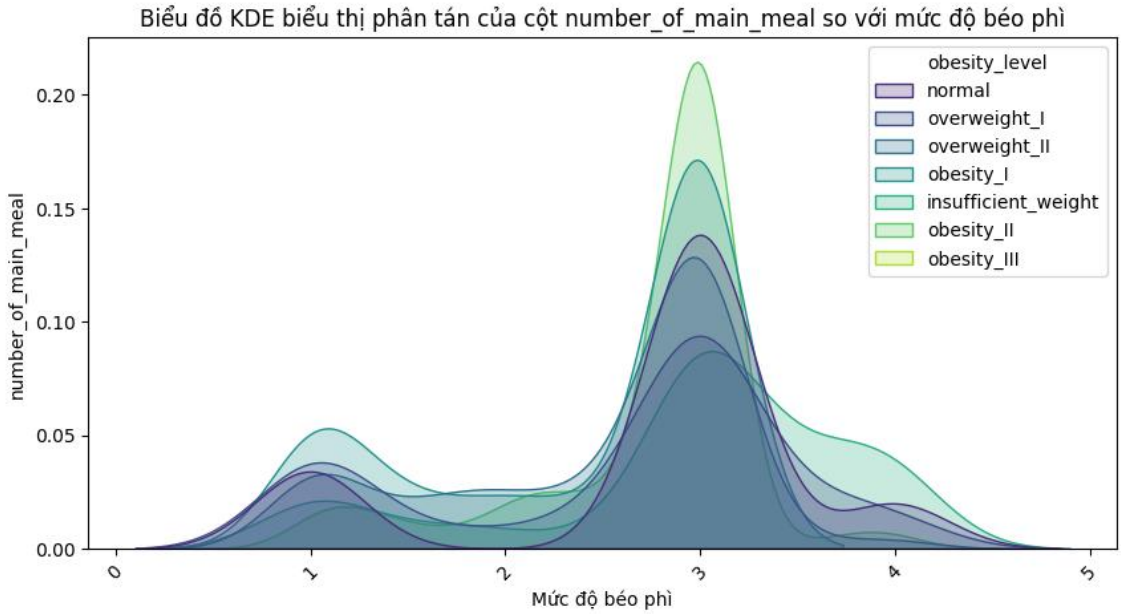
Đối với lối sống

- Tần suất tiêu thụ rau: việc tiêu thụ rau ở mỗi lớp trong biến mục tiêu cũng không có quá nhiều điều cần để tâm đến tuy nhiên giống như phân bố thì tần suất tiêu thụ rau đa số nằm trong khoảng từ 2 đến dưới 3 đạt cao nhất, điều đáng chú ý là ở mức thừa cân lại một lại có tuần suất ăn rau ở mức 2 cao nhất so với những lớp còn lại



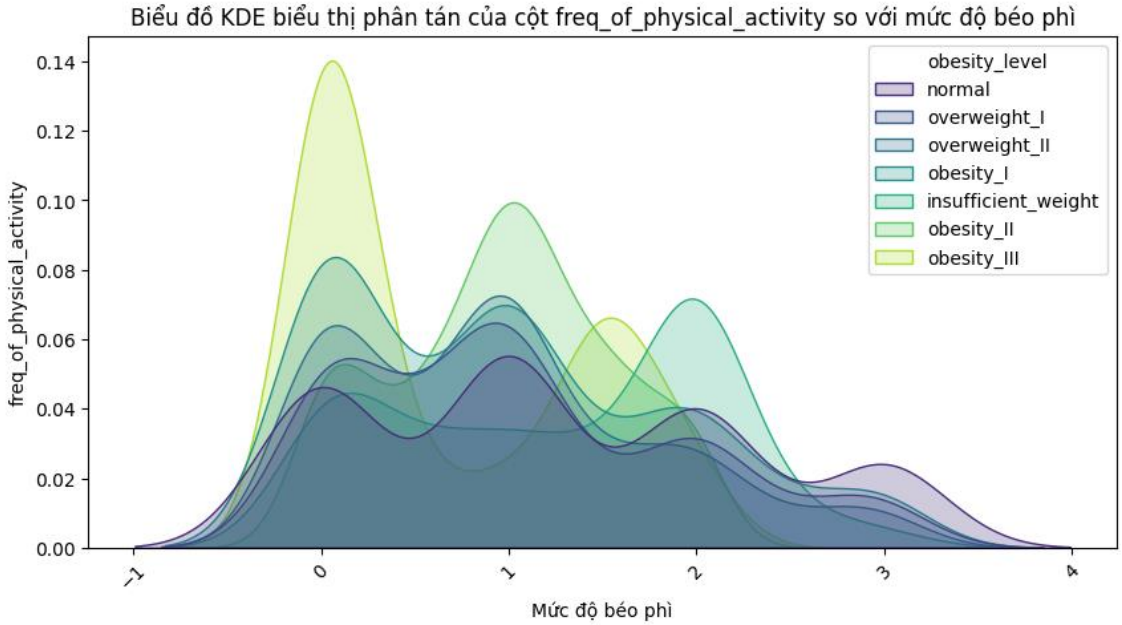
Hình 5.4: Biểu đồ thể thị sự phân tán của cột "Freq of vegetable" với mức độ béo phì

- Số lượng bữa ăn chính: Tập trung nhiều vào. Lượng bữa ăn được nạp vào cơ thể Chủ yếu nhiều nhát vẫn là ở các mức 1, 2, 3 lít bữa ăn mỗi ngày tuy nhiên có một chú ý nhỏ rằng ở mức độ béo phì loại 3 lượng bữa ăn chính được nạp vào cơ thể nhiều hơn hết thảy cũng không ít những bữa ăn từ 4 bữa trở lên



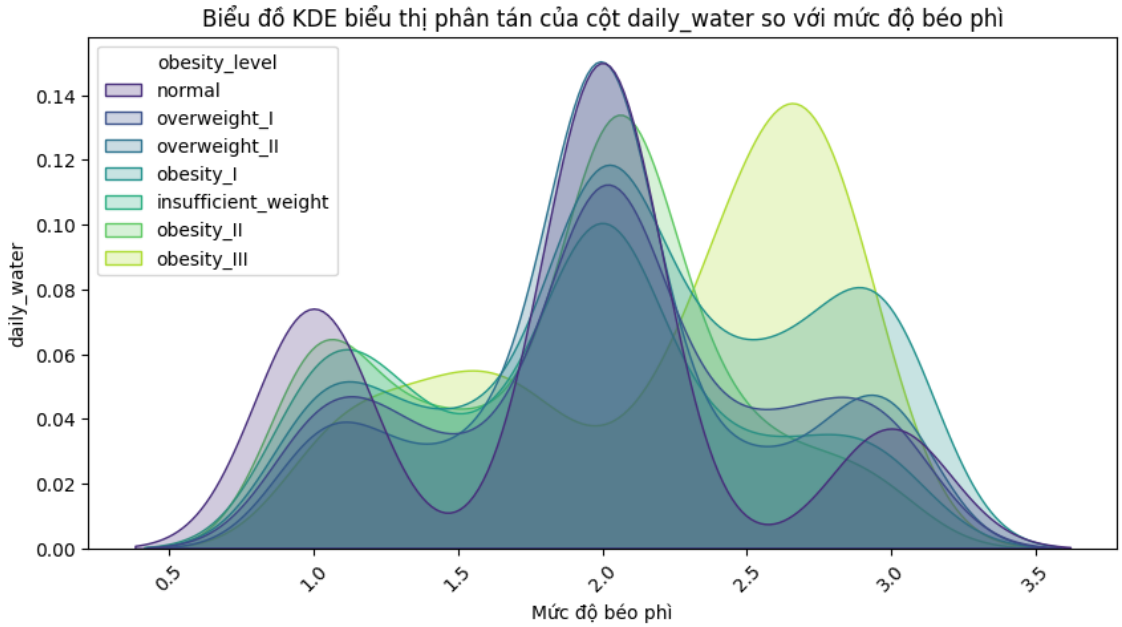
Hình 5.5: Biểu đồ thể thị sự phân tán của cột "Number of main meal" với mức độ béo phì

- Tần suất vận động: Dường như cũng đánh giá nhiều về tần suất vận động với mỗi loại béo phì khác nhau. Đặc biệt với người béo phì cao hay chính xác hơn là béo loại 3 thường có tần suất vận động rất thấp phân bố lệch nhiều về phía bên trái và đạt đỉnh cao nhất trong khoảng từ -1 đến 1. Tuy nhiên với người có cân nặng bình thường hay thưa cân nhẹ có mức vận động cao hơn và nghiêng về bên phải



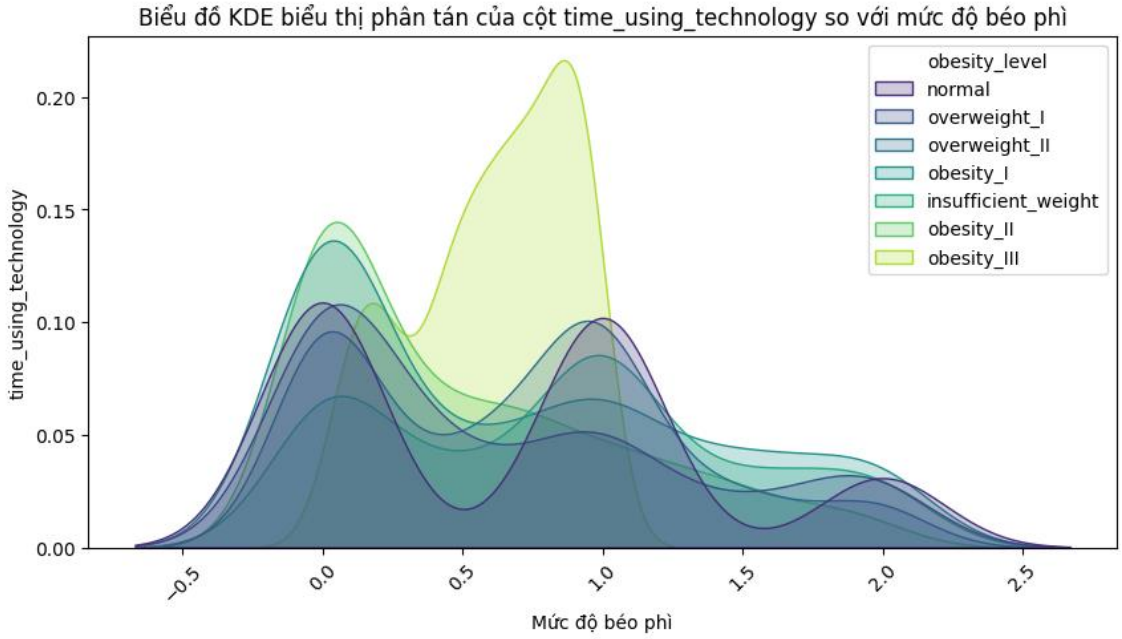
Hình 5.6: Biểu đồ thể thị sự phân tán của cột "Freq of physical activity" với mức độ béo phì

- Lượng nước nạp vào cơ thể: Các nhóm thuộc mức động bình thường hay thừa cân nhẹ sẽ có xu hướng tiêu thụ nước đầy đủ và đạt yêu cầu lượng nước tiêu thụ trong 1 ngày điều này đáp ứng việc đảm bảo lượng nước vừa đủ cung cấp để duy trì cơ thể



Hình 5.7: Biểu đồ thể thị sự phân tán của cột "Daily water" với mức độ béo phì

- Thời gian sử dụng công nghệ: Nhóm béo phì cao có tần suất sử dụng điện thoại cao hơn và tập trung nhiều vào các khoảng 0.5 đến 1 tương ứng 5-10 giờ/ ngày, các nhóm khác có phân bố rộng những nhiều hơn và thường sẽ có xu hướng ít sử dụng các thiết bị điện tử hơn



Hình 5.8: Biểu đồ thể thị sự phân tán của cột "Time using technology" với mức độ béo phì

Kết luận

Có thể đánh giá được rằng những người thường hay bị thừa cân hay béo phì thì thường thuộc vào các độ tuổi khoảng từ 15 đến 35 tuổi. Đặc biệt béo phì loại 3 cao nhất trong khoảng độ tuổi từ 20 đến 30 tuổi

Với cân nặng thì có thể phân chia một khoảng cách khá rõ ràng thành các khoảng trên biểu đồ

Phản ánh rõ thực tế rằng người bị bé0 phì thường có tần suất tập luyện thể dục ít hơn nếu mức độ càng tăng thì tần suất tập thể dục càng thấp

Việc sử dụng thiết bị điện tủ phân bố khá đều. Tuy nhiên trong khoảng 0.0 đến 1.0 béo phì loại 3 lại có đỉnh cao nhất

3.4 Phân tích tương quan

Để có thể đánh giá sâu cũng như có cái nhìn tổng quát về mối tương quan giữa các biến đặc trưng với nhau, có thể dựa vào những tiêu chí như hệ số tương quan mục đích giúp biểu thị mối quan hệ mạnh hay yếu giữa các đặc trưng với nhau:

Công thức:

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Trong đó:

Cov(x, y) Hiệu phương sai của biến x và y

$\sigma_x \sigma_y$*: Tích phương sai của biến x và biến y*

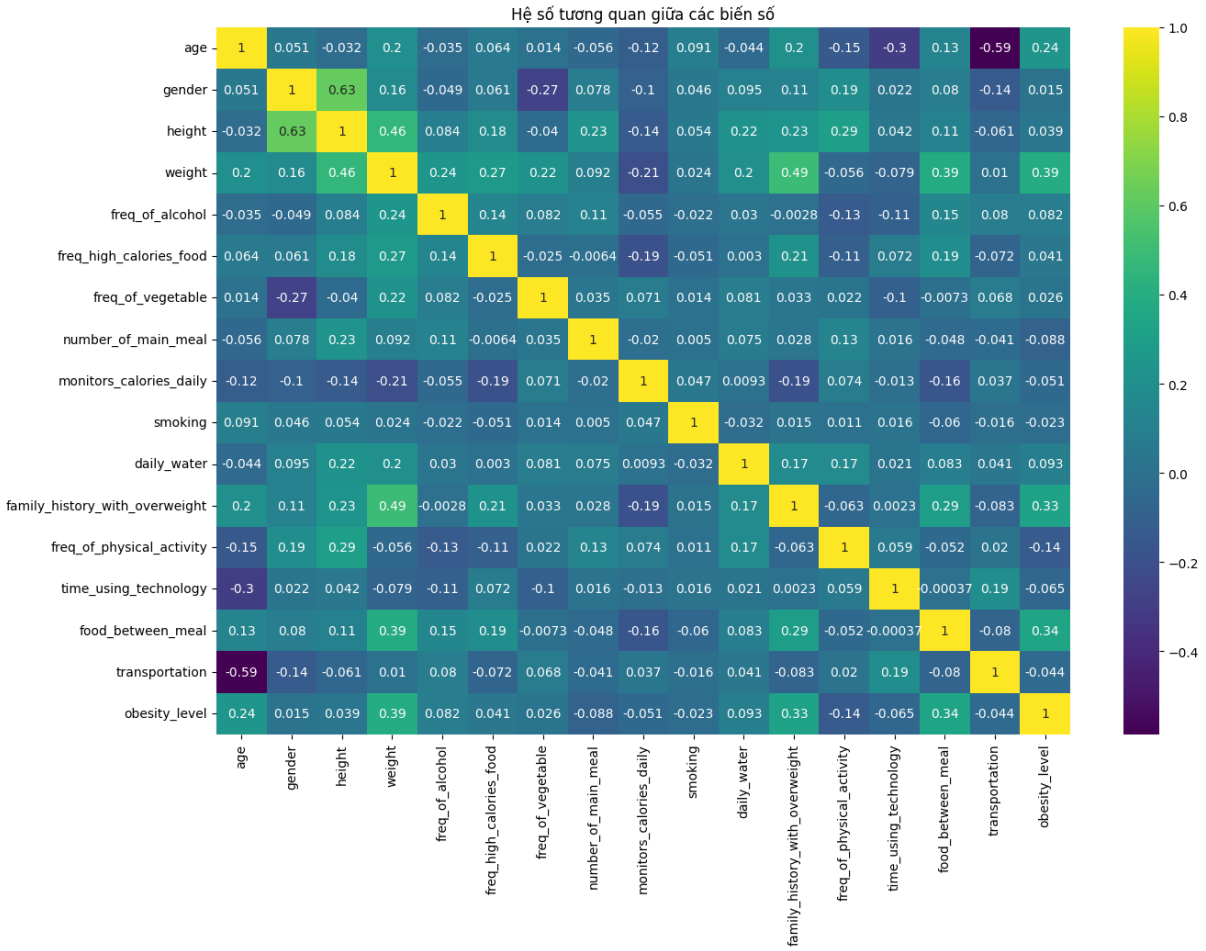
Nếu:

$|\rho_{xy}| > 0.7$*: Mối tương quan mạnh*

$0.5 < |\rho_{xy}| < 0.7$*: Mối tương quan trung bình*

$0.3 < |\rho_{xy}| < 0.5$*: Mối tương quan yếu*

$|\rho_{xy}| < 0.3$*: Tương quan rất yếu hoặc không có tương quan tuyến tính*



Hình 6: Biểu đồ nhiệt biểu thị mối quan hệ giữa các đặc trưng

Phân tích tương quan (correlation heatmap) cho thấy:

Biến mục tiêu (obesity_level) có mối quan hệ cụ thể với các đặc trưng

Tương quan mạnh:

Cân nặng (weight): Có mối tương quan dương rất mạnh với mức độ béo phì (r = 0.94). Điều này hoàn toàn hợp lý, vì cân nặng là một thành phần quan trọng trong việc tính toán chỉ số BMI – một chỉ số thường được dùng để phân loại mức độ béo phì.

