ỨNG DỤNG NGÔN NGỮ PYTHON VÀ ĐÁNH GIÁ CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU TRONG PHÂN TÍCH DỰ ĐOÁN BÉO PHÌ DỰA TRÊN CHỈ SỐ NHÂN TRẮC HỌC

Đinh Tuấn Long[1], Nguyễn Thị Tú Lan[1], Trần Thùy Ninh[1]
Email: dinhtuanlong@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 06/12/2024
Ngày phản biện đánh giá: 12/06/2025
Ngày bài báo được duyệt đăng: 26/06/2025

DOI: 10.59266/houjs.2025.582

Tóm tắt: *Trong bối cảnh tỷ lệ béo phì gia tăng nhanh chóng tại Việt Nam (tăng 38% trong giai đoạn 2010-2014) và trên toàn cầu, việc áp dụng các phương pháp khoa học dữ liệu tiên tiến mang lại tiềm năng đáng kể trong việc cải thiện độ chính xác chẩn đoán và chiến lược can thiệp cá nhân hóa. Nghiên cứu của chúng tôi hướng tới việc triển khai các mô hình học máy phân tích các dữ liệu y tế để dự đoán nguy cơ mắc bệnh béo phì, kết hợp với việc sử dụng các thư viện của ngôn ngữ lập trình Python trên năm lĩnh vực chức năng: xử lý dữ liệu (Pandas, Dask), tính toán số học (NumPy, SymPy), trực quan hóa (Matplotlib, Seaborn, Plotly), học máy (Scikit-learn, TensorFlow, PyTorch) và phân tích thống kê (Statsmodels, SciPy). Sử dụng bộ dữ liệu đa quốc gia (n=33.610) gồm 17 biến nhân trắc học và lối sống, chúng tôi đã phát triển và tối ưu hóa nhiều mô hình phân loại thông qua phương pháp tối ưu hóa Bayesian. Kết quả cho thấy thuật toán LightGBM đạt hiệu suất vượt trội (độ chính xác=93,07%, F1-score=92,48%, PR-AUC=96,30%), vượt trội đáng kể so với mô hình Random Forest (độ chính xác=92,25%) và Multi-Layer Perceptron (độ chính xác=89,05%). Việc triển khai các công cụ này đã tạo điều kiện cho việc phát triển hệ thống tích hợp cung cấp khuyến nghị sức khỏe cá nhân hóa dựa trên mức độ rủi ro béo phì dự đoán. Nghiên cứu này đóng góp vào cả sự tiến bộ về phương pháp phân tích dữ liệu sức khỏe và ứng dụng thực tiễn trong phòng chống béo phì thông qua can thiệp cá nhân hóa dựa trên công nghệ.*

Từ khóa: *dự đoán béo phì, khai phá dữ liệu, ngôn ngữ lập trình python, phương pháp bayesian, mô hình LightGBM, mô hình Random Forest, mô hình MLP*

I. Đặt vấn đề

Quá trình chuyển đổi số trong lĩnh vực y tế đã xác định dữ liệu y khoa như một nguồn tài nguyên quan trọng cho việc nâng cao độ chính xác trong chẩn đoán và hiệu quả điều trị. Hiện nay, bệnh béo phì đang gia tăng nhanh chóng tại Việt Nam, với tỷ lệ tăng 38% trong giai đoạn 2010-2014 *(theo thông tin trong Quyết định 2892/QĐ-BYT của Bộ Y tế)*. Dữ liệu dịch

[1] Trường Đại học Mở Hà Nội

tễ học từ năm 2024 cho thấy béo phì ảnh hưởng đến 22,1% người trưởng thành tại các trung tâm đô thị như Hà Nội và Thành phố Hồ Chí Minh - gần gấp đôi tỷ lệ ở khu vực nông thôn (11,2%). Xu hướng này có mối tương quan đáng kể với hồ sơ rủi ro gia tăng đối với bệnh tiểu đường, bệnh tim mạch và các bệnh mãn tính khác (Tạp chí Bảo hiểm Xã hội, 2022).

