

ỨNG DỤNG TRÍ TUỆ NHÂN TẠO DỰ ĐOÁN TÌNH TRẠNG DINH DƯỠNG CỦA THANH NIÊN VIỆT NAM

Bùi Hồng Ngọc^{1,✉}, Lê Minh Giang¹, Lê Thị Thanh Xuân¹
Lê Thị Hương¹, Bùi Anh Phong³, Nguyễn Thị Hương Giang¹
Nguyễn Thị Thu Liễu¹, Nguyễn Thanh Tùng²

¹Trường Đại học Y Hà Nội

²Trường Đại học CMC

³Công ty điểm sáng công nghệ

Nghiên cứu này nhằm xây dựng và đánh giá mô hình trí tuệ nhân tạo (AI) dự đoán tình trạng dinh dưỡng của thanh niên Việt Nam dựa trên dữ liệu lớn. Dữ liệu được thu thập từ khảo sát 12.042 sinh viên đại học trên toàn quốc. Các mô hình học máy hiện đại như XGBoost, LightGBM và CatBoost được ứng dụng để phân loại tình trạng dinh dưỡng theo chỉ số BMI. Mô hình XGBoost cho kết quả tốt nhất với độ chính xác đạt 75,3% và F1-score là 73,75%. Những biến đầu vào quan trọng trong dự đoán bao gồm vòng eo, vòng mông, thời gian ngủ, giới tính, thói quen ăn đêm, sử dụng thức ăn nhanh, và khẩu phần ăn hàng ngày (bao gồm năng lượng, protein, chất béo và vi chất dinh dưỡng). Kết quả cho thấy tiềm năng ứng dụng AI trong phân tích dữ liệu sức khỏe cộng đồng, góp phần phát hiện sớm các nguy cơ dinh dưỡng và xây dựng các khuyến nghị chuyên biệt về dinh dưỡng – thể lực cho nhóm thanh niên. Nghiên cứu góp phần mở rộng khả năng cá thể hóa các can thiệp y tế trong bối cảnh chuyển đổi số và y học chính xác tại Việt Nam.

Từ khóa: Trí tuệ nhân tạo, dinh dưỡng, thanh niên, BMI, dự đoán.

I. ĐẶT VẤN ĐỀ

Tình trạng thừa cân, béo phì, và suy dinh dưỡng ở thanh niên đang ngày càng trở thành vấn đề đáng lo ngại tại Việt Nam, đặc biệt trong nhóm sinh viên đại học – đối tượng dễ chịu ảnh hưởng từ thay đổi môi trường sống, thói quen ăn uống và lối sống ít vận động.¹ Các khảo sát gần đây tại Việt Nam cho thấy, tỉ lệ sinh viên có tình trạng dinh dưỡng không hợp lý (bao gồm cả gầy và thừa cân – béo phì) dao động từ 15% đến 30%, tùy theo vùng miền và đặc điểm ngành học.²⁻⁴ Tình trạng này không chỉ ảnh hưởng đến hiệu suất học tập và chất lượng cuộc sống, mà còn là yếu tố nguy cơ của

nhiều bệnh lý mạn tính không lây trong tương lai như đái tháo đường type 2, bệnh tim mạch, tăng huyết áp và rối loạn lipid máu.⁵

Hiện nay các chương trình can thiệp dinh dưỡng truyền thống tại Việt Nam chủ yếu tập trung vào trẻ em và phụ nữ mang thai, nhóm thanh niên – đặc biệt là sinh viên – vẫn chưa nhận được sự quan tâm tương xứng về mặt nghiên cứu cũng như hành động chính sách. Một trong những thách thức lớn khi xây dựng chương trình dự phòng và can thiệp cho nhóm này là thiếu các công cụ tiên lượng sớm, cá thể hóa và có thể áp dụng rộng rãi.