Tiền sử gia đình béo phì (family_history_with_overweight): Cho thấy mối tương quan dương tương đối mạnh với "obesity_level" (r = 0.83). Kết quả này nhấn mạnh vai trò của yếu tố di truyền hoặc môi trường gia đình trong việc hình thành xu hướng béo phì, vượt qua cả các yếu tố hành vi cá nhân.

Tương quan vừa và yếu:

Tần suất ăn thực phẩm giàu calo (freq_high_calories_food): Có tương quan dương ở mức tương đối với mức độ béo phì (r = 0.34). Điều này phản ánh rằng tiêu thụ thường xuyên các loại thực phẩm có hàm lượng calo cao có thể góp phần đáng kể vào việc tăng cân và nguy cơ béo phì.

Tần suất hoạt động thể chất (freq_of_physical_activity): Có tương quan âm tương đối với "obesity_level" (r = -0.34). Người ít vận động có xu hướng có mức độ béo phì cao hơn.

Tuổi (age): Tương quan dương yếu với mức độ béo phì (r = 0.24), cho thấy rằng yếu tố tuổi tác có thể ảnh hưởng đến nguy cơ béo phì, mặc dù mức độ ảnh hưởng là không cao.

Giới tính (gender): Có mối tương quan dương yếu (r = 0.13), cho thấy sự khác biệt về mức độ béo phì giữa nam và nữ là không đáng kể trong tập dữ liệu này**.**

Thời gian sử dụng công nghệ (time_using_technology): Tương quan dương yếu (r = 0.19). Việc dành nhiều thời gian cho các thiết bị công nghệ có thể làm giảm hoạt động thể chất, từ đó làm tăng nguy cơ béo phì.

Ăn vặt giữa các bữa (food_between_meal) và tần suất uống rượu (freq_of_alcohol): Đều có tương quan dương yếu (r lần lượt là 0.08 và 0.08). Các yếu tố này có thể góp phần nhỏ vào việc gia tăng mức độ béo phì.

Tương quan âm yếu:

Tần suất ăn rau (freq_of_vegetable), số bữa ăn chính mỗi ngày (number_of_main_meal), lượng nước uống hàng ngày (daily_water), theo dõi lượng calo hàng ngày (monitors_calories_daily) và hút thuốc (smoking) đều có tương quan âm yếu (r dao động từ -0.01 đến -0.06). Mặc dù các yếu tố này có thể ảnh hưởng đến tình trạng cân nặng, những mối liên hệ thống kê trong tập dữ liệu hiện tại là không đáng kể**.**

Mối tương quan khác

Cân nặng và chiều cao (weight - height): Có mối tương quan dương mạnh (r = 0.46), phản ánh mối quan hệ tự nhiên giữa chiều cao và cân nặng ở con người.

Giới tính với chiều cao và cân nặng: Giới tính có tương quan dương tương đối với chiều cao (r = 0.63) và cân nặng (r = 0.63), cho thấy nam giới thường cao và nặng hơn nữ giới trong dữ liệu thu thập.

Tần suất hoạt động thể chất và thời gian sử dụng công nghệ: Có mối tương quan âm yếu (r = -0.07), cho thấy việc sử dụng công nghệ nhiều có thể làm giảm thời gian vận động.

Tiền sử gia đình béo phì với các hành vi cá nhân: Các biến hành vi như ăn uống, hút thuốc, hoặc uống nước cho thấy mối tương quan yếu hoặc không đáng kể với "family_history_with_overweight". Điều này gợi ý rằng yếu tố di truyền hoặc ảnh hưởng môi trường gia đình có thể tác động trực tiếp đến nguy cơ béo phì hơn là thông qua các hành vi cụ thể.

Kết luận

Phân tích tương quan cho thấy cân nặng và tiền sử gia đình là hai yếu tố ảnh hưởng mạnh nhất đến mức độ béo phì. Một số hành vi như tiêu thụ thực phẩm giàu calo hoặc thiếu hoạt động thể chất cũng góp phần đáng kể, trong khi các yếu tố như

giới tính, hút thuốc hay theo dõi calo hàng ngày có vai trò yếu hơn. Những phát hiện này cung cấp nền tảng quan trọng cho việc xây dựng mô hình dự đoán béo phì cũng như thiết kế các chương trình can thiệp phù hợp trong thực tế.

3.5 Phát hiện giá trị ngoại lai (Outlier)

Thông qua biểu đồ boxplot, một số mẫu có giá trị ngoại lệ ở các đặc trưng như **tuổi, cân nặng, chiều cao** được phát hiện. Tuy nhiên, không có ngoại lệ nghiêm trọng đến mức cần loại bỏ, vì chúng phản ánh sự đa dạng thực tế của dữ liệu.

4. Lựa chọn và xây dựng mô hình

Sau khi hoàn thành các bước tiền xử lý dữ liệu và khai phá mối tương quan giữa các đặc trưng và biến mục tiêu, cùng với việc phân tích phân bố của các lớp trong biến mục tiêu, giai đoạn tiếp theo là lựa chọn và xây dựng mô hình học máy phù hợp cho bài toán.

Trong nghiên cứu này, biến mục tiêu obesity_level là một biến phân loại (categorical variable) bao gồm bảy lớp distinct (distinct classes), đại diện cho các mức độ cân nặng khác nhau từ thiếu cân đến béo phì cấp độ III. Do đó, bài toán này được định nghĩa là một bài toán **phân loại đa lớp (multi-class classification)**.

Dựa trên các nguyên lý hoạt động, ưu điểm và nhược điểm của các thuật toán phân loại đã được trình bày ở các chương trước, năm mô hình học máy sau đây đã được lựa chọn và thử nghiệm để đánh giá hiệu suất trong bài toán này:

***Hồi quy Logistic (Logistic Regression - đa lớp)*:** Được chọn làm mô hình cơ sở (baseline model) do tính đơn giản, khả năng diễn giải cao và hiệu quả trong việc thiết lập một điểm tham chiếu ban đầu cho các mô hình phức tạp hơn.

Cây Quyết định (Decision Tree): Lựa chọn này dựa trên khả năng của thuật toán trong việc mô hình hóa các mối quan hệ phi tuyến tính giữa đặc trưng và nhãn lớp. Ngoài ra, tính dễ trực quan hóa và phân tích của Decision Tree cũng là một yếu tố quan trọng.

Rừng Ngẫu nhiên (Random Forest): Là một thuật toán học tập tổ hợp (ensemble learning) mạnh mẽ, Random Forest giúp giảm thiểu hiện tượng quá khớp (overfitting) thường gặp ở cây quyết định đơn lẻ thông qua việc kết hợp nhiều cây quyết định.

K-Nearest Neighbors (KNN): Đây là một thuật toán phân loại không tham số (non-parametric) dựa trên khoảng cách, được đánh giá là hiệu quả trong các bài toán phân loại khi dữ liệu đã được chuẩn hóa tốt.

***Support Vector Machine - SVM*:** SVM được xem xét do khả năng tìm kiếm siêu phẳng phân tách tối ưu trong không gian đa chiều, đặc biệt thích hợp cho các bài toán phân loại có biển phân cách rõ ràng. Mô hình này được thử nghiệm với các kernel phi tuyến để nắm bắt các mối quan hệ phức tạp.

Trước khi tiến hành huấn luyện các mô hình, tập dữ liệu đã được chia thành tập huấn luyện (training set) và tập kiểm tra (test set) theo tỷ lệ 80:20. Đồng thời, kỹ thuật **chuẩn hóa dữ liệu (standardization)** đã được áp dụng trên tất cả các đặc trưng để

đảm bảo các đặc trưng có cùng thang đo, từ đó tăng cường độ ổn định và hiệu suất cho các thuật toán nhạy cảm với thang đo như SVM và KNN. Quá trình tiền xử lý này đã được thực hiện và chi tiết trong các bước trước của nghiên cứu.

Sau quá trình huấn luyện trên tập dữ liệu đã được xử lý, hiệu suất của mỗi mô hình được đánh giá thông qua một bộ các chỉ số đo lường hiệu suất tiêu chuẩn trong học máy, bao gồm:

Độ chính xác (Accuracy): Đo lường tỷ lệ các mẫu được phân loại đúng trên tổng số mẫu.

Ma trận nhầm lẫn (Confusion Matrix): Cung cấp cái nhìn chi tiết về số lượng True Positives, True Negatives, False Positives và False Negatives cho từng lớp.

Precision: Đo lường tỷ lệ các dự đoán dương tính thực sự là dương tính.

Recall (Sensitivity): Đo lường tỷ lệ các trường hợp dương tính thực sự được mô hình nhận diện.

F1-Score: Là trung bình điều hòa của Precision và Recall, cung cấp một độ đo cân bằng khi có sự mất cân bằng giữa hai chỉ số này.

Macro Average và Weighted Average: Được sử dụng để tổng hợp các chỉ số Precision, Recall, và F1-Score trên toàn bộ các lớp, giúp phản ánh hiệu suất trên toàn bộ các lớp một cách khách quan, đặc biệt quan trọng trong trường hợp dữ liệu có thể có sự mất cân bằng nhẹ về số lượng mẫu giữa các lớp.

Việc sử dụng đa dạng các chỉ số đánh giá này nhằm đảm bảo một cái nhìn toàn diện và khách quan về hiệu suất của mô hình, giúp tránh những thiên lệch có thể phát sinh khi chỉ dựa vào một chỉ số duy nhất, đặc biệt trong các tập dữ liệu có khả năng mất cân bằng nhẹ như trong bài toán này.

5. Tiêu chí đánh giá mô hình

- Độ chính xác (Accuracy): là một chỉ số đánh giá hiệu suất phổ biến trong xá mô hình phân loại. Nó thể hiện tỷ lệ phần trăm số lượng dự đoán đúng trên tổng số dự đoán. Accuracy cho biết mô hình hoạt động tốt như thế nào trên toàn bộ tập dữ liệu. Nếu độ chính xác càng cao đồng nghĩa với việc mô hình dự đoán càng chính xác

Công thức tính:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Trong đó:

TP (True Positive): Dự đoán đúng trường hợp dương tính

TN (True Negative): Dự đoán đúng trường hợp âm tính

FP (False Positive): Dự đoán sai – dự đoán dương tính nhưng thực tế là âm

FN (False Negative): Dự đoán sai – dự đoán âm tính nhưng thực tế là dương

- **Precision (độ chính xác theo nghĩa hẹp)** là một chỉ số đo lường độ chính xác của dự đoán dương tính — tức là trong số tất cả các dự đoán là một lớp nào đó, thì bao nhiêu trong số đó là đúng. Precision càng cao thì mô hình ít bị nhầm lẫn khi dự đoán. Ngược lại nếu chúng càng thấp thì mô hình có khả năng cao bị dự đoán sai tức nhầm lẫn với các lớp khác

Công thức tính:

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

Trong đó:

TP (True Positive): Dự đoán đúng lớp đó

FP (False Positive): Dự đoán sai rằng là lớp đó (nhưng thực ra không phải)

Bên cạnh đó nếu ta áp dụng phương pháp tính toàn bộ Precision của tất cả các lớp phân loại có trong biến mục tiêu, rồi sau đó cộng chúng lại rồi lấy trung bình đều. Phương pháp này giúp tìm ra **Precision macro**, giúp cho ta có cái nhìn tổng quát chung về tính chính xác đối với từng lớp

Công thức tính:

$$MacroPrecision = \frac{1}{N}\sum_{i=1}^{N} Precision_i$$

Trong đó:

N: Số lượng lớp trong biến mục tiêu

Precision: Precision của từng lớp

Để xác định một mô hình có thể dự đoán đúng hay sai cần rất nhiều yếu tố khác nhau. Việc chỉ dựa vào một giá trị có thể gây nên các đánh giá sai. Ngoài accuracy, precision cũng là một yếu tố cần phải được quan tâm đặc biệt là trong các bài toán khi chi phí của việc dự đoán sai là cao và xuất hiện nhiều các "dự báo giá" .

Ngoài phương pháp tính Precision của từng lớp rồi sau đó lấy trung bình để tìm ra Macro Precision thì còn phương pháp lấy giá trị trung bình của Precision ứng với từng lớp, trong đó mỗi lớp được gán một trọng số dựa trên số lượng mẫu thực tế của lớp đó trong tập dữ liệu. Phương pháp này được gọi là Weighted Precision

Công thức tính:

$$WeightedPrecision = \sum_{c=1}^{N} (Precision_c \; x \; \frac{|D_c|}{\sum_{i=1}^{N} |D_i|})$$

Trong đó:

N: Số lượng lớp trong biến mục tiêu

Precision: Precision của lớp c

$|D_c|$ *: Số lượng mẫu của tập hợp các trường hợp thực tế thuộc lớp c*

$\sum_{i=1}^{N} |D_i|$: *Tổng số lượng mẫu trong toàn bộ tập dữ liệu (tập dữ liệu kiểm thử)*

$\frac{|D_c|}{\sum_{i=1}^{N} |D_i|}$: *Đây chính là trọng số của lớp c, phản ánh tỷ lệ phần trăm số lượng mẫu của lớp đó so với tổng số mẫu trong tập dữ liệu.*

Phương pháp này được sử dụng trong các trường hợp nhằm xử lý việc mất cân bằng dữ liệu giúp đảm bảo rằng các lớp có số lượng mẫu lớn hơn sẽ có ảnh hưởng lớn hơn đến giá trị Precision của tổng thể. Điều này có nghĩa là hiệu suất của mô hình trên các lớp phổ biến hơn sẽ được ưu tiên hơn trong chỉ số cuối cùng

Cung cấp một cái nhìn về độ chính xác trung bình của mô hình trong đó hiệu suất trên các lớp có tần suất cao hơn sẽ "quan trọng" hơn trong việc quyết định giá trị tổng thể

Macro được sử dụng khi tất cả các lớp đều được coi là quan trọng như nhau macro precision rất nhạy cảm với hiệu suất lớp thiểu số , còn Weight hoạt động tốt với dữ liệu bị mất cân bằng rằng lớp nào nhiều hơn thì ưu tiên hơn

- Recall (Độ phủ/Độ nhạy): Là một trong những thước đo hiệu suất then chốt của một mô hình phân loại. Được xác định là tỷ lệ các trường hợp đúng phân loại chính xác bởi mô hình trên tổng số các trường hợp đúng có trong tập dữ liệu. Recall càng cao mô hình dự đoán càng thiếu chính xác (nhầm lẫn) và ngược lại

Công thức tính:

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

Trong đó:

TP (True Positives): Dự đoán đúng của lớp đó

FN (False Negatives): Số lượng các mẫu được phân loại là sai nhưng thực tế lại là đúng. Đây là những trường hợp đúng mà mô hình đã bỏ sót

Mẫu số (TP+FN) đại diện cho tổng số lượng các trường hợp đúng thực sự tồn tại trong tập dữ liệu

Tương tự như Precision đối với bài toán phân loại đa lớp ta cũng có thể sinh Recall của từng lớp trong biến mục tiểu rồi sau đó lấy trung bình cộng để tìm ra MacroRecall

Công thức tính:

$$MacroRecall = \frac{1}{N}\sum_{i=1}^{N} Recall_i$$

Trong đó:

N: Số lượng lớp trong biến mục tiêu

$Recall_i$: Recall của từng lớp

Recall thường được xem xét cùng với Precision trong khi Recall tập trung vào việc bao nhiêu phần trăm các trường hợp đúng được tìm thấy thì Precision tập trung vào việc bao nhiêu phần trăm các dự đoán đúng của mô hình là đúng

Tương tự như Precision ta cũng có Weighted Recall được định nghĩa là tổng có trọng số của Recall của từng lớp, với trọng số là tỷ lệ lượng mẫu thực tế của lớp đó trong tập dữ liệu

Công thức tính:

$$WeightedRecall = \sum_{c=1}^{N} (Recall_c \; x \; \frac{|D_c|}{\sum_{i=1}^{N} |D_i|})$$

Trong đó:

N: Số lượng lớp trong biến mục tiêu

$Recall_c$: Recall của lớp c

$|D_c|$: Số lượng mẫu của tập hợp các trường hợp thực tế thuộc lớp c

$\sum_{i=1}^{N} |D_i|$: Tổng số lượng mẫu trong toàn bộ tập dữ liệu (tập dữ liệu kiểm thử)

$\frac{|D_c|}{\sum_{i=1}^{N} |D_i|}$: Đây chính là trọng số của lớp c, phản ánh tỷ lệ phần trăm số lượng mẫu của lớp đó so với tổng số mẫu trong tập dữ liệu.