Việc áp dụng các phương pháp tính toán để phân tích dữ liệu nhân trắc học giúp đề xuất hướng giải quyết vấn đề sức khỏe cộng đồng ngày càng gia tăng này. Tuy nhiên, các đánh giá về hiệu quả của những công cụ này cho dự đoán béo phì và hỗ trợ can thiệp sớm vẫn còn hạn chế, đặc biệt trong bối cảnh các quốc gia đang phát triển như Việt Nam.

Nghiên cứu này triển khai triển khai các kỹ thuật học máy thông qua các thư viện của ngôn ngữ lập trình Python để phân tích chỉ số khối cơ thể và các số đo nhân trắc học liên quan, đánh giá nguy cơ béo phì và khuyến nghị các vấn đề cá nhân hóa. Các kỹ thuật học máy và khai phá dữ liệu khi được tối ưu hóa thích hợp với các dữ liệu y tế có thể đạt được hiệu suất dự đoán tốt hơn so với các phương pháp thống kê truyền thống và cung cấp kết quả có thể diễn giải hỗ trợ quá trình ra quyết định lâm sàng.

II. Cơ sở lý thuyết

2.1. Khái niệm về béo phì và các chỉ số đánh giá

Béo phì là tình trạng tích tụ mỡ thừa trong cơ thể đến mức có thể gây tác động tiêu cực đến sức khỏe. Tổ chức Y tế Thế giới (WHO) định nghĩa chỉ số khối cơ thể (BMI) được tính bằng cân nặng (kg) chia cho bình phương chiều cao (m^2). Trong nghiên cứu này, chúng tôi áp dụng phân loại béo phì với 7 mức độ theo chỉ số BMI, từ thiếu cân đến béo phì độ III.

Bảng 1. Danh sách phân loại mức độ béo phì

Mức độ	Mô tả
Insufficient_Weight	BMI < 18,5
Normal_Weight	18,5 ≤ BMI < 24,9
Overweight_Level_I	25,0 ≤ BMI < 27,5
Overweight_Level_II	27,5 ≤ BMI < 30,0
Obesity_Type_I	30,0 ≤ BMI < 34,9
Obesity_Type_II	35,0 ≤ BMI < 39,9
Obesity_Type_III	BMI ≥ 40

2.2. Ứng dụng học máy trong phân tích dữ liệu y tế

Học máy đã trở thành công cụ không thể thiếu trong phân tích dữ liệu y tế, mang lại khả năng phát hiện mẫu phức tạp và mối quan hệ phi tuyến tính mà các phương pháp thống kê truyền thống khó nhận biết. Trong lĩnh vực dự đoán béo phì, các thuật toán học máy như Random Forest, Support Vector Machines, và mạng nơ-ron nhân tạo đã được áp dụng rộng rãi để phân tích dữ liệu nhân trắc học, thói quen ăn uống, và yếu tố di truyền.

Theo nghiên cứu của Wiyono và cộng sự (2019), các thuật toán học máy như K-Nearest Neighbors (KNN), Support Vector Machine (SVM), và Decision Tree có thể đạt độ chính xác từ 92% đến 95% trong dự đoán, cho thấy tiềm năng của các phương pháp này trong việc phân tích dữ liệu phức tạp. Tương tự, Sheth và cộng sự (2022) đã chứng minh rằng Naïve Bayes có hiệu quả nổi bật trong một số tập dữ liệu nhất định, trong khi SVM thường đạt hiệu suất tốt với dữ liệu có nhiều đặc trưng.

III. Phương pháp nghiên cứu

3.1. Phương pháp nghiên cứu

Nghiên cứu này áp dụng phương pháp luận định lượng theo quy trình chuẩn của khoa học dữ liệu, bao gồm năm giai

đoạn chính: thu thập và tiền xử lý dữ liệu, khám phá và phân tích dữ liệu, lựa chọn đặc trưng, xây dựng và huấn luyện mô hình, đánh giá và giải thích kết quả.

Thiết kế nghiên cứu tuân theo mô hình thử nghiệm với phân chia tập dữ liệu thành tập huấn luyện (80%) và tập kiểm thử (20%) để đánh giá khả năng tổng quát hóa của mô hình.