Với sự phát triển nhanh chóng của công nghệ dữ liệu và trí tuệ nhân tạo (AI), ngày càng có nhiều nghiên cứu quốc tế ứng dụng học máy (machine learning) để phân tích dữ liệu sức khỏe cộng đồng nhằm xây dựng mô hình dự đoán nguy cơ bệnh tật.⁶ Các thuật toán học

Tác giả liên hệ: Bùi Hồng Ngọc

Trường Đại học Y Hà Nội

Email: buihongngoc@hmu.edu.vn

Ngày nhận: 07/07/2025

Ngày được chấp nhận: 21/07/2025

máy như XGBoost, LightGBM, CatBoost đã chứng minh hiệu quả trong việc xử lý dữ liệu phức tạp, có nhiều chiều, nhiều kiểu định dạng và cho phép đưa ra dự đoán với độ chính xác cao vượt trội so với các phương pháp thống kê truyền thống. Tuy nhiên, tại Việt Nam, việc ứng dụng AI vào lĩnh vực dự báo tình trạng dinh dưỡng cho thanh niên còn rất hạn chế.

Trong bối cảnh đó, nghiên cứu này được thực hiện với mục tiêu xây dựng và đánh giá mô hình trí tuệ nhân tạo (AI) để dự đoán tình trạng dinh dưỡng của sinh viên Việt Nam, dựa trên dữ liệu nhân khẩu học, thói quen sinh hoạt, hành vi ăn uống và khẩu phần dinh dưỡng. Việc phát triển mô hình AI không chỉ góp phần đổi mới phương pháp đánh giá nguy cơ sức khỏe trong y tế công cộng, mà còn tạo tiền đề để xây dựng các hệ thống cảnh báo sớm và tư vấn dinh dưỡng cá thể hóa – hướng đến tăng cường sức khỏe toàn diện cho thanh niên Việt Nam trong thời đại số.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Đối tượng

Dữ liệu khảo sát 12.042 sinh viên từ 10 trường đại học tại 3 miền Bắc, Trung, Nam của Việt Nam (Trường Đại học Y Hà Nội, Trường Đại học Kinh tế - Đại học Quốc gia Hà Nội, Trường Đại học Bách khoa Hà Nội, Trường Đại học Xây dựng Hà Nội, Học viện Ngân hàng Hà Nội, Đại học Nông Lâm Thái Nguyên, Đại học Sư phạm Đà Nẵng, Đại học Thủ Dầu Một, Đại học Sư phạm Thành phố Hồ Chí Minh, Học viện Thanh thiếu niên Việt Nam). Tại thời điểm xây dựng mô hình, dữ liệu từ Trường Đại học Luật Hà Nội chưa được hoàn tất nên không được đưa vào phân tích. Dữ liệu được thu thập bằng cách sinh viên tự khai báo qua biểu mẫu trực tuyến (website).

2. Phương pháp

Nghiên cứu áp dụng các phương pháp học

máy (machine learning) để xây dựng mô hình trí tuệ nhân tạo (AI) có khả năng phân loại tình trạng dinh dưỡng (BMI) của sinh viên Việt Nam. Ba thuật toán chính được sử dụng gồm: XGBoost, LightGBM, CatBoost – đây là những mô hình dạng cây tăng cường (boosting trees) hiệu quả trong xử lý dữ liệu lớn, nhiều biến và có độ phức tạp cao. Các mô hình được huấn luyện và đánh giá dựa trên tập dữ liệu khảo sát quy mô lớn (12.042 sinh viên đại học trên toàn quốc), với các chỉ số đánh giá bao gồm: độ chính xác (accuracy), độ tin cậy (precision) và F1-score.

Biến số

Biến phụ thuộc: Tình trạng dinh dưỡng được phân loại dựa trên chỉ số khối cơ thể (BMI) theo tiêu chuẩn của Tổ chức Y tế Thế giới (WHO) dành cho khu vực Châu Á:

Gầy: $BMI < 18,5 \text{ kg/m}^2$

Bình thường: $18,5 \leq BMI < 23 \text{ kg/m}^2$

Thừa cân: $23 \leq BMI < 25 \text{ kg/m}^2$

Béo phì độ I: $25 \leq BMI < 30 \text{ kg/m}^2$

Béo phì độ II: $BMI \geq 30 \text{ kg/m}^2$

Biến độc lập: 180 biến đầu vào bao gồm:

Thông tin nhân khẩu học: giới tính, nơi cư trú, dân tộc, ngành học...