Thông thường, có một sự đánh đổi vốn có giữa Precision và Recall. Việc cố gắng tối đa hóa Recall (ví dụ, bằng cách làm cho mô hình nhạy hơn, hạ thấp ngưỡng phân loại) có thể dẫn đến việc tăng số lượng FP, từ đó làm giảm Precision. Ngược lại, việc tối đa hóa Precision (bằng cách chỉ đưa ra dự đoán dương tính khi rất chắc chắn, nâng cao ngưỡng phân loại) có thể làm tăng số lượng FN, dẫn đến giảm Recall.

Sự lựa chọn giữa việc ưu tiên Recall hay Precision phụ thuộc vào chi phí tương đối của False Negatives so với False Positives trong ngữ cảnh bài toán cụ thể. Các chỉ số tổng hợp như **F1-score** (trung bình của Precision và Recall) thường được sử dụng để cân bằng và đánh giá toàn diện khi cả hai đều quan trọng.

- F1-score: F1-score là một chỉ số đánh giá mô hình phân loại, đặc biệt hữu ích khi dữ liệu mất cân bằng (tức là một số lớp xuất hiện nhiều hơn hẳn các lớp khác). F1-score cân bằng giữa Precision và Recall, giúp đánh giá mô hình một cách công bằng hơn so với Accuracy. Nếu F1-score có giá trị bằng 1 mô hình hoàn hảo ngược lại nếu bằng 0 thì mô hình hoàn toàn sai. Giá trị F1-Score càng cao thì mô hình càng tốt.

Khi Precision và Recall không cho được các đánh giá mong muốn thì F1-score chính là sự lựa chọn hợp lý, giúp vừa hạn chế dự đoán sai vừa không bỏ sót các trường hợp thực sự đúng, điểm cộng là F1 có thể sử dụng được với các tập dữ liệu bị mất cân bằng tránh việc mô hình chỉ học tốt với các lớp lớn mà bỏ qua các lớp bị hạn chế

Công thức tính:

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

Trong đó:

Precision: Tỷ lệ dự đoán đúng trong số các dự đoán dương tính

Recall: Tỷ lệ dự đoán đúng trong số các mẫu thực tế dương tính.

Khi mô hình phân loại nhiều lớp (như phân loại mức độ béo phì thành 7 mức khác nhau), mỗi lớp sẽ có một F1-score riêng. Macro F1-score là cách tính trung bình F1-score đều nhau cho từng lớp, không quan tâm lớp đó có bao nhiêu mẫu.

Công thức tính:

$$MacroF1 - score = \frac{1}{N}\sum_{i=1}^{N} F1 - score_i$$

Trong đó:

N: Số lượng lớp trong biến mục tiêu

F1-score: F1-score của từng lớp

Đối với tập dữ liệu mất cân bằng khi một mô hình hoạt động tốt trên các lớp chiếm đa số nhưng lại hoạt động kém với các lớp chiếm thiểu số, Weighted F1-score sẽ gán trọng số cao cho F1-score của các lớp đa số và trọng số thấp cho các lớp thiểu số. Điều này phản ánh hiệu suất của mô hình trên các lớp lớn hơn sẽ có ảnh hưởng lớn đến Weighted F1-score cuối cùng

Công thức tính:

$$WeightedF1 - score = \sum_{c=1}^{N} (F1 - score_c \; x \; \frac{|D_c|}{\sum_{i=1}^{N} |D_i|})$$

Trong đó:

N: Số lượng lớp trong biến mục tiêu

F1-score$_c$: Recall của lớp c

$|D_c|$ *: Số lượng mẫu của tập hợp các trường hợp thực tế thuộc lớp c*

$\sum_{i=1}^{N} |D_i|$*: Tổng số lượng mẫu trong toàn bộ tập dữ liệu (tập dữ liệu kiểm thử)*

$\frac{|D_c|}{\sum_{i=1}^{N} |D_i|}$*: Đây chính là trọng số của lớp c, phản ánh tỷ lệ phần trăm số lượng mẫu của lớp đó so với tổng số mẫu trong tập dữ liệu.*

- Ma trận nhầm lẫn (Confusion matrix) là một bảng tóm tắt kết quả dự đoán của mô hình phân loại, giúp bạn thấy rõ mô hình **dự đoán đúng và sai như thế nào** với từng lớp là công cụ cực kỳ quan trọng để phân tích hiệu suất mô hình phân loại, đặc biệt là với bài toán phân loại nhiều lớp như dự đoán mức độ béo phì. Có cái nhìn sâu hơn vào hiệu suất từng lớp thay vì chỉ dựa vào mỗi con số accuracy.

Vì **accuracy tổng thể có thể đánh lừa** — nếu lớp "normal" chiếm 80% thì đoán đại cũng được 80% đúng

Ma trận nhầm lẫn **cho thấy rõ mô hình sai ở đâu**, nhầm lẫn giữa những lớp nào. Từ đó giúp mô hình hoạt động hiệu quả hơn. Ma trận cũng chính là tiền đề để có thể tính toán các giá trị Precision hay Recall, …

Cấu trúc cơ bản (2 lớp):

	Dự đoán: Positive	**Dự đoán: Negative**
Thực tế: Positive	True Positive (TP)	False Negative (FN)
Thực tế: Negative	False Positive (FP)	True Negative (TN)

Bảng 3.4 Cấu trúc cơ bản của ma trận nhầm lẫn

Trong đó:

TP (True Positive): Dự đoán đúng dương

FP (False Positive): Dự đoán sai là dương

FN (False Negative): Dự đoán sai là âm

TN (True Negative): Dự đoán đúng âm

Trong ma trận nhầm lẫn:

- *Giá trị **đường chéo chính**: là số lượng dự đoán đúng*
- *Các **giá trị ngoài đường chéo**: là số lần mô hình dự đoán sai (nhầm lớp)*

Chương IV

Kết quả và thảo luận

1. Kết quả huấn luyện và kiểm tra mô hình

Sau khi đã thực hiện huấn luyện tập dữ liệu trên mô các mô hình khác nhau ta đã có những đánh giá sau theo từng mô hình

Mô hình Logistic Regression

Lớp	**Precision**	**Recall**	**F1-Score**	**Support**
Thiếu cân	0.86	0.86	0.86	59
Bình thường	0.74	0.79	0.76	61
Béo phì loại 1	0.80	0.73	0.76	70
Béo phì loại 2	0.93	0.89	0.91	64

Béo phì loại 3	0.97	0.98	0.98	60
Thừa cân loại 1	0.62	0.42	0.50	55
Thừa cân loại 2	0.41	0.59	0.48	49
Accuracy			0.76	418
Macro Avg	0.76	0.75	0.75	418
Weighted Avg	0.77	0.76	0.76	418

Bảng 4.1.1 Báo cáo hiệu suất phân loại của mô hình Logistic Regression

Tổng quan mô hình đạt độ chính xác tổng thể là 0.76 tương đương 76% số lượng trường hợp trên tập dữ liệu. Các chỉ số trung bình như macro hay weighted của các đơn vị như precision, recall, f1-score đã được đề cập ở chương trên đạt hiệu suất trong khoảng quanh 0.75 – 0.77

Cụ thể:

Lớp ***Thiếu cân*** gồm 59/418 mẫu có: *Precision (0.86), Recall (0.86), F1-score (0.86)* Đây là một trong những lớp mà Logistic Regression hoạt động khá tốt, với sự cân bằng giữa khả năng nhận diện đúng các mẫu của lớp này và tỷ lệ các dự đoán tích cực thực sự thuộc về lớp này.

Lớp ***Bình thường*** gồm 61/418 mẫu có: *Precision (0.74), Recall (0.79), F1-score (0.76)* Hiệu suất ở mức chấp nhận được. Recall cao hơn precision một chút, cho thấy mô hình có xu hướng nhận diện được nhiều mẫu của lớp bình thường, nhưng đôi khi cũng dự đoán nhầm các mẫu không thuộc lớp bình thường thành lớp bình thường.

Lớp ***Béo phì Loại I*** gồm 70/418 mẫu có: *Precision (0.80), Recall (0.73), F1-score (0.76)* Precision cao hơn recall. Điều này có nghĩa là khi mô hình dự đoán một mẫu là lớp béo phì loại I, khả năng cao là đúng, nhưng mô hình có thể bỏ sót một số mẫu thực sự thuộc lớp béo phì loại I.

Lớp ***Béo phì Loại II*** gồm 64/418 *Precision (0.93), Recall (0.89), F1-score (0.91)*: Đây là lớp mà Logistic Regression hoạt động rất tốt, với precision và recall cao, dẫn đến F1-score cao. Mô hình rất đáng tin cậy khi dự đoán lớp này.

Lớp ***Béo phì Loại III*** gồm 60/418 *Precision (0.97), Recall (0.98), F1-score (0.98)*: Đây là lớp có hiệu suất tốt nhất của mô hình. Precision và recall gần như hoàn hảo, cho thấy mô hình cực kỳ hiệu quả trong việc phân loại các mẫu thuộc lớp này.

Lớp ***Thừa cân Loại I*** gồm 55/418 *Precision (0.62), Recall (0.42), F1-score (0.50)*: Đây là một trong những lớp yếu nhất của mô hình. Đặc biệt, recall rất thấp (0.42) cho thấy mô hình bỏ sót rất nhiều mẫu thực sự thuộc lớp 5. F1-score thấp (0.50) phản ánh hiệu suất kém tổng thể cho lớp này.

Lớp ***Thừa cân Loại II*** gồm 49/418 *Precision (0.41), Recall (0.59), F1-score (0.48)*: Đây là lớp yếu nhất. Precision cực kỳ thấp (0.41) cho thấy khi mô hình dự đoán một mẫu là lớp thừa cân loại II, khả năng cao là sai. Mặc dù recall tốt hơn precision một chút, nhưng F1-score vẫn rất thấp (0.48), chỉ ra rằng mô hình gặp khó khăn đáng kể trong việc phân loại lớp này một cách chính xác.

Đánh giá tổng thể:

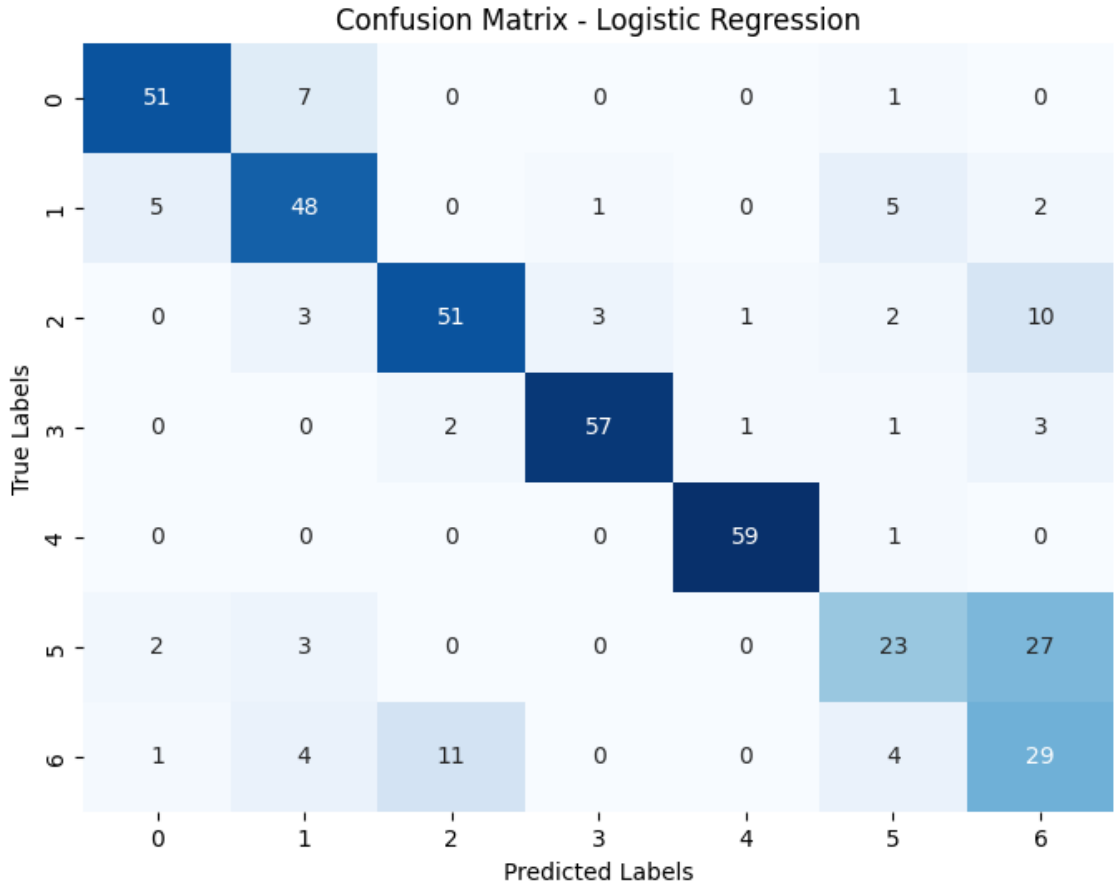
Điểm mạnh:

- Logistic Regression hoạt động rất hiệu quả đối với Lớp ***Béo phì Loại II*** **và** Lớp ***Béo phì Loại III***, với các chỉ số precision, recall và f1-score rất cao.
- Hoạt động khá tốt đối với Lớp ***Thiếu cân*** và ở mức chấp nhận được đối với Lớp ***Bình thường*** **và** Lớp ***Béo phì loại I***.

Điểm yếu:

- Mô hình gặp khó khăn nghiêm trọng với Lớp ***Thừa cân loại I*** **và** Lớp ***Thừa cân loại II***. Precision và recall thấp ở các lớp này kéo F1-score xuống rất thấp. Điều này cho thấy mô hình có thể không học được đủ đặc trưng hoặc có sự nhầm lẫn giữa các lớp này với các lớp khác.
- Accuracy tổng thể ở mức 76% là khá tốt, những việc có những lớp hoạt động kém như lớp ***Thừa cân loại I*** và ***Thừa cân loại II*** cho thấy có tiềm năng cải thiện đáng kể.