3.2. Nguồn dữ liệu sử dụng

3.2.1. Nguồn dữ liệu

Do các vấn đề về bảo vệ thông tin cá nhân cũng như khó khăn trong việc tiếp cận các nguồn dữ liệu y tế chính thức tại Việt Nam, chúng tôi lựa chọn bộ dữ liệu từ Kaggle, một nền tảng uy tín cho nguồn tài nguyên khoa học dữ liệu. Quy trình lựa chọn ưu tiên chất lượng dữ liệu, độ tin cậy và tuân thủ quy định bảo mật thông tin (Kaggle, n.d.). Dữ liệu gốc từ 33.610 người tham gia, được thu thập với sự chấp thuận từ Hội đồng Đạo đức Nghiên cứu Y tế của Đại học Inonu (số phê duyệt 2023/4677), đảm bảo tuân thủ các tiêu chuẩn đạo đức cho nghiên cứu liên quan đến sức khỏe.

3.2.2. Đặc điểm của dữ liệu

Bộ dữ liệu gốc của nghiên cứu bao gồm dữ liệu từ nhiều người tham gia khác nhau, với các thuộc tính sau:

Bảng 2. Danh sách thuộc tính trong tập dữ liệu

Thuộc tính	Mô tả	Giá trị
Gender	Biến phân loại	male female
Age	Biến số	Số tự nhiên (từ 14 - 61)
Height	Biến số	Số thập phân (mét)
Weight	Biến số	Số thập phân (kg)
Family history of overweight	Biến phân loại cho biết có người nào trong gia đình bị thừa cân không	yes no
FAVC	Biến phân loại cho biết cá nhân có thường xuyên ăn thực phẩm calo cao không	yes no
FCVC	Biến thứ bậc cho biết mức độ thường xuyên cá nhân ăn rau	1 : không bao giờ 2 : thỉnh thoảng 3 : luôn luôn
NCP	Biến thứ bậc cho biết số bữa ăn chính mỗi ngày của cá nhân	1 : từ 1 đến 2 2 : bằng 3 3 : lớn hơn 3 4 : không trả lời
MTRANS	Biến phân loại về loại phương tiện giao thông mà cá nhân sử dụng	automobile motorbike bike public transportation walking
CAEC	Biến thứ bậc về mức độ thường xuyên ăn vặt giữa các bữa	no sometimes frequently always
SMOKE	Biến phân loại về việc có hút thuốc hay không	yes no

Thuộc tính	Mô tả	Giá trị
CH2O	Biến thứ bậc cho biết lượng nước uống hàng ngày	1 : ít hơn 1 lít 2 : từ 1 đến 2 lít 3 : hơn 2 lít
SCC	Biến phân loại cho biết liệu cá nhân có theo dõi lượng calo tiêu thụ của mình không	yes no
FAF	Biến thứ bậc cho biết mức độ thường xuyên tham gia hoạt động thể chất	1 : không bao giờ 2 : từ 1-2 lần/tuần 3 : 3 lần/tuần 4 : từ 4-5 lần/tuần
TUE	Biến thứ bậc về thời gian sử dụng thiết bị điện tử	0 : không 1 : ít hơn 1 giờ 2 : từ 1 đến 3 giờ 3 : hơn 3 giờ
CALC	Biến thứ bậc về mức độ thường xuyên uống rượu	no sometimes frequently always
NObesity	Biến thứ bậc về mức độ béo phì của cá nhân theo chỉ số BMI của họ	Insufficient_Weight Normal_Weight Overweight_Level_I Overweight_Level_II Obesity_Type_I Obesity_Type_II Obesity_Type_III

3.3. Giới thiệu cách triển khai các mô hình

3.3.1. Chuẩn bị dữ liệu huấn luyện

Bộ dữ liệu gồm 33.610 mẫu được chia thành các tập huấn luyện (training), kiểm tra (testing) và xác thực (validation) theo phương pháp phân tầng (stratified sampling) để đảm bảo phân phối đều giữa 7 lớp mục tiêu:

- **Tập kiểm tra (Testing set)**: 20% bộ dữ liệu (6.722 mẫu)
- **Tập huấn luyện (Training set)**: 80% còn lại (26.888 mẫu), với mỗi lớp có khoảng 3.841 mẫu.