Hành vi và thói quen ăn uống: thói quen ăn sáng/ăn đêm, tần suất sử dụng các loại thực phẩm...

Khẩu phần ăn: năng lượng, protein, lipid, glucid, vi chất dinh dưỡng...

Lối sống và sinh hoạt: mức độ hoạt động thể lực, thời gian ngủ trong ngày, thời gian ngồi trong ngày...

Thời gian nghiên cứu: từ 8/2024 đến 2/2025.

Xử lý và phân tích dữ liệu

Làm sạch dữ liệu: loại bỏ các biến có tỉ lệ thiếu > 15%, điền giá trị còn thiếu bằng trung bình, loại bỏ các giá trị ngoại lai và sai lệch rõ rệt.

Chia tập dữ liệu: Tập huấn luyện (Train) 80%, Tập kiểm định (Validation) 10%, Tập kiểm tra (Test) 10%.

Ngôn ngữ và công cụ lập trình: Sử dụng Python 3.10.16 với các thư viện học máy: Scikit-learn, XGBoost, LightGBM, CatBoost để huấn luyện mô hình.

Quy trình xây dựng mô hình học máy dự báo tình trạng dinh dưỡng

Quy trình xây dựng mô hình học máy dự báo tình trạng dinh dưỡng được thực hiện theo 6 bước:

Bước 1: Làm sạch và khai phá dữ liệu.

Bước 2: Đề xuất các mô hình học máy phù

$$\text{Accuracy} = \frac{\text{Số dự đoán đúng}}{\text{Tổng số mẫu}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Trong đó:

TP (True Positive): Dương tính đúng – Dự đoán đúng khi mẫu thực sự là dương tính.

TN (True Negative): Âm tính đúng – Dự đoán đúng khi mẫu thực sự là âm tính.

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

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

Độ tin cậy: trong số các mẫu mà mô hình dự đoán là dương tính, có bao nhiêu mẫu thực sự là dương tính. Độ tin cậy cao tức là tổng số các mẫu mà mô hình dự đoán dương tính phần lớn là đúng.

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

F1 score: là chỉ số tổng hợp giữa độ tin cậy và độ nhạy, giúp đánh giá hiệu quả thực sự của mô hình thay vì chỉ nhìn từng chỉ số riêng lẻ.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong đó: Độ nhạy (Recall) được tính bằng công thức:

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

hợp dựa vào việc khai phá dữ liệu.

Bước 3: Lựa chọn các biến đầu vào ban đầu.

Bước 4: Huấn luyện các mô hình.

Bước 5: Lựa chọn mô hình tốt nhất với các biến đầu vào ý nghĩa.

Bước 6: Kiểm thử sản phẩm.

Đánh giá mô hình học máy dự báo tình trạng dinh dưỡng

Sử dụng các chỉ số như độ tin cậy, độ chính xác, F1-score.

Độ chính xác: là tỷ lệ dự đoán đúng (cả dương tính và âm tính) trên tổng số mẫu. Độ chính xác cao tức là mô hình dự đoán đúng phần lớn các mẫu.

3. Đạo đức nghiên cứu

Toàn bộ dữ liệu dùng để xây dựng mô hình được lấy từ nghiên cứu “Ứng dụng sử dụng khoa học dữ liệu lớn và trí tuệ nhân tạo trong việc xây dựng mô hình can thiệp chăm sóc dinh dưỡng tối ưu và nâng cao thể lực cho thanh niên Việt Nam” được phê duyệt bởi Hội đồng đạo đức y sinh của Trường Đại học Y Hà Nội (số 991/GCN-HMUIRB ngày 30/10/2023).