Ma trận nhầm lẫn



Hình 7.1: Ma trận nhầm lẫn mô hình Logistic Regression

Mô hình Decision Tree

Lớp	Precision	Recall	F1-Score	Support
Thiếu cân	0.92	0.92	0.92	59
Bình thường	0.86	0.90	0.88	61
Béo phì loại 1	0.97	0.93	0.95	70
Béo phì loại 2	0.98	0.97	0.98	64
Béo phì loại 3	0.97	1.00	0.98	60
Thừa cân loại 1	0.98	0.96	0.97	55
Thừa cân loại 2	0.94	0.94	0.94	49
Accuracy			0.94	418
Macro Avg	0.95	0.95	0.95	418
Weighted Avg	0.92	0.92	0.92	59

Bảng 4.1.2 Báo cáo hiệu suất phân loại của mô hình Decision Tree

Mô hình Cây Quyết định đạt độ chính xác tổng thể là **0.94**, tương đương **94%** số lượng trường hợp trên tập dữ liệu. Các chỉ số trung bình như macro hay weighted của các đơn vị Precision, Recall, F1-score đã được đề cập ở chương trên đạt hiệu suất trong khoảng quanh **0.92 – 0.95**.

Cụ thể:

Lớp ***Thiếu cân*** gồm 59/418 mẫu có: *Precision (0.92), Recall (0.92tiênmf), F1-score (0.92).* Đây là một trong những lớp mà Cây Quyết định hoạt động rất tốt, với sự cân bằng cao giữa khả năng nhận diện đúng các mẫu của lớp này và tỷ lệ các dự đoán tích cực thực sự thuộc về lớp này.

Lớp ***Bình thường*** gồm 61/418 mẫu có: *Precision (0.86), Recall (0.90), F1-score (0.88)*. Hiệu suất ở mức rất tốt. Recall cao hơn Precision một chút, cho thấy mô hình có xu hướng nhận diện được nhiều mẫu của lớp "Bình thường", nhưng đôi khi cũng có thể dự đoán nhầm các mẫu không thuộc lớp "Bình thường" thành lớp "Bình thường".

Lớp ***Béo phì Loại I*** gồm 70/418 mẫu có: *Precision (0.97), Recall (0.93), F1-score (0.95)*. Precision cao hơn Recall. Điều này có nghĩa là khi mô hình dự đoán một

mẫu là lớp "Béo phì Loại I", khả năng cao là đúng, và mô hình vẫn nhận diện được phần lớn các mẫu thực sự thuộc lớp này.

Lớp ***Béo phì Loại II*** gồm 64/418 mẫu có: *Precision (0.98), Recall (0.97), F1-score (0.98).* Đây là lớp mà Cây Quyết định hoạt động rất tốt, với Precision và Recall cao, dẫn đến F1-score cao. Mô hình rất đáng tin cậy khi dự đoán lớp này.

Lớp ***Béo phì Loại III*** gồm 60/418 mẫu có: *Precision (0.97), Recall (1.00), F1-score (0.98)*. Đây là lớp có hiệu suất tốt nhất của mô hình. Precision rất cao và Recall gần như hoàn hảo (1.00), cho thấy mô hình cực kỳ hiệu quả trong việc phân loại tất cả các mẫu thuộc lớp này mà ít mắc lỗi.

Lớp ***Thừa cân Loại I*** gồm 55/418 mẫu có: *Precision (0.98), Recall (0.96), F1-score (0.97)*. Đây là một trong những lớp mà mô hình Cây Quyết định hoạt động rất xuất sắc, khác biệt rõ rệt so với kết quả của mô hình Hồi quy Logistic. Precision và Recall đều rất cao, phản ánh hiệu suất tổng thể vượt trội cho lớp này.

Lớp ***Thừa cân Loại II*** gồm 49/418 mẫu có*: Precision (0.94), Recall (0.94), F1-score (0.94)*. Đây cũng là một lớp mà Cây Quyết định cho thấy hiệu suất rất mạnh, hoàn toàn khác biệt so với mô hình Hồi quy Logistic trước đó. Precision và Recall đều cao, chỉ ra rằng mô hình có khả năng phân loại chính xác đáng kể cho lớp này.

Đánh giá tổng thể:

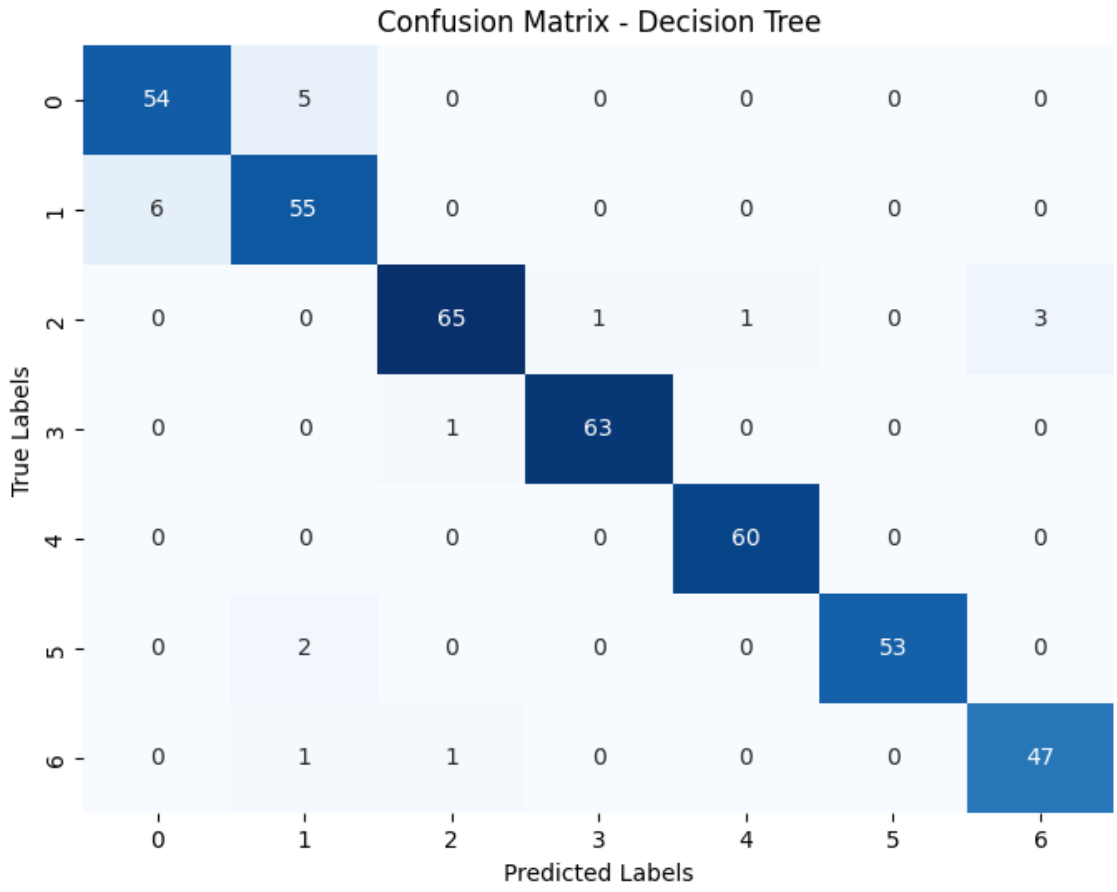
Điểm mạnh:

- **Cây Quyết định hoạt động rất hiệu quả đối với hầu hết các lớp**, đặc biệt là Lớp "Béo phì Loại I, II, III", Lớp "Thiếu cân", và quan trọng nhất là Lớp "Thừa cân Loại I" và "Thừa cân Loại II", với các chỉ số Precision, Recall và F1-score rất cao (hầu hết trên 0.90).
- **Khắc phục đáng kể điểm yếu của mô hình Logistic Regression** trên các lớp "Thừa cân Loại I" và "Thừa cân Loại II", cho thấy khả năng của Cây Quyết định trong việc nắm bắt các mối quan hệ phức tạp và phi tuyến tính trong dữ liệu.
- **Accuracy tổng thể ở mức 94% là rất tốt**, phản ánh hiệu suất phân loại tổng thể vượt trội của mô hình trên tập dữ liệu.

Điểm yếu:

- Mặc dù hiệu suất tổng thể rất cao, lớp "Bình thường" có F1-Score (0.88) thấp hơn một chút so với các lớp khác (hầu hết trên 0.90). Điều này cho thấy vẫn có thể có một số trường hợp nhầm lẫn nhỏ xảy ra trong việc phân loại lớp này so với các lớp khác. Tuy nhiên, đây là một điểm yếu rất nhỏ so với hiệu suất tổng thể.

Ma trận nhầm lẫn:



Hình 7.2: Ma trận nhầm lẫn mô hình Decision Tree

Mô hình Random Forest

Lớp	Precision	Recall	F1-Score	Support
Thiếu cân	1.00	0.88	0.94	59
Bình thường	0.82	0.98	0.90	61
Béo phì loại 1	0.96	0.94	0.95	70
Béo phì loại 2	1.00	0.97	0.98	64
Béo phì loại 3	1.00	1.00	1.00	60
Thừa cân loại 1	0.98	0.89	0.93	55
Thừa cân loại 2	0.88	0.94	0.91	49
Accuracy			0.94	418
Macro Avg	0.95	0.94	0.94	418

Weighted Avg	0.95	0.94	0.95	418

Bảng 4.1.3 Báo cáo hiệu suất phân loại của mô hình Random Forest

Tổng quan mô hình Random Forest đạt độ chính xác tổng thể là **0.94**, tương đương 94% số lượng trường hợp trên tập dữ liệu. Các chỉ số trung bình như macro hay weighted của các đơn vị Precision, Recall, F1-score đạt hiệu suất trong khoảng quanh **0.94 – 0.95**.

Cụ thể:

*Lớp **Thiếu cân*** gồm 59/418 mẫu có: *Precision (1.00), Recall (0.88), F1-score (0.94)*. Đây là một trong những lớp mà Random Forest hoạt động rất tốt, với Precision tuyệt đối cho thấy không có dự đoán sai, nhưng Recall thấp hơn một chút gợi ý mô hình bỏ sót một số trường hợp.

Lớp ***Bình thường*** gồm 61/418 mẫu có: *Precision (0.82), Recall (0.98), F1-score (0.90)*. Hiệu suất ở mức rất tốt. Recall cao hơn Precision đáng kể, cho thấy mô hình có xu hướng nhận diện được hầu hết các mẫu của lớp "Bình thường", nhưng đôi khi cũng có thể dự đoán nhầm các mẫu không thuộc lớp này thành lớp "Bình thường".

Lớp ***Béo phì Loại I*** gồm 70/418 mẫu có: *Precision (0.96), Recall (0.94), F1-score (0.95)*. Precision cao hơn Recall một chút, cho thấy khi mô hình dự đoán một mẫu là lớp "Béo phì Loại I", khả năng cao là đúng, và mô hình vẫn nhận diện được phần lớn các mẫu thực sự thuộc lớp này.

Lớp ***Béo phì Loại II*** gồm 64/418 mẫu có: *Precision (1.00), Recall (0.97), F1-score (0.98)*. Đây là lớp mà Random Forest hoạt động rất tốt, với Precision và Recall rất cao, dẫn đến F1-score cao. Mô hình rất đáng tin cậy khi dự đoán lớp này.

Lớp ***Béo phì Loại III*** gồm 60/418 mẫu có: *Precision (1.00), Recall (1.00), F1-score (1.00)*. Đây là lớp có hiệu suất hoàn hảo của mô hình. Precision và Recall đều đạt 1.00, cho thấy mô hình cực kỳ hiệu quả trong việc phân loại tất cả các mẫu thuộc lớp này mà không mắc lỗi.

Lớp ***Thừa cân Loại I*** gồm 55/418 mẫu có: *Precision (0.98), Recall (0.89), F1-score (0.93)*. Đây là một trong những lớp mà mô hình Random Forest hoạt động rất xuất sắc, với Precision cao và Recall tốt, phản ánh hiệu suất tổng thể rất mạnh cho lớp này.

Lớp ***Thừa cân Loại II*** gồm 49/418 mẫu có: *Precision (0.88), Recall (0.94), F1-score (0.91)*. Đây cũng là một lớp mà Random Forest cho thấy hiệu suất rất mạnh. Recall cao hơn Precision một chút, chỉ ra rằng mô hình có khả năng nhận diện được nhiều mẫu của lớp này, mặc dù có thể có một số dự đoán nhầm lẫn.

Đánh giá tổng thể:

Điểm mạnh:

- Random Forest hoạt động rất hiệu quả đối với hầu hết các lớp, đặc biệt là Lớp "Béo phì Loại I, II, III", Lớp "Thiếu cân", Lớp "Thừa cân Loại I" và Lớp "Thừa

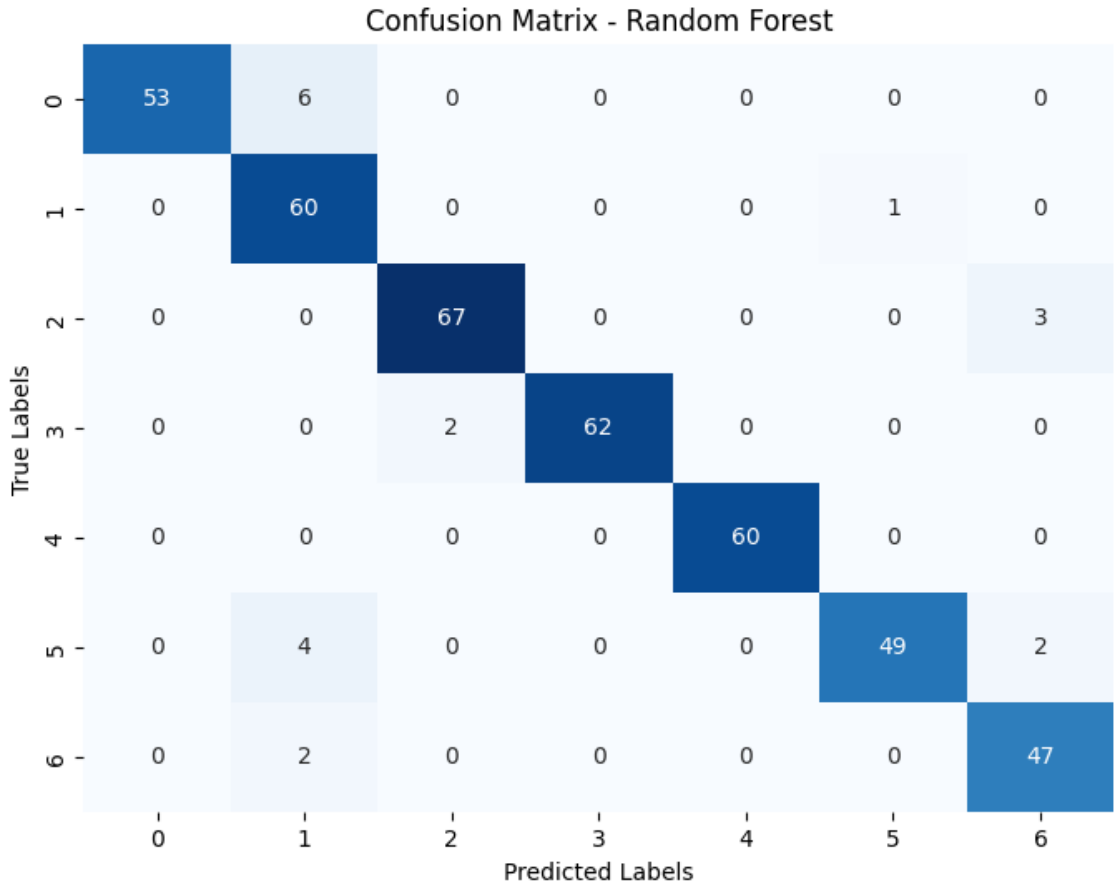
cân Loại II", với các chỉ số Precision, Recall và F1-score rất cao (hầu hết trên 0.90).