Tập huấn luyện được sử dụng trong quá trình kiểm định chéo 5 phần (5 folds cross-validation) để đánh giá hiệu suất của nhiều mô hình học máy khác nhau.

3.3.2. Lựa chọn mô hình

Sau khi phân tích dữ liệu, chúng tôi đã lựa chọn ba thuật toán phân loại chính để dự đoán mức độ béo phì:

1. **Random Forest (RF)**: Một thuật toán học tập tổng hợp kết hợp nhiều cây quyết định để cải thiện độ chính xác và giảm hiện tượng quá khớp (overfitting). RF có khả năng xử lý dữ liệu có nhiều đặc trưng, bao gồm cả biến số và biến phân loại.

2. **LightGBM (LGBM)**: Một khung học tập tăng cường theo độ dốc (gradient) sử dụng thuật toán học tập dựa trên cấu trúc dữ liệu cây, nổi bật với chiến lược phát triển cây theo hướng ưu tiên nhánh lá (leaf-wise). LGBM được chọn vì hiệu quả tính toán cao với tập dữ liệu lớn và khả năng xử lý dữ liệu không cân bằng.

3. **Multi-Layer Perceptron (MLP)**: Một lớp mạng thần kinh nhân tạo lan truyền thẳng (Feedforward Neural Network) bao gồm nhiều lớp perceptron với các hàm kích hoạt phi tuyến. MLP được chọn vì khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các đặc trưng.

3.3.3. Tối ưu hóa siêu tham số

Chúng tôi áp dụng phương pháp tối ưu hóa Bayesian để xác định bộ siêu tham số tối ưu cho các mô hình học máy với quy trình:

1. **Định nghĩa không gian tham số**: Xác định phạm vi khả thi cho mỗi siêu tham số.

2. **Mô hình hóa Gaussian Process**: Hàm mục tiêu:

$$f(D))\sim GP(\ (x),k(x,x'))$$

Với hàm kernel RBF:

$$k(x,x') = exp(-(||x-x'||^2)/(2l^2))$$

3. **Định nghĩa hàm mục tiêu**:

$$f(\theta) = -Macro\ F1(\theta)$$

Trong đó θ là một bộ siêu tham số cần tối ưu hóa.

4. **Chiến lược thu nhận**: Expected Improvement (EI) được sử dụng để xác định các điểm lấy mẫu tối ưu:

$$EI(x) = E[max(0,f(x^*) - f(x))]\ )$$

Mỗi cấu hình siêu tham số được đánh giá thông qua kiểm định chéo 5-fold trên tập huấn luyện, với F1-score được chọn làm chỉ tiêu đánh giá chính.

IV. Kết quả và thảo luận

4.1. Xử lý tiền dữ liệu

4.1.1. Xử lý giá trị thiếu và dữ liệu trùng lặp

Chúng tôi đã sử dụng các phương thức *duplicated(), info()* và *isna()* từ thư viện Pandas để phát hiện và xử lý dữ liệu trùng lặp và giá trị thiếu. Do tỷ lệ dữ liệu trùng lặp và thiếu không lớn, chúng tôi đã áp dụng phương pháp loại bỏ hoàn toàn để duy trì tính toàn vẹn của dữ liệu.

4.1.2. Phát hiện và xử lý ngoại lai

Đối với biến phân loại, chúng tôi sử dụng biểu đồ *boxplot* để xác định ngoại lai, với lỗi chính tả nhỏ được sửa trực tiếp và các giá trị bất thường còn lại được loại bỏ. Đối với biến liên tục như Tuổi, Chiều cao, Cân nặng, chúng tôi áp dụng các phương pháp sau:

Phương pháp Z-score:

$$Z = \frac{X-\mu}{\sigma}$$

Với X là điểm dữ liệu, μ là giá trị trung bình, và σ là độ lệch chuẩn. Giá trị được coi là ngoại lai nếu $|Z| > 3$.