III. KẾT QUẢ

1. Xây dựng mô hình học máy dự báo tình trạng dinh dưỡng

Làm sạch và khai phá dữ liệu

Dữ liệu để xây dựng mô hình là bộ dữ liệu thu thập online từ 12.042 sinh viên của 10 trường đại học trên cả nước. Việc kiểm soát độ tin cậy của các giá trị khai báo online được thực hiện từ khi xây dựng website thu thập số liệu như thiết kế bộ câu hỏi tốt, mỗi số điện thoại di động chỉ được đăng nhập và trả lời một lần nên tránh được tình huống một người làm hộ cho nhiều người, các câu hỏi quan trọng đều để yêu cầu bắt buộc phải điền mới submit được, các câu hỏi định lượng được giới hạn khoảng nhập dữ

liệu theo điều kiện định sẵn (ví dụ tuổi được giới hạn từ 16 - 30)... Ngoài ra số liệu điều tra online được so sánh với số liệu đo thực tế (ví dụ chiều cao, cân nặng tự khai được so sánh với chiều cao, cân nặng được điều tra viên đo)... Bộ dữ liệu thô được làm sạch và lọc dữ liệu để loại bỏ các dữ liệu nhiễu và lỗi không hợp lệ. Dữ liệu thiếu được xử lý bằng các phương pháp như: điền giá trị trung bình, trung vị, loại bỏ cột có trên 15–25% dữ liệu thiếu. Quá trình giúp làm sạch dữ liệu và đảm bảo dữ liệu đúng định dạng để đưa vào mô hình học máy.

Lọc các giá trị không xác định: Đầu tiên, tiến hành loại bỏ những giá trị chuỗi trống (») hoặc «undefined» xuất hiện trong bộ dữ liệu. Tuy nhiên, lượng không xác định này khá nhiều, nên team AI đã thử một số chiến thuật điền dữ liệu còn thiếu: điền số 0, điền giá trị trung bình, trung vị, điền giá trị dựa trên các điểm dữ liệu lân cận... Điều này đảm bảo rằng dữ liệu còn lại không có các giá trị không rõ ràng, để có thể đưa vào mô hình học máy.

Loại bỏ các hàng thiếu các giá trị quan trọng: Các hàng thiếu bất kỳ giá trị nào trong các cột height, weight, waist, và hip đều bị loại bỏ. Đây là các cột chứa thông tin quan trọng cần thiết cho quá trình tính toán chỉ số cơ thể (BMI) và các đặc tính liên quan khác. Việc giữ lại chỉ những hàng có đầy đủ các giá trị này giúp đảm bảo tính toàn vẹn và độ chính xác của bộ dữ liệu, đồng thời tránh việc phải sử dụng phương pháp nội suy hoặc ước lượng không cần thiết.

Loại bỏ các cột thiếu trên 15% dữ liệu: Những cột có hơn 15% giá trị bị thiếu sẽ bị loại bỏ khỏi bộ dữ liệu. Tỷ lệ 15% được chọn nhằm loại trừ các cột chứa nhiều giá trị thiếu, vì chúng có thể làm giảm hiệu quả của mô hình khi sử dụng. Việc giữ lại chỉ các cột có dữ liệu đầy đủ sẽ giúp mô hình đạt được độ chính xác cao hơn.

Tính giá trị trung bình hàng ngày cho các

cột dinh dưỡng: Đối với các cột liên quan đến dinh dưỡng, chúng tôi tính toán giá trị trung bình hàng ngày dựa trên dữ liệu ban đầu. Mỗi hàng dữ liệu sau khi tính trung bình sẽ đại diện cho lượng dinh dưỡng trung bình mà đối tượng đã tiêu thụ trong ngày, giúp chuẩn hóa và đơn giản hóa dữ liệu dinh dưỡng để sử dụng trong phân tích.

Sau đó, các kỹ thuật khai phá dữ liệu được áp dụng để hiểu rõ cấu trúc, phân bố cũng như các mối quan hệ tiềm ẩn giữa các biến số, nhằm đưa ra những nhận định ban đầu về dữ liệu, phục vụ cho việc lựa chọn các mô hình phù hợp.