- Accuracy tổng thể ở mức 94% là rất tốt, phản ánh hiệu suất phân loại tổng thể vượt trội của mô hình trên tập dữ liệu.
- Đạt Precision tuyệt đối (1.00) trên một số lớp, cho thấy độ tin cậy cực cao của các dự đoán dương tính cho những lớp này.

Điểm yếu:

- Mặc dù hiệu suất tổng thể rất cao, lớp "Thiếu cân" có Recall thấp hơn Precision (0.88 so với 1.00), và lớp "Bình thường" có Precision thấp hơn Recall (0.82 so với 0.98). Điều này cho thấy có sự đánh đổi giữa việc giảm thiểu False Negatives và False Positives tùy theo lớp, mặc dù F1-score vẫn ở mức cao.

Ma trận nhầm lẫn:



Hình 7.3: Ma trận nhầm lẫn mô hình Random Forest

Mô hình K-Nearest Neighbors (KNN)

Lớp	Precision	Recall	F1-Score	Support
Thiếu cân	0.94	0.83	0.88	59
Bình thường	0.57	0.85	0.68	61

Béo phì loại 1	0.98	0.83	0.90	70
Béo phì loại 2	1.00	0.94	0.97	64
Béo phì loại 3	1.00	1.00	1.00	60
Thừa cân loại 1	0.94	0.84	0.88	55
Thừa cân loại 2	0.85	0.82	0.83	49
Accuracy			0.87	418
Macro Avg	0.90	0.87	0.88	418
Weighted Avg	0.90	0.87	0.88	418

Bảng 4.1.4 Báo cáo hiệu suất phân loại của mô hình KNN

Tổng quan mô hình K-Nearest Neighbors (KNN) đạt độ chính xác tổng thể là **0.87**, tương đương 87% số lượng trường hợp trên tập dữ liệu. Các chỉ số trung bình như macro hay weighted của các đơn vị Precision, Recall, F1-score đạt hiệu suất trong khoảng quanh **0.87 – 0.90**.

Cụ thể:

Lớp ***Thiếu cân*** gồm 59/418 mẫu có**:** *Precision (0.94), Recall (0.83), F1-score (0.88).* Đây là một trong những lớp mà KNN hoạt động khá tốt, với Precision cao nhưng Recall thấp hơn một chút, cho thấy khả năng nhận diện đúng các mẫu của lớp này tốt nhưng có thể bỏ sót một số trường hợp.

Lớp ***Bình thường*** gồm 61/418 mẫu có: *Precision (0.57), Recall (0.85), F1-score (0.68)*. Hiệu suất của lớp này ở mức yếu. Precision rất thấp (0.57) chỉ ra rằng mô hình thường dự đoán nhầm các mẫu không thuộc lớp "Bình thường" thành lớp này, mặc dù Recall (0.85) khá hơn, F1-score thấp (0.68) phản ánh hiệu suất tổng thể kém cho lớp này.

Lớp ***Béo phì Loại I*** gồm 70/418 mẫu có: *Precision (0.98), Recall (0.83), F1-score (0.90)*. Precision rất cao, cho thấy khi mô hình dự đoán một mẫu là lớp "Béo phì Loại I", khả năng cao là đúng, nhưng Recall thấp hơn một chút gợi ý mô hình có thể bỏ sót một số mẫu thực sự thuộc lớp này.

Lớp ***Béo phì Loại II*** gồm 64/418 mẫu có: *Precision (1.00), Recall (0.94), F1-score (0.97)*. Đây là lớp mà KNN hoạt động rất tốt, với Precision và Recall cao, dẫn đến F1-score cao. Mô hình rất đáng tin cậy khi dự đoán lớp này.

Lớp ***Béo phì Loại III*** gồm 60/418 mẫu có: *Precision (1.00), Recall (1.00), F1-score (1.00)*. Đây là lớp có hiệu suất tốt nhất của mô hình. Precision và Recall gần như hoàn hảo, cho thấy mô hình cực kỳ hiệu quả trong việc phân loại các mẫu thuộc lớp này.

Lớp ***Thừa cân Loại I*** gồm 55/418 mẫu có: *Precision (0.94), Recall (0.84), F1-score (0.88)*. Hiệu suất khá tốt, với Precision cao hơn Recall, cho thấy độ tin cậy của các dự đoán là cao nhưng có thể bỏ sót một số trường hợp.

Lớp ***Thừa cân Loại II*** gồm 49/418 mẫu có: *Precision (0.85), Recall (0.82), F1-score (0.83)*. Hiệu suất ở mức chấp nhận được, với Precision và Recall khá cân bằng, cho thấy mô hình có khả năng phân loại tương đối tốt cho lớp này.

Đánh giá tổng thể:

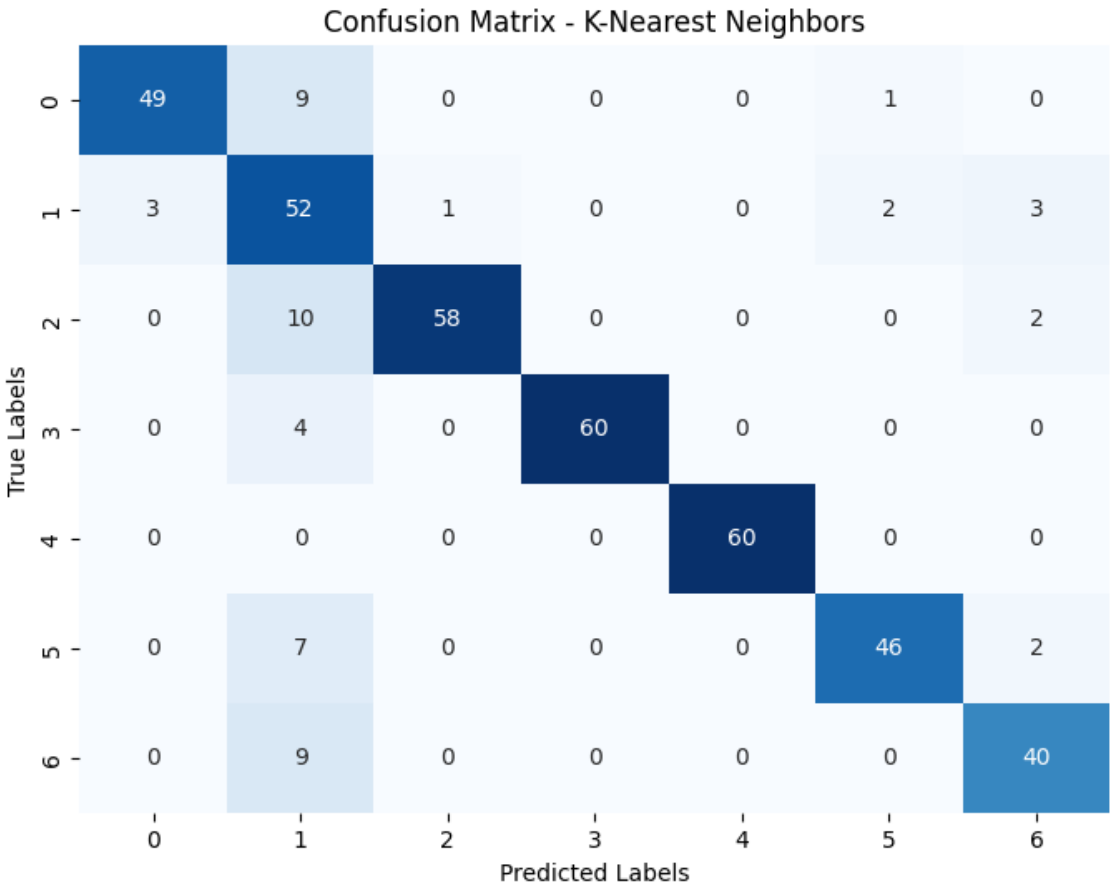
Điểm mạnh:

- KNN hoạt động rất hiệu quả đối với Lớp "Béo phì Loại II" và Lớp "Béo phì Loại III", với các chỉ số Precision, Recall và F1-score rất cao (Precision 1.00).
- Hoạt động khá tốt đối với Lớp "Thiếu cân", "Béo phì Loại I", "Thừa cân Loại I" và "Thừa cân Loại II".
- Accuracy tổng thể ở mức 87% là khá tốt, phản ánh hiệu suất phân loại tổng thể chấp nhận được của mô hình trên tập dữ liệu.

Điểm yếu:

- Mô hình gặp khó khăn nghiêm trọng với Lớp "Bình thường". Precision rất thấp (0.57) và F1-score thấp (0.68) ở lớp này. Điều này cho thấy mô hình có thể không học được đủ đặc trưng hoặc có sự nhầm lẫn đáng kể giữa lớp này với các lớp khác.

Ma trận nhầm lẫn:



Hình 7.4: Ma trận nhầm lẫn mô hình K-Nearest Neighbors (KNN)

Mô hình Support Vector Machine (SVM)

Lớp	Precision	Recall	F1-Score	Support
Thiếu cân	0.77	0.81	0.79	59
Bình thường	0.55	0.87	0.67	61
Béo phì loại 1	0.97	0.81	0.88	70
Béo phì loại 2	1.00	0.91	0.95	64
Béo phì loại 3	1.00	0.98	0.99	60
Thừa cân loại 1	1.00	0.78	0.88	55
Thừa cân loại 2	0.97	0.80	0.88	49
Accuracy			0.85	418
Macro Avg	0.89	0.85	0.86	418

Weighted Avg	0.89	0.85	0.86	418

Bảng 4.1.5 Báo cáo hiệu suất phân loại của mô hình SVM

Mô hình Support Vector Machine (SVM) đạt độ chính xác tổng thể là **0.85**, tương đương **85%** số lượng trường hợp trên tập dữ liệu. Các chỉ số trung bình như macro hay weighted của các đơn vị Precision, Recall, F1-score đã được đề cập ở chương trên đạt hiệu suất trong khoảng quanh **0.85 – 0.89**.

Cụ thể:

Lớp ***Thiếu cân*** gồm 59/418 mẫu có: *Precision (0.77), Recall (0.81), F1-score (0.79)*. Hiệu suất của lớp này ở mức chấp nhận được. Precision và Recall khá cân bằng, cho thấy mô hình có khả năng nhận diện một tỷ lệ hợp lý các mẫu của lớp này và độ tin cậy của các dự đoán dương tính là tương đối tốt.

Lớp ***Bình thường*** gồm 61/418 mẫu có: *Precision (0.55), Recall (0.87), F1-score (0.67)*. Đây là lớp có hiệu suất yếu nhất của mô hình. Precision rất thấp (0.55) là một điểm đáng lo ngại, chỉ ra rằng khi mô hình dự đoán một mẫu là lớp "Bình thường", khả năng cao là dự đoán đó sai. Mặc dù Recall (0.87) khá hơn, cho thấy mô hình nhận diện được phần lớn các mẫu thực sự "Bình thường", F1-score rất thấp (0.67) phản ánh hiệu suất tổng thể kém cho lớp này, cho thấy sự khó khăn đáng kể trong việc phân loại chính xác lớp "Bình thường".

Lớp ***Béo phì Loại I*** gồm 70/418 mẫu có: *Precision (0.97), Recall (0.81), F1-score (0.88). Precision rất cao (0.97)* cho thấy độ tin cậy của các dự đoán là "Béo phì Loại I" là xuất sắc. Tuy nhiên, Recall tương đối thấp hơn (0.81) cho thấy mô hình có thể bỏ sót một số trường hợp thực sự "Béo phì Loại I".

Lớp ***Béo phì Loại II*** gồm 64/418 mẫu có: *Precision (1.00), Recall (0.91), F1-score (0.95)*. Đây là lớp mà SVM hoạt động rất tốt, với Precision tuyệt đối (1.00) và Recall cao, dẫn đến F1-score cao. Mô hình rất đáng tin cậy khi dự đoán lớp này.

Lớp ***Béo phì Loại III*** gồm 60/418 mẫu có: *Precision (1.00), Recall (0.98), F1-score (0.99)*. Đây là lớp có hiệu suất gần như hoàn hảo của mô hình. Precision tuyệt đối (1.00) và Recall rất cao (0.98), cho thấy mô hình cực kỳ hiệu quả trong việc phân loại các mẫu thuộc lớp này.

Lớp ***Thừa cân Loại I*** gồm 55/418 mẫu có: *Precision (1.00), Recall (0.78), F1-score (0.88)*. Precision tuyệt đối (1.00) là một điểm rất mạnh, cho thấy tất cả các mẫu được dự đoán là "Thừa cân Loại I" đều đúng. Tuy nhiên, Recall tương đối thấp (0.78) là một điểm yếu, cho thấy mô hình bỏ sót một phần đáng kể các trường hợp thực sự "Thừa cân Loại I".

Lớp ***Thừa cân Loại II*** gồm 49/418 mẫu có: *Precision (0.97), Recall (0.80), F1-score (0.88)*. Precision rất cao (0.97) cho thấy độ tin cậy của các dự đoán là "Thừa cân Loại II" là xuất sắc. Tuy nhiên, Recall tương đối thấp (0.80) gợi ý rằng mô hình bỏ sót một số trường hợp thực sự thuộc lớp này.

Đánh giá tổng thể:

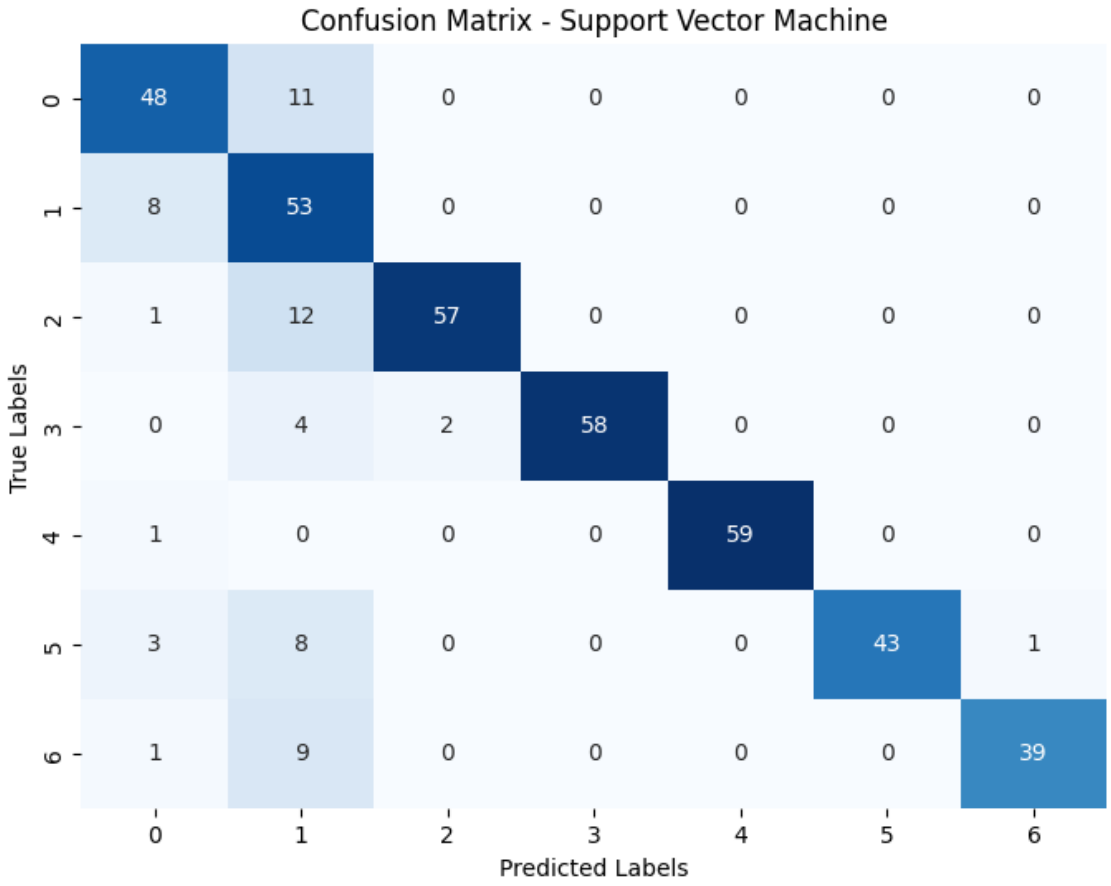
Điểm mạnh:

- SVM hoạt động rất hiệu quả đối với Lớp "Béo phì Loại II" và Lớp "Béo phì Loại III", với các chỉ số Precision, Recall và F1-score rất cao (Precision 1.00).
- Hoạt động tốt đối với Lớp "Béo phì Loại I", Lớp "Thừa cân Loại I" và Lớp "Thừa cân Loại II", đặc biệt nổi bật với Precision rất cao (trên 0.97).
- Accuracy tổng thể ở mức 85% là khá tốt, phản ánh hiệu suất phân loại tổng thể chấp nhận được của mô hình trên tập dữ liệu.