Phương pháp IQR (Interquartile Range): $IQR= Q3 -Q1$

Với Q_1 và Q_3 lần lượt là phân vị thứ 25 và 75. Giá trị được coi là ngoại lai nếu nằm ngoài khoảng:

$$[Q1-1,5\times IQR;\ Q3+1,5\times IQR]$$

4.1.3. Chuẩn hóa dữ liệu

Dữ liệu được chuẩn hóa qua các kỹ thuật:

- **One-Hot Encoding**: Áp dụng cho các biến phân loại danh nghĩa để chuyển đổi chúng thành các vector nhị phân phù hợp với các mô hình học máy.
- **Label Encoding**: Áp dụng cho các biến thứ tự để bảo toàn mối quan hệ thứ tự vốn có trong khi chuyển đổi sang định dạng số.

Quá trình tiền xử lý toàn diện này đã tạo ra một bộ dữ liệu sạch, chuẩn hóa, phù hợp cho việc mô hình hóa và phân tích tiếp theo.

4.2. Cơ sở khoa học trong việc lựa chọn các đặc trưng

Căn cứ trên các nghiên cứu đã có của Wang và Beydoun (2007), Tapera và cộng sự (2017), Panuganti và cộng sự (2023), Al-Hazzaa và cộng sự (2012), Carlson và cộng sự (2012), Specht và cộng sự (2022), Ohlsson và Manjer (2020), Gozukara (Bag

& cộng sự, 2023) về các vấn đề có thể ảnh hưởng tới nguy cơ béo phì, chúng tôi đã lựa chọn ra 17 đặc trưng trên các nhóm:

- Đặc trưng nhân khẩu học và sinh lý: giới tính, tuổi, chiều cao và cân nặng.

- Tiền sử gia đình và di truyền: Tiền sử gia đình có người thừa cân.

- Chế độ ăn uống và thói quen sinh hoạt: FAVC (tiêu thụ thực phẩm giàu calo), FCVC (tiêu thụ rau củ), NCP (số bữa ăn chính), CAEC (ăn giữa các bữa), CH2O (lượng nước tiêu thụ), CALC (tiêu thụ rượu).

- Hoạt động thể chất và lối sống: FAF (tần suất hoạt động thể chất), TUE (thời gian sử dụng thiết bị điện tử), MTRANS (phương tiện di chuyển), SMOKE (hút thuốc), SCC (theo dõi lượng calo tiêu thụ).

17 đặc trưng trong mô hình của chúng tôi cho phép nắm bắt đầy đủ các yếu tố ảnh hưởng đến béo phì, bao gồm cả các yếu tố sinh học, hành vi và môi trường. Điều này cung cấp một cách tiếp cận toàn diện để dự đoán và hiểu về nguy cơ béo phì, phù hợp với mô hình sinh thái của béo phì được đề xuất bởi các tổ chức y tế hàng đầu.

4.3. Triển khai các mô hình và đánh giá kết quả

Các dữ liệu được đưa vào ba mô hình: RandomForest, LightGBM và MLP, với tập huấn luyện gồm 26.888 mẫu, sử dụng nhiều chỉ số đánh giá để cung cấp cái nhìn toàn diện về hiệu suất của mô hình. Mỗi bộ siêu tham số được kiểm định 5-fold cross-validation, với F1-score được chọn làm chỉ tiêu đánh giá chính do khả năng cân bằng tốt giữa precision và recall cho cả giai đoạn ban đầu và kiểm định.

Bảng 3. Bảng thống kê các chỉ số ban đầu của mô hình

Mô hình	Accuracy	Precision	Recall	F1-score	PR-AUC
LGBM	92,44%	92,48%	92,44%	92,45%	96,08%
Random Forest	92,25%	83,32%	82,25%	82,24%	85,28%
MLP	88,90%	89,55%	88,90%	88,82%	83,34%

Bảng 4. Bảng thống kê các chỉ số của mô hình trong quá trình kiểm định

Mô hình	Accuracy	Precision	Recall	F1-score	PR-AUC
LGBM	92,44%	92,60%	92,55%	92,57%	96,42%
Random Forest	91,65%	91,72%	91,65%	91,66%	92,95%
MLP	83,03%	82,74%	83,76%	83,68%	88,40%