Để xử lý và làm sạch dữ liệu nhóm nghiên cứu đã sử dụng thư viện NumPy - version 1.24.3 và Pandas - version 1.5.3 sau đó trực quan hoá dữ liệu bằng thư viện Matplotlib - Version 3.9.2 và Seaborn - Version 0.13.2.

Đề xuất các mô hình học máy phù hợp dựa vào việc khai phá dữ liệu

Vì dữ liệu dạng bảng, trong đó bao gồm rất nhiều trường với các thang đo khác nhau, các mô hình dạng cây là lựa chọn tối ưu bởi:

Mô hình cây có khả năng xử lý số liệu với thang đo khác nhau mà không cần phải tiền xử lý.

Tính giải thích được của mô hình cây: mô hình cây có thể được đem ra phân tích để xác định độ quan trọng của các biến.

Xử lý được nhiễu: ở một mức nhiễu cho phép, mô hình cây có khả năng tự loại bỏ các biến dữ liệu đầu vào bị nhiễu, bằng cách không tập trung vào các biến dữ liệu đầu vào này.

Trong khi đó, nếu sử dụng một số loại mô hình phổ biến khác, điển hình có thể kể đến mô hình Mạng nơ-ron, các mô hình này sẽ tồn tại rất nhiều nhược điểm:

Buộc phải tiền xử lý dữ liệu cẩn thận: tất cả các cột dữ liệu phải chuẩn hóa về cùng một thang đo, chẳng hạn như từ 0 đến 1.

Mô hình mạng nơ-ron thường được gọi là một “hộp đen” - rất khó để phân tích và giải thích kết quả của mạng nơ-ron.

Mô hình dạng cây cơ bản nhất là cây quyết định (Decision Tree). Ở đó, mô hình sẽ xây dựng một cây để đưa ra dự đoán. Tại mỗi lá, cây sẽ sử dụng một biến đầu vào để tách thành các nhánh nhỏ hơn. Để tăng độ chính xác cho mô hình người ta xây dựng nhiều cây hơn và tính trung bình. Rừng ngẫu nhiên (Random Forest): bằng cách xây dựng nhiều cây quyết định để tạo thành một rừng cây, rừng ngẫu nhiên có thể tăng độ chính xác đáng kể, bằng cách tăng sự đa dạng và giảm thiểu sai lệch. XGBoost, LightGBM, CatBoost cải tiến thuật toán của rừng ngẫu nhiên: các cây được xây dựng sau sẽ học từ các cây được xây dựng từ trước, để từ đó tạo ra một rừng cây tối ưu.

Vì vậy 3 thuật toán XGBoost, LightGBM, CatBoost được chọn để xây dựng mô hình.

Lựa chọn các biến đầu vào ban đầu

Việc lựa chọn biến đầu vào được kết hợp giữa thuật toán AI và ý kiến của chuyên gia dinh dưỡng. Đầu tiên, chạy mô hình với 180 biến đầu vào, sau đó lựa chọn các biến đầu vào dựa theo độ quan trọng trong mô hình và ý kiến chuyên gia dinh dưỡng.

Độ quan trọng dựa vào mức độ đóng góp của biến đó đối với giá trị đầu ra của mô hình và được tính bằng Importance Score. Biến có Importance Score càng lớn thì biến đó giúp tăng độ chính xác của mô hình càng nhiều,

Importance Score được thuật toán học máy tính toán ra từ việc huấn luyện mô hình.

Chuyên gia dinh dưỡng mô hình xác giáo sư, phó giáo sư, tiến sĩ của bộ môn Dinh dưỡng và An toàn thực phẩm, Viện Đào tạo Y học dự phòng & Y tế công cộng, Trường Đại học Y Hà Nội. Là những người có chuyên môn sâu về dinh dưỡng, đều có thâm niên công tác trong lĩnh vực dinh dưỡng trên 10 năm.