Điểm yếu:

- Mô hình gặp khó khăn nghiêm trọng với Lớp "Bình thường". Precision rất thấp (0.55) và F1-score thấp (0.67) ở lớp này. Điều này cho thấy mô hình có thể không học được đủ đặc trưng hoặc có sự nhầm lẫn đáng kể giữa lớp này với các lớp khác.
- Recall tương đối thấp trên một số lớp, bao gồm "Thiếu cân" (0.81), "Béo phì Loại I" (0.81), "Thừa cân Loại I" (0.78) và "Thừa cân Loại II" (0.80). Mặc dù Precision cao, việc bỏ sót các trường hợp thực sự (False Negatives) ở những lớp này là một điểm cần cân nhắc.

Ma trận nhầm lẫn:



Hình 7.5: Ma trận nhầm lẫn mô hình Support Vector Machine (SVM)

2. So sánh hiệu năng giữa các mô hình

Sau khi đã thực hiện huấn luyện các mô hình học máy trên tập dữ liệu ta có thể dựa vào các đơn vị cũng như các chỉ số đánh giá nhằm có cái nhìn tổng quát so sánh hiệu năng giữa chúng

Trong bài toán này, năm mô hình học máy khác nhau—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), và Support Vector Machine (SVM)—đã được đánh giá trên cùng một tập dữ liệu phân loại trạng thái cân nặng. Bảng tổng hợp dưới đây sẽ so sánh hiệu suất tổng thể và chi tiết theo từng lớp của các mô hình.

Mô hình	Accuracy	Macro Avg	Weighted Avg	Lớp yếu nhất	Lớp tốt nhất
Logistic Regression	0.76	0.77	0.77	Thừa cân Loại II (0.48)	Béo phì Loại III (0.98)
KNN	0.87	0.88	0.88	Bình thường (0.68)	Béo phì Loại III (1.00)
SVM	0.85	0.86	0.86	Bình thường (0.67)	Béo phì Loại III (0.99)
Decision Tree	0.94	0.95	0.92	Bình thường (0.88)	Béo phì Loại III (0.98)
Random Forest	0.94	0.94	0.95	Bình thường (0.90)	Béo phì Loại III (1.00)

Bảng 4.2 Tóm tắt hiệu suất của các mô hình được sử dụng

Hiệu suất Tổng thể (Accuracy và F1-Score trung bình):

- **Decision Tree** và **Random Forest** thể hiện hiệu suất vượt trội với Accuracy đều đạt 0.94. Đặc biệt, Random Forest có Weighted Avg F1-Score là 0.95, cao hơn Decision Tree (0.92) và các mô hình khác. Điều này cho thấy các mô hình dựa trên cây (tree-based models) và phương pháp học tập tổ hợp (ensemble learning) có khả năng tổng quát hóa tốt nhất trên tập dữ liệu này.

- **KNN** đứng thứ hai với **Accuracy** 0.87 và F1-Score trung bình 0.88.

- **SVM** đạt Accuracy 0.85 và **F1-Score** trung bình 0.86.

- **Logistic Regression** cho thấy hiệu suất thấp nhất với Accuracy 0.76 và F1-Score trung bình khoảng 0.75-0.77, thiết lập một baseline khởi điểm thấp hơn đáng kể so với các mô hình phi tuyến tính.

Khả năng xử lý các Lớp riêng lẻ:

- Lớp ***Béo phì Loại III***: Tất cả các mô hình ngoại trừ Logistic Regression đều cho thấy hiệu suất gần như hoàn hảo hoặc hoàn hảo (F1-Score 0.98-1.00). Điều này cho thấy lớp này có đặc trưng rất rõ ràng và dễ phân tách.

- Lớp ***Bình thường***: Đây là lớp thách thức nhất đối với hầu hết các mô hình.

  Logistic Regression, KNN và SVM đều gặp khó khăn đáng kể với lớp này, thể hiện qua F1-Score tương đối thấp (0.67 - 0.68). Đặc biệt, Precision của KNN (0.57) và SVM (0.55) trên lớp này rất thấp, chỉ ra tỷ lệ dương tính giả cao.

  Ngược lại, Decision Tree (F1-Score 0.88) và Random Forest (F1-Score 0.90) thể hiện khả năng phân loại vượt trội cho lớp "Bình thường", với Random Forest đạt hiệu suất tốt nhất. Điều này củng cố ưu điểm của các mô hình dựa trên cây trong việc nắm bắt các ranh giới phức tạp.

- Lớp ***Thừa cân Loại I*** và ***Thừa cân Loại II***:

  Logistic Regression đã gặp khó khăn nghiêm trọng ở các lớp này, đặc biệt là "Thừa cân Loại II" với F1-Score chỉ 0.48.

  Decision Tree (F1-Score 0.97 và 0.94) và Random Forest (F1-Score 0.93 và 0.91) đã khắc phục điểm yếu này một cách ngoạn mục, đạt hiệu suất rất cao.

  SVM cũng thể hiện Precision cao (1.00 và 0.97) trên các lớp này, nhưng Recall thấp hơn (0.78 và 0.80) cho thấy mô hình bỏ sót một số trường hợp.

  KNN cũng hoạt động khá tốt trên hai lớp này (F1-Score 0.88 và 0.83).

Sự ổn định và Khả năng tổng quát hóa:

Random Forest, với bản chất là một mô hình ensemble, cho thấy sự cân bằng tốt giữa Precision và Recall trên hầu hết các lớp, đồng thời giảm thiểu rủi ro quá khớp so với một Decision Tree đơn lẻ.

Decision Tree mặc dù đạt Accuracy cao, nhưng đôi khi có xu hướng quá khớp với dữ liệu huấn luyện nếu không được điều chỉnh sâu hợp lý. Tuy nhiên, trong trường hợp này, hiệu suất trên tập kiểm tra là rất mạnh.

KNN và SVM mặc dù có thể đạt hiệu suất tốt trên các lớp được phân tách rõ ràng, nhưng nhạy cảm hơn với các đặc trưng và sự chồng lấn giữa các lớp, đặc biệt là lớp "Bình thường".

Logistic Regression, là một mô hình tuyến tính, rõ ràng gặp khó khăn trong việc phân tách các ranh giới phi tuyến tính và phức tạp của dữ liệu.

Kết luận so sánh:

Random Forest và Decision Tree là hai mô hình có hiệu suất vượt trội nhất cho bài toán phân loại trạng thái cân nặng này, với Accuracy cao và F1-Score xuất sắc trên hầu hết các lớp, đặc biệt là khả năng xử lý tốt các lớp "Thừa cân" và "Bình thường" mà Logistic Regression gặp vấn đề.

KNN và SVM là những lựa chọn khá tốt nhưng vẫn còn điểm yếu đáng kể ở lớp "Bình thường", gợi ý rằng các ranh giới phân loại cho lớp này phức tạp hơn so với khả năng của chúng trong cấu hình hiện tại.

Logistic Regression chỉ nên được xem xét là một mô hình baseline ban đầu do hiệu suất thấp hơn đáng kể so với các thuật toán phi tuyến và ensemble.

3. Phân tích vai trò của đặc trưng đầu vào

3.1 Phương pháp lựa chọn đặc trưng quan trọng

Phương pháp chọn lọc đặc trưng Chi-squared (χ^2) là một kỹ thuật thống kê được sử dụng phổ biến trong học máy để đánh giá mối quan hệ giữa các đặc trưng phân loại (categorical features) và biến mục tiêu phân loại (categorical target variable). Mục tiêu của phương pháp này là xác định những đặc trưng có ý nghĩa thống kê nhất trong việc phân biệt các lớp của biến mục tiêu, từ đó giúp giảm chiều dữ liệu và cải thiện hiệu suất của mô hình học máy.

Nguyên lý hoạt động:

Phương pháp χ^2 dựa trên kiểm định Chi-square độc lập (Chi-squared test of independence). Nguyên lý cơ bản là kiểm tra giả thuyết vô hiệu (null hypothesis - H0) rằng không có mối quan hệ (tức là độc lập) giữa đặc trưng và biến mục tiêu. Giả thuyết đối (alternative hypothesis - H1) là có mối quan hệ (tức là phụ thuộc) giữa chúng.

Độc lập (Independence): *Nếu một đặc trưng độc lập với biến mục tiêu, thì sự phân bố của các giá trị của đặc trưng đó sẽ gần như giống nhau trên tất cả các lớp của biến mục tiêu.*

Phụ thuộc (Dependence): *Nếu một đặc trưng phụ thuộc vào biến mục tiêu, thì sự phân bố của các giá trị của đặc trưng đó sẽ khác biệt đáng kể giữa các lớp của biến mục tiêu.*

Giá trị χ^2 càng lớn, xác suất để bác bỏ giả thuyết vô hiệu càng cao, điều này ngụ ý rằng có mối quan hệ phụ thuộc mạnh mẽ hơn giữa đặc trưng và biến mục tiêu. Do đó, các đặc trưng có giá trị χ^2 cao sẽ được chọn vì chúng mang nhiều thông tin hữu ích cho việc phân loại.

*Công thức tính**:***

Đối với một đặc trưng phân loại và một biến mục tiêu phân loại, giá trị χ2 được tính toán từ bảng tần suất quan sát (observed frequencies) và bảng tần suất kỳ vọng (expected frequencies).

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Trong đó:

R: Số lượng hàng trong bảng tần suất (số lượng giá trị độc nhất của đặc trưng).

C: Số lượng cột trong bảng tần suất (số lượng lớp độc nhất của biến mục tiêu).

Oij: Tần suất quan sát (observed frequency) của các mẫu thuộc hàng i và cột j.

Eij: Tần suất kỳ vọng (expected frequency) của các mẫu thuộc hàng i và cột j, được tính như sau: Eij=Tổng số mẫu = (Tổng hàng i)×(Tổng cột j) Tần suất kỳ vọng Eij

đại diện cho số lượng mẫu dự kiến trong ô (i,j) nêu đặc trưng và biến mục tiêu hoàn toàn độc lập.

Cách áp dụng trong chọn lọc đặc trưng:

- **Tính toán giá trị χ2:** Đối với mỗi đặc trưng phân loại, tính toán giá trị χ2 của nó với biến mục tiêu.
- **Xếp hạng đặc trưng:** Sắp xếp các đặc trưng theo thứ tự giảm dần của giá trị χ2. Đặc trưng có giá trị χ2 cao nhất là đặc trưng có mối quan hệ phụ thuộc mạnh nhất với biến mục tiêu.
- **Chọn lọc:** Lựa chọn k đặc trưng hàng đầu có giá trị χ2 cao nhất, hoặc lựa chọn các đặc trưng có giá trị χ2 lớn hơn một ngưỡng nhất định. Một cách khác là sử dụng p-value tương ứng với giá trị χ2 và bậc tự do (degrees of freedom) để bác bỏ giả thuyết vô hiệu ở một mức ý nghĩa nhất định (ví dụ: α=0.05). Các đặc trưng có p-value nhỏ hơn α được coi là có mối quan hệ thống kê đáng kể với biến mục tiêu.

Ưu điểm:

- **Dễ diễn giải:** Giá trị χ2 cung cấp một thước đo trực tiếp về mức độ phụ thuộc giữa hai biến phân loại.
- **Không cần giả định phân phối:** Không yêu cầu giả định về phân phối của dữ liệu.
- **Hiệu quả cho dữ liệu phân loại:** Đặc biệt phù hợp và hiệu quả khi làm việc với các đặc trưng và biến mục tiêu dạng phân loại.

Hạn chế:

- **Chỉ áp dụng cho đặc trưng phân loại:** Không thể trực tiếp áp dụng cho các đặc trưng liên tục (numerical features). Đối với đặc trưng liên tục, cần phải thực hiện rời rạc hóa (discretization) trước khi áp dụng χ2.
- **Nhạy cảm với tần suất thấp:** Nếu có các ô trong bảng tần suất mà tần suất kỳ vọng (Eij) quá nhỏ (thường dưới 5), giá trị χ2 có thể trở nên không đáng tin cậy. Trong trường hợp này, có thể cần kết hợp các nhóm hoặc sử dụng các kiểm định thay thế (ví dụ: Kiểm định Fisher).
- **Không tính đến tương tác đa biến:** χ2 đánh giá mối quan hệ của từng đặc trưng độc lập với biến mục tiêu, không xem xét các tương tác phức tạp giữa nhiều đặc trưng với nhau.

Theo phương pháp chi2 (chi-square) thì top 10 đặc trưng quan trọng ứng với từng số điểm là

Đặc trưng	Điểm
weight	436.59
gender	326.05

monitors_calories_daily	122.14
food_between_meal	113.21
number_of_main_meal	110.43
family_history_with_overweight	108.03
time_using_technology	105.37
age	91.84
transportation	42.82
freq_of_physical_activity	39.59

Bảng 4.3: Bảng so sánh tầm quan trọng của đặc trưng

Bên cạnh phương pháp chọn lọc đặc trưng chi2 thì trong thư viện sklearn còn hỗ trợ thêm *importance()* là hàm được thiết kế nhằm tính toán độ quan trọng của các đặc trưng dựa trên mô hình đã cho từ đó giúp ta chọn lọc đánh giá, sắp xếp cách mà các mô hình xem trọng đặc trưng nào là quan trọng để có thể đánh giá cũng như đưa ra dự đoán. Dưới đây là biểu đồ biểu thị mức độ quan trọng của các đặc trưng theo từng mô hình

Trong bài tiểu luận này chỉ có hai mô hình là Decision Tree và Random Forest là có hỗ trợ hàm *importance*(). Tuy nhiên cũng có thể sử dụng các phương pháp gián tiếp nhằm quy đổi ước tính ra các điểm số nhằm xác định đặc trưng chẳng hạn như sử dụng mô hình GradientBoostingClassifier, nhưng tầm quan trọng được hiển thị sẽ là của mô hình GradientBoostingClassifier chứ không còn là của mô hình ban đầu. Điều này dẫn tới cả ba mô hình còn lại Logistic Regression, SVM, KNN đều có chung một biểu đồ mức độ quan trọng

Phân tích đặc trưng của 3 mô hình đầu (sử dụng GradientBoostingClassifier)

Như đã đề cập ở phần trên do cả ba mô hình này đều không hỗ trợ hàm *importance().* Nên một cách thay thế là sử dụng mô hình GradientBoostingClassifier để có thể thực hiện được việc chọn lọc đặc trưng kết quả ta có 3 biểu đồ bên dưới (Hình 8.1, Hình 8.2, Hình 8.3)

Đặc trưng quan trọng

- weight (Cân nặng): Chiếm tầm quan trọng vượt trội so với tất cả các đặc trưng khác, với giá trị khoảng 0.6. Điều này là hoàn toàn hợp lý và có ý nghĩa lâm sàng/sinh học sâu sắc. Cân nặng là một yếu tố trực tiếp và cốt lõi trong việc xác định mức độ béo phì. Bất kỳ mô hình nào dự đoán béo phì đều kỳ vọng cân nặng sẽ là đặc trưng quan trọng nhất.