Bảng 5. Bảng thống kê các chỉ số của mô hình sau khi tinh chỉnh tham số

Mô hình	Ban đầu	Validation Accuracy	Sau tinh chỉnh	Đánh giá tổng quan
LGBM	92,44%	92,44%	93,07%	Ban đầu đã có độ chính xác cao (92,44%), sau tinh chỉnh tăng nhẹ lên 92,07%. Điều này cho thấy mô hình đã gần tối ưu từ đầu, và tinh chỉnh giúp cải thiện nhẹ nhưng vẫn đáng ghi nhận
Random Forest	92,25%	91,65%	92,25%	Độ chính xác sau tinh chỉnh giảm nhẹ, tuy nhiên vẫn duy trì ở mức ổn định, không có sự cải thiện rõ rệt.
MLP	88,90%	83,03%	89,05%	Mô hình MLP có sự cải thiện so tinh chỉnh, tuy nhiên độ chính xác vẫn thấp hơn. Các mô hình khác cho thấy cần cải thiện thêm.

Trong giai đoạn tối ưu hóa, các mô hình được huấn luyện sử dụng tập dữ liệu và đánh giá hiệu suất thông qua cross-validation 5-fold. Kết quả tối ưu được thể hiện qua độ chính xác (LGBM: 92,44%, RF: 91,65%, MLP: 83,03%). Sau khi xác định được bộ siêu tham số tối ưu, các mô hình được tinh chỉnh, huấn luyện lại trên toàn bộ tập dữ liệu.

4.3.1. So sánh hiệu suất các mô hình

Sử dụng mô hình đề xuất với ba thuật toán MLP, Random Forest và LightGBM được huấn luyện trên tập dữ liệu 24,000 mẫu và kiểm thử trên tập 6,000 mẫu độc lập, chúng tôi thu được kết quả như sau:

Bảng 6. Thống kê kết quả kiểm thử các mô hình

Mô hình	Độ chính xác	Precision	Recall	F1-score	PR-AUC
LGBM	93,07%	92,54%	92,46%	92,48%	96,30%
Random Forest	91,25%	91,76%	91,50%	91,58%	95,28%
MLP	89,05%	88,45%	87,90%	87,92%	92,94%

Kết quả thử nghiệm cho thấy:

• **LightGBM** đạt hiệu suất cao nhất với độ chính xác 93,07%, chỉ số F1-Score 92,48% và PR-AUC 96,30%, cho thấy khả năng cân bằng tốt giữa độ chính xác và độ bao phủ. Đặc biệt, tỷ lệ Recall đạt 92,46% cho thấy mô hình ít bỏ sót các trường hợp dương tính.

• **Random Forest** có hiệu suất ổn định với độ chính xác 91,25% và F1-Score 91,58%, phù hợp cho các ứng dụng yêu cầu độ tin cậy cao. Tuy nhiên, PR-AUC (95,28%) thấp hơn LightGBM, cho thấy khả năng phân loại các lớp thiểu số kém hơn một chút.

• **MLP** đạt độ chính xác thấp nhất (89,05%) nhưng vẫn duy trì PR-AUC 92,94%, phản ánh hiệu suất chấp nhận được nhưng kém hơn hai mô hình dựa trên cây quyết định.

4.3.2. Đánh giá tầm quan trọng của đặc trưng

Phân tích tầm quan trọng của đặc trưng từ mô hình LightGBM cho thấy:

1. **Tiền sử gia đình thừa cân** (20,3%): Yếu tố quan trọng nhất, phản ánh tác động mạnh mẽ của di truyền và môi trường gia đình đối với nguy cơ béo phì.

2. **FAF - Tần suất hoạt động thể chất** (15,7%): Yếu tố quan trọng thứ hai, khẳng định vai trò bảo vệ của hoạt động thể chất đối với béo phì.

3. **FAVC - Tiêu thụ thực phẩm giàu calo** (13,2%): Ảnh hưởng đáng kể đến tình trạng béo phì, phù hợp với hiểu biết hiện tại về vai trò của chế độ ăn uống.

4. **Tuổi** (10,5%): Xác nhận rằng nguy cơ béo phì tăng theo tuổi, phản ánh những thay đổi chuyển hóa và lối sống.

5. **MTRANS - Phương tiện di chuyển** (9,8%): Cho thấy tác động của vận động hàng ngày đến cân nặng.