Sau rất nhiều lần dùng thuật toán AI và ý kiến chuyên gia dinh dưỡng, 38 biến quan trọng được lựa chọn làm biến số đầu vào của mô hình gồm các nhóm biến số nhân khẩu học, nhân trắc học, khẩu phần ăn, hành vi ăn uống và lối sống.

Huấn luyện các mô hình

Sử dụng tập dữ liệu train tiến hành huấn luyện các mô hình XGBoost, LightGBM, CatBoost với các tham số khác nhau để tìm ra được mô hình ý nghĩa và mang lại kết quả tốt nhất. Sau đó nhóm nghiên cứu sử dụng tập dữ liệu validation để kiểm tra xem các mô hình có bị mắc các lỗi ảnh hưởng xấu đến hiệu suất không.

Lựa chọn mô hình tốt nhất với các biến đầu vào ý nghĩa

Sau nhiều lần chạy thử nghiệm, trong ba mô hình đã xây dựng thì mô hình XGBoost mang lại kết quả đánh giá tốt nhất được lựa chọn để dự đoán cho các tập dữ liệu mới. Có 11 biến số đóng góp không nhiều cho mô hình cũng bị loại bỏ, cuối cùng nhóm biến số đầu vào quan trọng nhất được lựa chọn gồm 27 biến số (bảng 1):

Bảng 1. Các biến số được lựa chọn đưa vào mô hình dự đoán tình trạng dinh dưỡng

TT	Nhóm biến	Biến số được lựa chọn
1	Nhân khẩu học	Giới tính
2	Nhân trắc học	Vòng eo Vòng mông

TT	Nhóm biến	Biến số được lựa chọn
3	Hành vi ăn uống và lối sống	Tần suất ăn khuya Tần suất sử dụng thức ăn béo Tần suất ăn trái cây Tần suất ăn hải sản Thói quen sử dụng các ứng dụng trên Smart phone để cải thiện tình trạng rèn luyện thể lực Thời gian ngủ trung bình mỗi ngày Tự đánh giá hành vi ăn uống trong 1 năm quan Áp dụng chế độ ăn tăng cân hay không Thời gian tập thể dục hàng ngày bao nhiêu phút thì tốt cho sức khỏe Di chuyển bằng phương tiện cơ giới trong 7 ngày qua.
4	Khẩu phần ăn	Năng lượng Potein Lipid Glucid Chất xơ Vitamin A Vitamin C Sắt Canxi Natri Kali Kẽm Beta_caroten Cholesterol

Kiểm thử sản phẩm

Kiểm thử hệ thống và tính chính xác của mô hình đã chọn được tiến hành trên tập dữ liệu test. Việc kiểm thử bao gồm thử nghiệm tính ổn định và đảm bảo mô hình hoạt động một cách bình thường, kết quả dự đoán ở mức hợp lý.

Qua đó, những lỗi phát sinh sẽ được phát hiện và sửa chữa nhằm đảm bảo sản phẩm hoạt động chính xác, hiệu quả trước khi đưa vào sử dụng rộng rãi.

2. Đánh giá mô hình học máy dự báo tình trạng dinh dưỡng

Bảng 2. Chỉ số đánh giá mô hình học máy dự báo tình trạng dinh dưỡng

Mô hình	Độ chính xác (Accuracy)	Độ tin cậy (Precision)	F1 Score
XGBoost	75,3%	72,7%	73,75%
LightGBM	69,6%	65,9%	67,5%
CatBoost	72,5%	68,7%	70,1%

Trong ba mô hình học máy được so sánh: XGBoost, LightGBM và CatBoost, mô hình XGBoost cho kết quả tốt nhất, với độ chính xác, độ tin cậy và F1-score cao hơn các mô hình còn lại.