- height (Chiều cao): Có tầm quan trọng đáng kể tiếp theo, khoảng 0.1. Tương tự như cân nặng, chiều cao là một thành phần trực tiếp của chỉ số BMI. Mặc dù cân

nặng đóng góp nhiều hơn, chiều cao vẫn là một yếu tố quan trọng trong việc định lượng tỷ lệ cơ thể.

- freq_of_vegetable (Tần suất ăn rau): Đặc trưng này có tầm quan trọng đáng chú ý, khoảng 0.09. Điều này gợi ý rằng thói quen ăn uống lành mạnh, đặc biệt là việc tiêu thụ rau thường xuyên, có ảnh hưởng đáng kể đến mức độ béo phì. Đây là một kết quả hợp lý trong bối cảnh dinh dưỡng và sức khỏe cộng đồng.

- gender (Giới tính): Cũng thể hiện tầm quan trọng gần với freq_of_vegetable, khoảng 0.085. Giới tính có thể ảnh hưởng đến các yếu tố như tỷ lệ mỡ cơ thể, chuyển hóa, thói quen sinh hoạt, hoặc thậm chí là các yếu tố văn hóa-xã hội liên quan đến béo phì.

- daily_water (Lượng nước uống hàng ngày): Có tầm quan trọng tương đối nhỏ hơn nhưng vẫn đáng chú ý, khoảng 0.04. Uống đủ nước có thể liên quan đến quá trình trao đổi chất và cảm giác no, gián tiếp ảnh hưởng đến cân nặng.

- age (Tuổi): Có tầm quan trọng khoảng 0.02. Tuổi tác là một yếu tố sinh học quan trọng ảnh hưởng đến chuyển hóa và hoạt động thể chất, do đó có mối liên hệ với béo phì.

- freq_of_alcohol (Tần suất uống rượu): Có tầm quan trọng tương tự như age, khoảng 0.02. Lượng calo từ rượu và ảnh hưởng của nó đến chuyển hóa có thể góp phần vào việc tăng cân.

- number_of_main_meal (Số bữa ăn chính): Có tầm quan trọng nhỏ, khoảng 0.015. Điều này có thể cho thấy số lượng bữa ăn chính ít quan trọng hơn so với chất lượng và tần suất các loại thực phẩm cụ thể.

Đặc trưng ít quan trọng:

Các đặc trưng ở phía bên trái của biểu đồ, với tầm quan trọng rất gần 0 (dưới 0.01), bao gồm:

- **smoking (Hút thuốc)**
- **family_history_with_overweight (Tiền sử gia đình béo phì)**
- **monitors_calories_daily (Theo dõi lượng calo hàng ngày)**
- **transportation (Phương tiện đi lại)**
- **time_using_technology (Thời gian sử dụng công nghệ)**
- **freq_high_calories_food (Tần suất ăn thức ăn nhiều calo)**
- **freq_of_physical_activity (Tần suất hoạt động thể chất)**
- **food_between_meal (Ăn vặt giữa các bữa)**

Việc các đặc trưng này có tầm quan trọng thấp có thể chỉ ra một số điều:

Một số đặc trưng được kỳ vọng có ảnh hưởng đến béo phì như tần suất tiêu thụ thực phẩm nhiều năng lượng (*freq_high_calories_food*) hay mức độ vận động thể chất (*freq_of_physical_activity*) lại thể hiện mức độ quan trọng thấp trong mô hình. Hiện tượng này có thể được lý giải qua một số khía cạnh.

Thứ nhất, ảnh hưởng của các đặc trưng này có thể mang tính gián tiếp hoặc phi tuyến, tức là mối quan hệ giữa chúng với mức độ béo phì không phải lúc nào cũng đơn giản và tuyến tính. Chúng có thể tương tác với các đặc trưng khác hoặc được biểu hiện thông qua các đặc trưng trung gian như cân nặng (*weight*) và chiều cao (*height*), vốn là các chỉ số đã phản ánh hệ quả của lối sống.

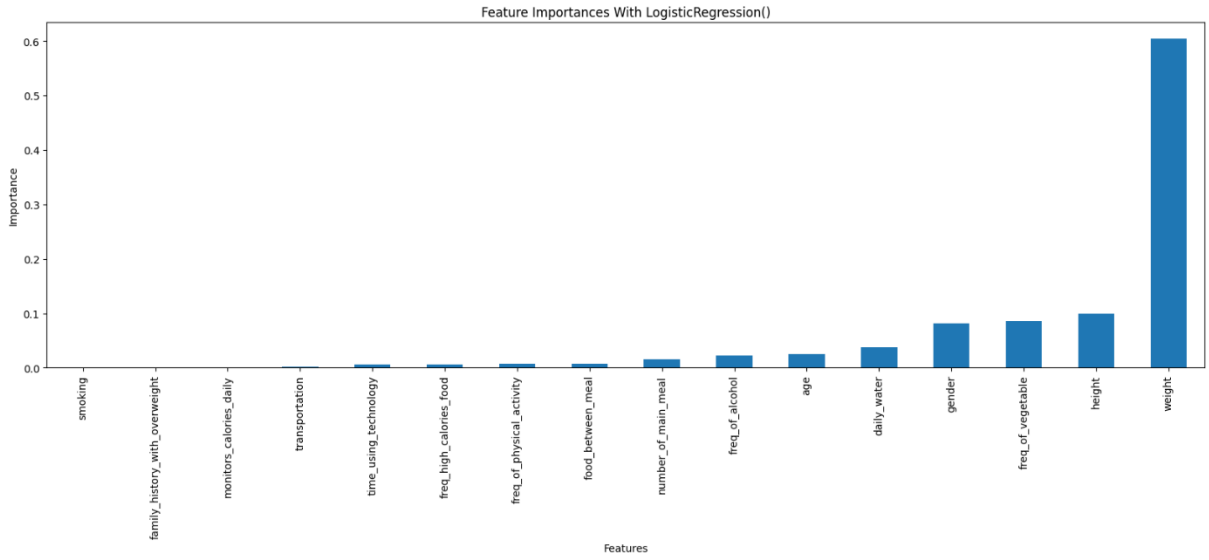
Thứ hai, dữ liệu thu thập có thể thiếu tính đa dạng, dẫn đến việc mô hình không nhận diện được đầy đủ tầm quan trọng của một số đặc trưng. Khi các nhóm dữ liệu đặc trưng cho thói quen sống không được đại diện đầy đủ, mô hình sẽ thiên lệch trong việc đánh giá vai trò của chúng.

Thứ ba, mối tương quan cao giữa các đặc trưng cũng là một yếu tố cần cân nhắc. Khi hai hoặc nhiều đặc trưng có mối quan hệ chặt chẽ, mô hình có xu hướng ưu tiên một trong số đó và giảm vai trò của các đặc trưng còn lại. Ví dụ, nếu *freq_high_calories_food* có tương quan cao với *weight*, mô hình có thể ưu tiên sử dụng *weight* như một đại diện hiệu quả hơn.

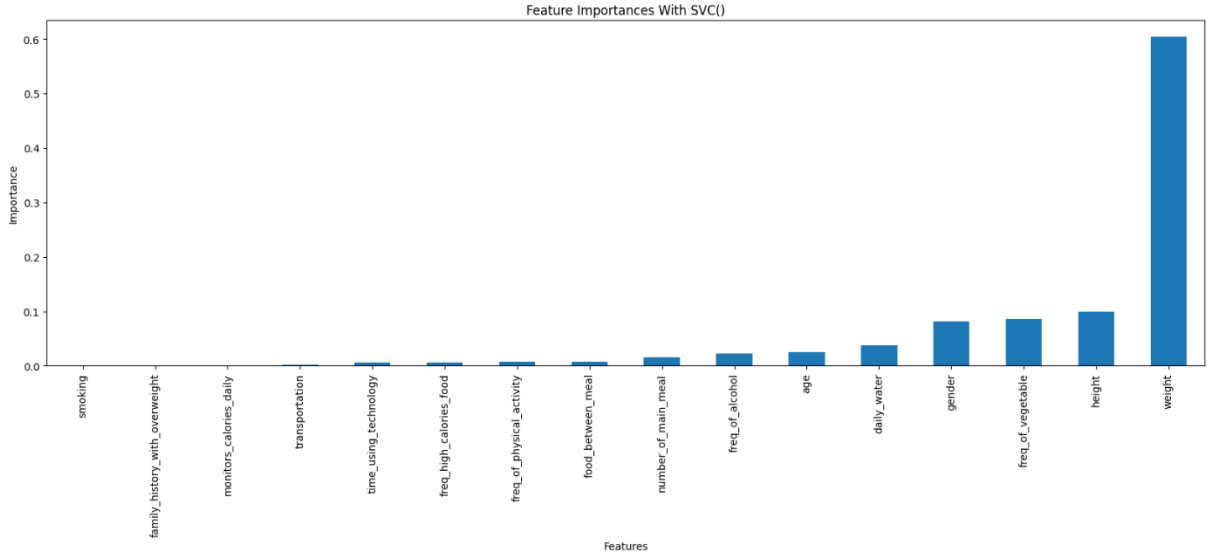
Thứ tư, việc sử dụng một mô hình học máy cụ thể làm proxy cũng có thể dẫn đến sai lệch trong việc đánh giá tầm quan trọng của các đặc trưng. Chẳng hạn, nếu sử dụng Gradient Boosting Classifier, mô hình có thể bỏ qua các đặc trưng có ảnh hưởng nhỏ nhưng ổn định mà các mô hình tuyến tính như Logistic Regression có khả năng ghi nhận.

Mặc dù vậy, kết quả phân tích cũng đồng thời xác nhận lại tầm quan trọng nổi bật của *weight* và *height*—các đặc trưng sinh trắc học có liên hệ trực tiếp đến tình trạng thể trọng. Điều này phù hợp với các kiến thức y học hiện đại, đặc biệt là trong việc sử dụng chỉ số BMI để phân loại mức độ béo phì. Ngoài ra, các đặc trưng liên quan đến thói quen sống lành mạnh như *freq_of_vegetable* hay *daily_water* cũng thể hiện vai trò đáng kể, cho thấy lối sống vẫn đóng vai trò then chốt trong dự báo béo phì.

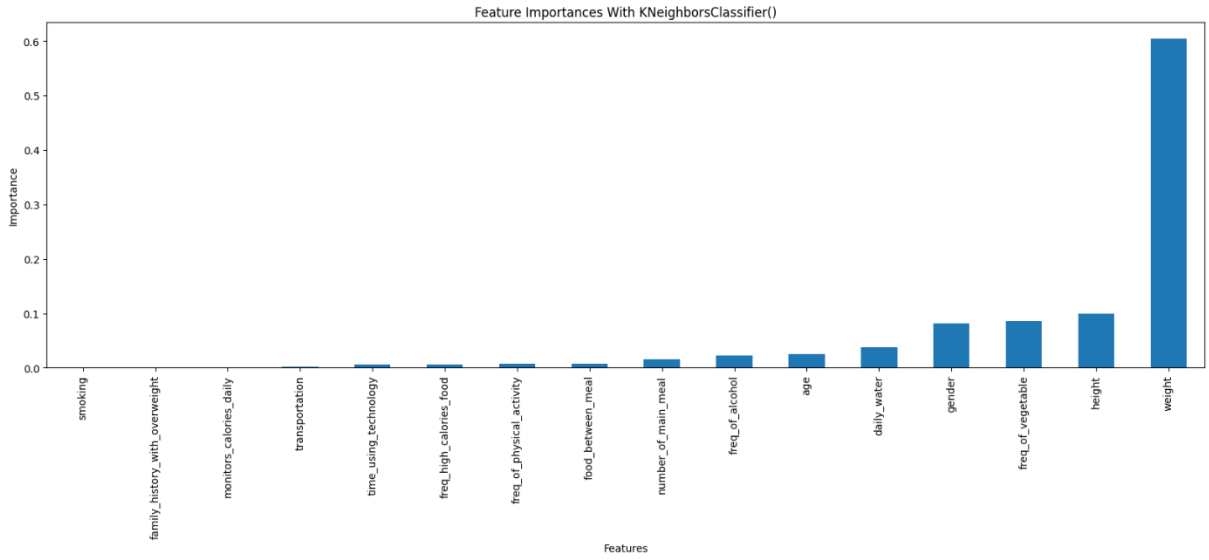
Đối với các đặc trưng thể hiện mức độ quan trọng thấp một cách bất ngờ, cần có các phân tích bổ sung như kiểm tra tương quan giữa các đặc trưng, phân tích đơn biến hoặc phân tích cặp với biến mục tiêu. Ngoài ra, việc thu thập dữ liệu chi tiết và chính xác hơn—chẳng hạn như phân loại cụ thể các loại thực phẩm giàu năng lượng hoặc mô tả rõ ràng hình thức vận động—có thể giúp cải thiện chất lượng mô hình và làm rõ hơn vai trò thực sự của các yếu tố này**.**



Hình 8.1: Biểu đồ đặc trưng quan trọng của mô hình Logistic Regression



Hình 8.2: Biểu đồ đặc trưng quan trọng của mô hình Support Vector Machine



Hình 8.3: Biểu đồ đặc trưng quan trọng của mô hình KNN

Với hai mô hình còn lại do chúng được hỗ trợ hàm *importance()* nên ít nhiều sẽ có sự khác biệt giữa việc chọn lọc các yếu tố đặc trưng

Phương pháp mà hàm *importance()* sử dụng trong hai mô hình bên dưới chính là dựa vào tổng mức giảm tạp chất trên từng nút hay từng cây liên kết với những chương trên chính là *Gini impurity hoặc entropy*. Nếu đặc trưng nào giảm *Gini* nhiều hơn thì tức đặc trưng đó quan trọng hơn

Quan sát hai biểu đồ bên dưới (*Hình 8.4, Hình 8.5*)

Mô hình Decision Tree:

Các đặc trưng được sắp xếp từ trái qua phải với mức độ quan trọng tầm ảnh hưởng tăng dần

*Các đặc trưng có tầm quan trọng trung bình (Moderately Important Features)***:**

- age (tuổi tác) có tầm quan trọng trong khoảng 0.05 cho thấy rằng tuổi tác phần nào đó cũng có phần ảnh hưởng đến khả năng gây béo phì (có thể tuổi tác cao khả năng vận động ít lượng mỡ tích tụ nhiều)

- gender (giới tính) có tầm quan trọng khoảng 0.16 **cho** thấy giới tính cũng gây ảnh hưởng đến mức độ gây béo phì chẳng hạn như đối với giới tính nam thường sẽ ở một số người khá ít vận động và thường xuyên có xu hướng tiêu thụ nhiều calories (có tác động đến những đặc trưng khác)

Hai đặc trưng này bắt đầu cho thấy mức độ đóng góp đáng kể hơn. Điều này gợi ý rằng age và gender có mối liên hệ nhất định với biến mục tiêu và được sử dụng bởi cây quyết định để tạo ra các phân tách có ý nghĩa. gender có tầm quan trọng cao hơn age, cho thấy nó có khả năng phân biệt tốt hơn.

Các đặc trưng quan trọng nhất (Most Important Features):

- height (chiều cao) với tầm quan trọng khoảng 0.25 không quá ngạc nhiên bởi lẽ đây chính là yếu tố ảnh hưởng sâu đến chỉ số lâm sàng cũng như sinh học cùng với weight (cân nặng) sẽ chính là những yếu tố quyết định sáu đến với chỉ số BMI (đánh giá một người mắc béo phì hay không)

-weight (cân nặng) với tầm quan trọng khoảng 0.45 cùng với height (chiều cao) đặc trưng này cũng chính là chìa khóa cũng như là yếu tố then chốt nhằm quyết định đến sự phân loại dự đoán của mô hình

Ý nghĩa: Trong nhiều ngữ cảnh, height và weight là các yếu tố cơ bản và có ảnh hưởng lớn đến nhiều thuộc tính hoặc tình trạng khác (ví dụ: sức khỏe, BMI, v.v.). Sự ưu tiên của mô hình đối với các đặc trưng này là hợp lý về mặt lý thuyết.