6. **CH2O - Lượng nước tiêu thụ** (7,6%): Phản ánh tầm quan trọng của hydrat hóa trong quản lý cân nặng.

7. **TUE - Thời gian sử dụng thiết bị điện tử** (6,2%): Nhấn mạnh mối liên hệ giữa lối sống ít vận động và nguy cơ béo phì.

8. **Giới tính** (5,4%): Xác nhận sự khác biệt về giới trong phân bố mỡ cơ thể và nguy cơ béo phì.

9. **FCVC - Tiêu thụ rau củ** (4,8%): Cho thấy tác động bảo vệ của chế độ ăn giàu rau củ.

Các đặc trưng còn lại (CAEC, CALC, SCC, SMOKE, NCP) chiếm tỷ lệ thấp hơn (tổng cộng 6,5%) nhưng vẫn đóng góp vào hiệu suất tổng thể của mô hình. Điều này chứng minh giá trị của việc sử dụng tất cả 17 đặc trưng trong xây dựng mô hình dự đoán toàn diện.

Căn cứ trên các kết quả đạt được, có thể nhận thấy các yếu tố như di truyền và gia đình đóng vai trò then chốt trong việc gây ra nguy cơ béo phì, và ngược lại, lối sống tích cực có tác động rất có ý nghĩa với việc giảm nguy cơ. Từ đó có thể đề xuất các chính sách phù hợp với các cá nhân có nguy cơ cũng như phương án điều chỉnh, phòng ngừa.

V. Kết luận

Dự đoán và đánh giá nguy cơ béo phì dựa trên chỉ số cơ thể là một nhu cầu cấp thiết trong bối cảnh tỷ lệ béo phì gia tăng nhanh chóng, đặc biệt tại các quốc gia đang phát triển như Việt Nam. Nghiên cứu này đã đánh giá một cách hệ thống các khung công cụ Python để phân tích dữ liệu nhân trắc học và xây dựng mô hình dự đoán béo phì toàn diện. Các kết quả thu được từ các thuật toán học máy có thể được sử dụng làm thông tin hỗ trợ cho quá trình phát hiện sớm, phòng ngừa nguy cơ béo phì, mang lại các giá trị tích cực cho xã hội.

Những phát hiện chính từ nghiên cứu bao gồm:

1. Xác nhận vai trò then chốt của yếu tố di truyền trong nguy cơ béo phì, cho thấy tầm quan trọng của sàng lọc và can thiệp sớm cho những người có tiền sử gia đình bị béo phì.

2. Tác động bảo vệ đáng kể của hoạt động thể chất thường xuyên và lối sống tích cực, ủng hộ việc tích hợp hoạt động thể chất vào cuộc sống hàng ngày.

3. Khả năng phân loại chính xác giữa các mức độ béo phì khác nhau, mặc dù vẫn có thách thức trong việc phân biệt giữa các mức độ kề cận.

4. Tính khả thi của việc phát triển công cụ dự đoán tự động để hỗ trợ sàng lọc và can thiệp béo phì trong môi trường lâm sàng.

Tuy nhiên, nghiên cứu vẫn có một số hạn chế. Dữ liệu chủ yếu từ khu vực Mỹ Latinh, có thể không hoàn toàn phản ánh đặc điểm dân số Việt Nam. Ngoài ra, mặc dù bộ dữ liệu khá lớn (33.610 mẫu), việc phân phối không đồng đều giữa các nhóm tuổi và vùng địa lý có thể ảnh hưởng đến khả năng tổng quát hóa của mô hình.

Để nâng cao tính ứng dụng thực tế, các hướng nghiên cứu trong tương lai bao gồm: (1) thu thập dữ liệu từ dân số Việt Nam; (2) tích hợp thêm các chỉ số sinh học như mỡ nội tạng, chỉ số đường huyết; và (3) phát triển ứng dụng di động tích hợp AI để hỗ trợ người dùng theo dõi sức khỏe. Những nỗ lực này sẽ góp phần vào việc phòng ngừa và quản lý béo phì hiệu quả hơn, giảm gánh nặng y tế công cộng.