IV. BÀN LUẬN

AI và học máy đang ngày càng được ứng dụng rộng rãi trong lĩnh vực y tế, từ chẩn đoán hình ảnh, sàng lọc ung thư đến dự báo bệnh không lây nhiễm (NCDs) và hành vi sức khỏe.⁶ Trong nghiên cứu này, AI được ứng dụng như một công cụ hỗ trợ dự đoán nhanh chóng, chi phí thấp và có thể mở rộng, đặc biệt phù hợp với bối cảnh nguồn lực y tế hạn chế tại Việt Nam. Với độ chính xác hơn 75%, độ tin cậy gần 73%, mô hình có thể giúp phát hiện sớm các nhóm có nguy cơ dinh dưỡng để thực hiện các can thiệp sớm và kịp thời, thay vì đợi đến khi các hậu quả sức khỏe xảy ra.

Một điểm đặc biệt quan trọng là việc mô hình dựa trên các biến số hành vi và dinh dưỡng tự khai báo, giúp giảm chi phí thu thập dữ liệu phức tạp. Chẳng hạn, biến "tần suất ăn đồ béo", "thói quen ăn đêm", hoặc "sử dụng thiết bị, ứng dụng trên Smart phone để cải thiện tình trạng rèn luyện thể lực" có khả năng giải thích cao trong mô hình, cho thấy rằng các hành vi đơn giản có thể là yếu tố dự báo mạnh mẽ về tình trạng sức khỏe.

Kết quả nghiên cứu này tương đồng và có phần vượt trội so với một số nghiên cứu tương tự trong khu vực. Nghiên cứu của Hengyan Liu và cộng sự tại Hồng Kông (2024) sử dụng thuật toán XGBoost, Random Forest, SVM... để dự đoán nhẹ cân, thừa cân, béo phì ở thanh thiếu niên, đạt độ chính xác chung từ 70 - 74%.⁷ Một nghiên cứu tại Thái Lan sử dụng C5.0 để dự đoán tình trạng dinh dưỡng ở sinh viên, độ chính xác chỉ đạt 60,78%.⁸ Mô hình XGBoost của nghiên cứu này không chỉ cho độ chính xác

cao hơn mà còn tích hợp được nhiều chiều dữ liệu (thói quen ăn uống, khẩu phần, thời gian ngủ, thời gian ngồi, hành vi sử dụng thiết bị công nghệ) – vốn chưa được tích hợp đầy đủ trong các nghiên cứu trước đó.

Một số biến như vòng eo, vòng hông, thói quen ăn đêm, thời gian ngủ, thời gian ngồi, và mức tiêu thụ năng lượng trung bình là những chỉ báo mạnh trong mô hình. Đây là những yếu tố đã được nhiều tài liệu y khoa ghi nhận là liên quan chặt chẽ tới hội chứng chuyển hóa và nguy cơ thừa cân béo phì ở người trẻ tuổi.⁹⁻¹²

Vì vậy mô hình không chỉ mang tính học thuật mà còn có khả năng ứng dụng thực tế cao, nhất là khi được tích hợp vào các nền tảng số như ứng dụng điện thoại thông minh, website chăm sóc sức khỏe cá nhân, hoặc hệ thống y tế dự phòng. Dựa trên các thông tin nhập vào, hệ thống có thể đưa ra cảnh báo sức khỏe cá thể hóa, gợi ý điều chỉnh hành vi (ví dụ: tăng thời gian ngủ, giảm thức ăn nhanh, tăng vận động nhẹ...).

V. KẾT LUẬN

Việc ứng dụng AI, đặc biệt là mô hình XGBoost, có khả năng hỗ trợ đắc lực trong việc dự đoán và quản lý nguy cơ dinh dưỡng của thanh niên Việt Nam. Kết quả mô hình cho thấy hiệu quả cao và có thể triển khai rộng rãi trong các nền tảng số.