Các đặc trưng ít quan trọng nhất (Most Trivial Features):

- **family_history_with_overweight (tiền xử gia đình thừa cân hay không)**
- **monitors_calories_daily (theo dõi lượng calories nàp vào cơ thể)**
- **freq_of_physical_activity (tần suất luyện tập thể dục)**

- **freq_of_alcohol (tần suất sử dụng đồ cồn)**
- **smoking (hút thuốc)**
- **transportation (phương tiện di chuyển)**
- **time_using_technology (tần suất sử dụng công nghệ)**
- **number_of_main_meal (số lượng bữa ăn chính)**
- **daily_water (lượng nước mỗi ngày)**
- **food_between_meal (ăn vặt giữa bữa ăn)**
- **freq_of_vegetable (tần suất tiêu thụ rau)**
- **freq_high_calories_food (tần suất tiêu thụ thức ăn nhiều calories)**

Các đặc trưng này có giá trị tầm quan trọng gần như bằng 0. Điều này cho thấy chúng không đóng góp đáng kể vào khả năng phân loại của mô hình DecisionTreeClassifier trong tập dữ liệu cụ thể này. Từ đó có thể xuất hiện một số lý do:

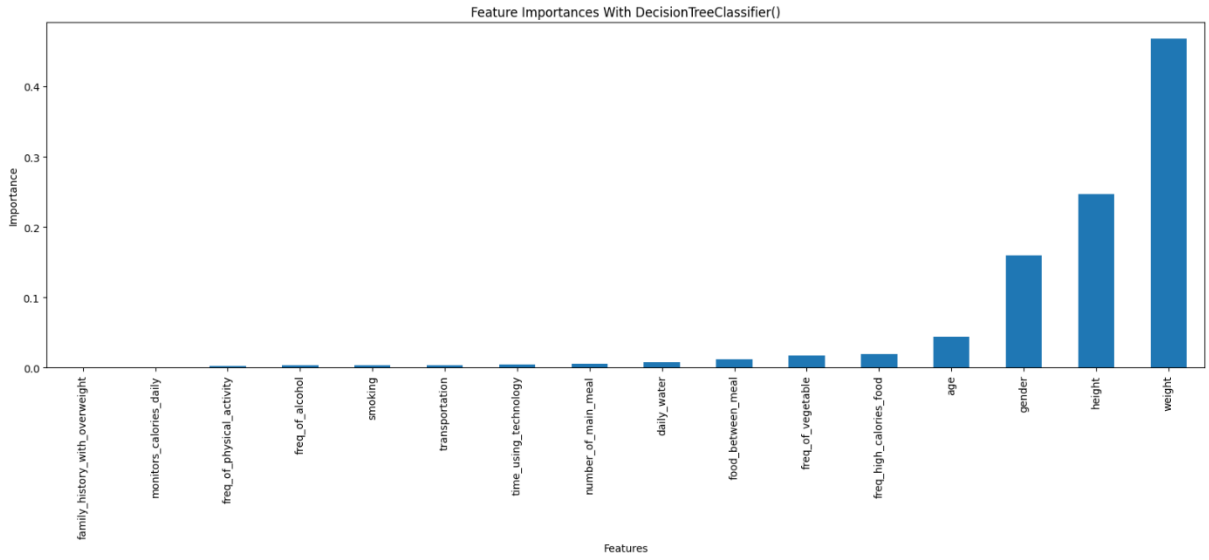
Thứ nhất, các đặc trưng này có thể không chứa đủ thông tin để phân biệt hiệu quả giữa các lớp hay quá ít để có thể ảnh hưởng đến phán đoán của mô hình

Thứ hai, dữ liệu trong các đặc trưng này có thể quá nhiễu, làm giảm khả năng cây quyết định sử dụng chúng một cách hiệu quả.

Thứ ba, ột đặc trưng có thể ít quan trọng nếu thông tin mà nó cung cấp đã được bao hàm bởi các đặc trưng khác có mối tương quan cao hơn với biến mục tiêu.

Thứ tư, DecisionTreeClassifier có xu hướng ưu tiên các đặc trưng có khả năng giảm tạp chất nhanh chóng ngay từ các nút trên của cây. Nếu các đặc trưng này không đáp ứng tiêu chí đó, chúng sẽ bị đánh giá thấp về tầm quan trọng.

Tương tự như với đánh giá rút ra từ mô hình Gradient Boosting ta cũng sẽ có những đặc trưng từ không quan trọng đến là yếu tố để quyết định dự đoán của một mô hình đồng thời cũng chính là yếu tố then chốt để đánh giá độ chính xác của mô hình



Hình 8.4: Biểu đồ đặc trưng quan trọng của mô hình Decision Tree

Mô hình Random Forest

Biểu đồ cũng trình bày các đặc trưng được sắp xếp theo thứ tự tăng dần về tầm quan trọng từ trái qua phải ta có lần lượt**.**

Các đặc trưng ít quan trọng nhất (Most Trivial Features):

- **smoking (Rất gần 0)**
- **monitors_calories_daily**
- **freq_high_calories_food**
- **transportation**
- **freq_of_alcohol**
- **food_between_meal**
- **family_history_with_overweight**

Các đặc trưng này có giá trị tầm quan trọng rất thấp, gần như bằng 0 đối với smoking. Điều này cho thấy chúng đóng góp không đáng kể vào khả năng phân loại của mô hình Random Forest. Có thể có một số lý do:

Thứ nhất, tương tự như Decision Tree các đặc trưng này có thể không chứa đủ thông tin để giúp mô hình phân biệt hiệu quả giữa các lớp.

Thứ hai, thông tin mà các đặc trưng này cung cấp có thể đã được nắm bắt bởi các đặc trưng khác có mối tương quan mạnh hơn với biến mục tiêu.

Thứ ba, các đặc trưng này có thể chứa nhiều nhiễu, làm giảm khả năng sử dụng chúng hiệu quả bởi các cây con.

Khả năng phân tán thông tin: Mặc dù một đặc trưng có vẻ quan trọng trong một cây đơn lẻ, nhưng nếu nó không nhất quán quan trọng trên nhiều cây trong rừng, tầm quan trọng trung bình của nó sẽ thấp.

Các đặc trưng có tầm quan trọng trung bình (Moderately Important Features):

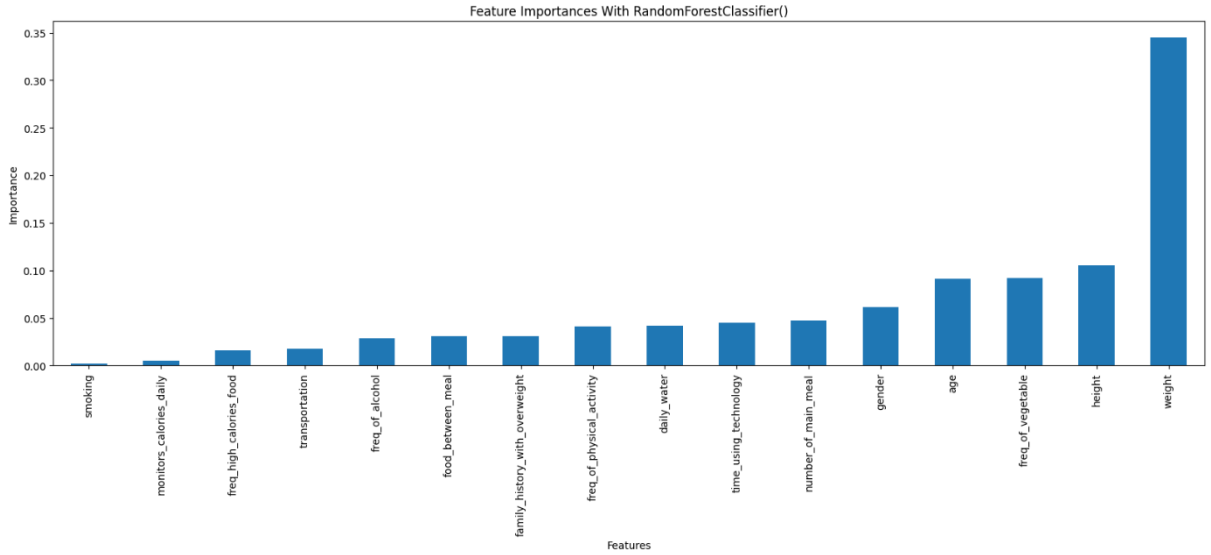
- **freq_of_physical_activity (Tẫn suất vận động)**
- **daily_water (Lượng nước nạp vào cơ thể mỗi ngày)**
- **time_using_technology (Thời gian sử dụng công nghệ)**
- **number_of_main_meal (Số lượng bữa ăn chính)**
- **gender (Giới tính)**
- **age (Tuổi)**
- **freq_of_vegetable (Tẫn suất ăn rau)**

Các đặc trưng này có giá trị tầm quan trọng ở mức trung bình, cho thấy chúng đóng góp ở mức độ vừa phải vào quá trình phân loại. Đặc biệt, gender, age, và freq_of_vegetable có tầm quan trọng đáng kể hơn so với nhóm trước đó, cho thấy chúng cung cấp thông tin hữu ích trong việc phân tách dữ liệu. Sự hiện diện của freq_of_vegetable ở đây, trong khi nó ít quan trọng hơn trong Decision Tree Classifier trước đó, có thể là một dấu hiệu của khả năng xử lý tốt hơn các mối quan hệ phức tạp của Random Forest.

Các đặc trưng quan trọng nhất (Most Important Features):

Tương tự như DecisionTreeClassifier, weight và height tiếp tục là hai đặc trưng quan trọng nhất, với weight có tầm quan trọng vượt trội. Điều này củng cố mạnh mẽ kết luận rằng:

Tính ổn định của tầm quan trọng: Các đặc trưng này là những yếu tố dự đoán mạnh mẽ và ổn định, được coi là quan trọng ngay cả khi mô hình được xây dựng trên các tập con dữ liệu ngẫu nhiên (bootstrapped samples) và các tập con đặc trưng (feature subsets) như trong Random Forest.



Hình 8.5: Biểu đồ đặc trưng quan trọng của mô hình Random Forest

Chương V

Kết luận và hướng phát triển

1. Tóm tắt kết quả

Sau khi đã thực hiện huấn luyện mô hình cũng như sử dụng các chỉ số để đánh giá kết quả nhận được từ các mô hình được rút ra như sau:

Mô hình hoạt động với độ chính xác (accuracy) cao nhất là hai mô hình Decision Tree và Random Forest. Hai mô hình này có độ chính xác cận tiệm hoàn hảo (trong khoảng 95%). Đánh giá tốt nhất ở lớp "***Béo phì loại III***" căn cứ theo Precision, Recall và F1-score mang lại kết quả như mong muốn

Mô hình hoạt động kém chính là những mô hình như Logistic Regression, SVM và KNN đặc biệt kém nhất là Logistic Regression. Ta có thể thấy được với bài toán phân loại đa lớp khá gây khó khăn cho mô hình này

Các đặc trưng quan trọng nhất theo cả ba phương pháp đặc trưng chính là chiều cao và cân nặng sở dĩ được như vậy chính là do chúng là những điều kiện thành phần nhằm quyết định chỉ số BMI nhằm đánh giá mức độ béo phì ở một con người, đồng thời cũng có thể rút ra được rằng hai đặc trưng này cũng có thể gây ảnh hưởng đến với những đặc trưng khác chứ không chỉ riêng biến mục tiêu (obesity_level)

Bên cạnh hai đặc trưng như chiều cao và cân nặng cũng có một số đặc trưng khá quan trọng như tuổi tác hay giới tính. Điều này cho thấy khi bạn có độ tuổi càng cao thì việc bạn tích tụ mỡ gia tăng nguy cơ béo phì cũng như khi là nam thì tỉ lệ bạn là nam mắc nguy cơ béo phì là khả thi

Tuy nhiên những kết luận trên chỉ mang tính chất tham khảo mô hình cũng có những hạn chế và thiếu sót như:

Đối với dữ liệu chưa thực sự đã "sạch" còn nhiều đặc trưng chứa dữ liệu bị nhiễu hay ngoại lai (outlier) có thể ảnh hưởng ít nhiều đến với kết quả của từng mô hình. Đặc biệt là với những mô hình bị nhạy cảm với dữ liệu ngoại lai

Chưa thể chọn ra được các siêu tham số để mô hình có thể dự đoán được tốt nhất. Việc so sánh trong bài tiểu luận này chỉ được thực hiện dựa trên các siêu tham số ngẫu nhiên, cơ bản

Chưa đề cập đến việc tối ưu hóa của các mô hình, nếu tối ưu hay loại bỏ các dữ liệu bị nhiễu thì kết quả có thể cúng sẽ thay đổi, chỉ đề cập đến các mô hình học máy cơ bản nếu thay thế bằng các mô hình học sâu thì kết quả có thể cải tiến hơn

Sử dụng các phương pháp chọn lọc đặc trưng còn khá hạn chế ở một số mô hình từ đó không thể đánh giá tổng quát mối tương quan giữa các đặc trưng, các đánh giá còn mang tính tham khảo và cần phải được đánh giá dựa trên nhiều yếu tố sinh trắc học khác như gen, môi trường, …

2. Ứng dụng thực tiễn

Một trong những ứng dụng nổi bật của mô hình dự đoán béo phì là hỗ trợ y tế dự phòng. Dựa trên dữ liệu đầu vào như độ tuổi, giới tính, chỉ số BMI, tần suất tập luyện, chế độ ăn uống và thói quen sinh hoạt, hệ thống có thể đưa ra dự đoán về nguy cơ béo phì của một cá nhân. Từ đó, bác sĩ có thể chủ động tư vấn cho bệnh nhân những giải pháp cụ thể như thay đổi lối sống, điều chỉnh thực đơn dinh dưỡng hoặc tăng cường vận động nhằm giảm thiểu rủi ro.

Hiện nay, nhiều ứng dụng chăm sóc sức khỏe trên điện thoại thông minh và thiết bị đeo tay (smartwatch, fitness tracker) đã tích hợp các thuật toán dự đoán béo phì để cá nhân hóa trải nghiệm người dùng. Ví dụ, ứng dụng MyFitnessPal hay Fitbit sử dụng dữ liệu từ người dùng để cảnh báo sớm nguy cơ béo phì và đề xuất kế hoạch luyện tập hoặc chế độ ăn uống hợp lý. Điều này giúp người dùng chủ động theo dõi và điều chỉnh sức khỏe của mình một cách khoa học và hiệu quả.

Trong lĩnh vực bảo hiểm sức khỏe, các công ty bảo hiểm có thể sử dụng mô hình dự đoán béo phì để đánh giá rủi ro tiềm ẩn của khách hàng. Dựa trên kết quả dự đoán, doanh nghiệp có thể thiết kế mức phí bảo hiểm phù hợp, đồng thời đưa ra các khuyến nghị chăm sóc sức khỏe để hỗ trợ khách hàng duy trì thể trạng tốt. Điều này không chỉ giúp doanh nghiệp tối ưu hóa chi phí mà còn nâng cao chất lượng dịch vụ và sự hài lòng của khách hàng.

Bên cạnh ứng dụng cá nhân, dự đoán béo phì còn đóng vai trò quan trọng trong công tác nghiên cứu và quản lý sức khỏe cộng đồng. Các nhà hoạch định chính sách có thể sử dụng dữ liệu dự đoán để theo dõi xu hướng béo phì theo vùng miền, độ tuổi, giới tính,... Từ đó, họ có thể triển khai các chương trình tuyên truyền, giáo dục hoặc hỗ trợ dinh dưỡng – vận động một cách phù hợp và hiệu quả hơn cho từng nhóm dân cư.

Tài liệu tham khảo

[1] W. H. Organization, "Obesity and overweight.," 2924. [Online]. Available: Obesity and overweight. https://www.who.int.

[2] P. e. al., Machine learning in obesity prediction. Journal of Biomedical Informatics., 2020.

[3] R. e. al, Predictive modeling of obesity using ML. Elsevier., 2021.

[4] T. Mitchell, Machine Learning. McGraw-Hill., 1997.

[5] "scikit-learn," [Online]. Available: https://scikit-learn.org/stable/.

[6] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2019.

[7] I. T. O. S. A. M. N. V. I. &. C. I. Kavakiotis, Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 2017.