Tài liệu tham khảo

[1]. Al-Hazzaa, H. M., Abahussain, N. A., Al-Sobayel, H. I., Qahwaji, D. M., & Musaiger, A. O. (2012*). Lifestyle factors associated with overweight and obesity among Saudi adolescents*. BMC Public Health, 12, 354. https://doi.org/10.1186/1471-2458-12-354.

[2]. Bag, H. G. G., Yagin, F. H., Gormez, Y., Prieto González, P., Colak, C., Gülü, M., Badicu, G., & Ardigò, L. P. (2023). Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits. *International Journal of Environmental Research and Public Health, 20*(3), 2094. https://doi.org/10.3390/ijerph20032094.

[3]. Carlson, J. A., Crespo, N. C., Sallis, J. F., Patterson, R. E., & Elder, J. P. (2012). Dietary-related and physical activity-related predictors of obesity in children: A 2-year prospective study. *Childhood Obesity, 8*(2), 110-115. https://doi.org/10.1089/chi.2011.0071.

[4]. Kaggle. (n.d.). *Privacy Policy*. https://www.kaggle.com/privacy.

[5]. Ohlsson, B., & Manjer, J. (2020). Sociodemographic and lifestyle factors in relation to overweight defined by BMI and “normal-weight obesity”.

Journal of Obesity, 2020, 2070297. https://doi.org/10.1155/2020/2070297.

[6]. Panuganti, K. K., Nguyen, M., & Kshirsagar, R. K. (2023). *Obesity*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK459357/.

[7]. Specht, I. O., Heitmann, B. L., & Larsen, S. C. (2022). Physical activity and subsequent change in body weight, composition and shape: Effect modification by familial overweight. *Frontiers in Endocrinology, 13*, Article 787827. https://doi.org/10.3389/fendo.2022.787827.

[8]. Tapera, R., Merapelo, M. T., Tumoyagae, T., Maswabi, T. M., Erick, P., Letsholo, B., & Mbongwe, B. (2017). The prevalence and factors associated with overweight and obesity among University of Botswana students. *Cogent Medicine, 4*(1), Article 1357249. https://doi.org/10.1080/2331205X.2017.1357249.

[9]. Tạp chí Bảo hiểm Xã hội. (2022). Bộ Y tế: Ban hành hướng dẫn chẩn đoán và điều trị bệnh béo phì. *Tạp chí Bảo hiểm Xã hội*. https://tapchibaohiemxahoi.gov.vn/bo-y-te-ban-hanh-huong-dan-chan-doan-va-dieu-tri-benh-beo-phi-93538.html.

[10]. Wang, Y., & Beydoun, M. A. (2007). The obesity epidemic in the United States—Gender, age, socioeconomic, racial/ethnic, and geographic characteristics: A systematic review and meta-regression analysis. *Epidemiologic Reviews, 29*(1), 6-28. https://doi.org/10.1093/epirev/mxm007.

APPLYING PYTHON AND EVALUATING DATA MINING TECHNIQUES FOR OBESITY PREDICTION BASED ON ANTHROPOMETRIC INDICATORS

Dinh Tuan Long[2], Nguyen Thi Tu Lan[2], Tran Thuy Ninh[2]

Abstract: *This study focuses on exploring and evaluating the effectiveness of various Python libraries in analyzing body mass index (BMI) data to support obesity prediction and provide relevant health recommendations. In the context of obesity becoming a global health issue, the application of data science tools such as Python enhances the accuracy of assessments, predictions, and personalized advice. The research compares Python libraries across multiple functional groups: Data processing (Pandas, Dask), Computation (NumPy, SymPy), Visualization (Matplotlib, Seaborn, Plotly), Machine learning and prediction (Scikit-learn, TensorFlow, PyTorch), and Statistics (Statsmodels, SciPy). Using regression, clustering, and machine learning techniques, the research team developed a model to predict obesity based on BMI-related indicators integrated with an automated health recommendation system. This work not only provides a comprehensive overview of the Python ecosystem for medical data analysis but also contributes to improving health monitoring and obesity prevention through technological solutions.*

Keywords*: Python libraries, obesity level prediction, linear regression, technological solution*

---

[2] Hanoi Open University