TÀI LIỆU THAM KHẢO

1. Avram C, Nyulas V, Onisor D, et al. Food Behavior and Lifestyle Among Students: The Influence of the University Environment. *Nutrients*. 2025;17(1):12. doi:10.3390/nu17010012
2. Vũ Thị Nhung. Tình trạng dinh dưỡng và một số chỉ số sinh học của sinh viên đại học chính quy năm thứ 2 trường đại học điều dưỡng nam định. *Tạp chí Y học Việt Nam*. 2022;516(1). doi:10.51298/vmj.v516i1.2963
3. Nguyễn Thành Chung, Trần Ngọc Anh,

Nguyễn Diệp Anh, và cs. Tình trạng dinh dưỡng của sinh viên năm nhất Phân hiệu Trường Đại học Y Hà Nội tại Thanh Hóa năm học 2023 - 2024. *Tạp chí Khoa học Điều dưỡng*. 2024;7(5):148-153. doi:10.54436/jns.2024.05.871

4. Nguyễn Trung Hiếu, Trần Khánh Thu, Vũ Duy Tùng. Một số chỉ số nhân trắc dinh dưỡng của sinh viên chính quy năm thứ nhất ngành y khoa và dược học trường đại học y dược thái bình năm 2023. *Tạp chí Dinh dưỡng và Thực phẩm*. 2023;19(4+5):54-61. doi:10.56283/1859-0381/614

5. Mousavi S, Bieber K, Zirpel H, et al. Large-scale analysis highlights obesity as a risk factor for chronic, non-communicable inflammatory diseases. *Front Endocrinol*. 2025;16:1516433. doi:10.3389/fendo.2025.1516433

6. Alhumaidi NH, Dermawan D, Kamaruzaman HF, et al. The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review. *JMIR Med Inform*. 2025;13(1):e68898. doi:10.2196/68898

7. Liu H, Wu YC, Chau PH, et al. Prediction of adolescent weight status by machine learning: a population-based study. *BMC Public Health*. 2024;24(1):1351. doi:10.1186/s12889-024-18830-1

8. Yusamran N, Chaitokkia S, Yusamran M. A Study of Factors Associated with Nutritional Status of Loei Rajabhat University Students by Machine Learning. *J Vongchavalitkul Univ*. 2024;37(2):96-113.

9. Taheri S, Lin L, Austin D, et al. Short Sleep Duration Is Associated with Reduced Leptin, Elevated Ghrelin, and Increased Body Mass Index. *PLOS Med*. 2004;1(3):e62. doi:10.1371/journal.pmed.0010062

10. Yoshida J, Eguchi E, Nagaoka K, et al. Association of night eating habits with metabolic syndrome and its components: a longitudinal study. *BMC Public Health*. 2018;18:1366. doi:10.1186/s12889-018-6262-3

11. Thorp AA, Owen N, Neuhaus M, et al. Sedentary Behaviors and Subsequent Health Outcomes in Adults: A Systematic Review of Longitudinal Studies, 1996–2011. *Am J Prev Med*. 2011;41(2):207-215. doi:10.1016/j.amepre.2011.05.004

12. Lopez-Lopez JP, Cohen DD, Ney-Salazar D, et al. The prediction of Metabolic Syndrome alterations is improved by combining waist circumference and handgrip strength measurements compared to either alone. *Cardiovasc Diabetol*. 2021;20(1):68. doi:10.1186/s12933-021-01256-z

Summary

APPLICATION OF ARTIFICIAL INTELLIGENCE IN PREDICTING THE NUTRITIONAL STATUS OF VIETNAMESE YOUTH

This study aims to develop and evaluate an artificial intelligence (AI) model for predicting the nutritional status of Vietnamese youth based on large-scale data. The dataset was derived from a nationwide survey involving 12,042 university students. Advanced machine learning algorithms, including XGBoost, LightGBM, and CatBoost, were applied to classify nutritional status using the Body Mass Index (BMI). Among these, the XGBoost model achieved the highest performance, with an accuracy of 75.3% and an F1-score of 73.75%. Key predictive variables included waist circumference, hip circumference, sleep duration, gender, late-night eating habits, fast food consumption, and daily dietary intake (energy, protein, fat, and micronutrients). The findings highlight the potential of AI in public health data analysis, particularly in identifying nutritional risks and generating personalized recommendations for diet and physical activity among youth. This research contributes to the development of precision nutrition approaches and supports the broader integration of digital health technologies in Vietnam's public health strategies.

Keywords: Artificial intelligence, nutrition, youth, BMI, prediction.