PHÂN LOẠI MỨC ĐỘ BÉO PHÌ DỰA TRÊN THÓI QUEN ĂN UỐNG VÀ TÌNH TRẠNG THỂ CHẤT

Phạm Đức Đại - 21130304

Võ Quốc Phong - 21130474

Khoa Công nghệ thông tin, Đại học Nông Lâm TP.HCM

TÓM TẮT

Dự án này tập trung vào việc phân loại mức độ béo phì bằng cách sử dụng các kỹ thuật học máy. Để giải quyết bài toán này, chúng tôi đã sử dụng 3 thuật toán chính: Logistics Regression, SVM, Random Forest. Dataset bao gồm 2111 mẫu và 16 đặc trưng.

Quá trình thực hiện bao gồm việc tiền xử lý dữ liệu, huấn luyện các mô hình và đánh giá hiệu quả của từng thuật toán. Kết quả cho thấy mô hình Random Forest đạt độ chính xác cao nhất trong số các phương pháp đã sử dụng.

1. GIỚI THIỆU

Béo phì là một trong những vấn đề sức khỏe nghiêm trọng nhất toàn cầu, với tỉ lệ gia tăng đáng kể trong vài thập kỷ qua, gây hậu quả như tăng nguy cơ bệnh tim mạch, đái tháo đường, và ung thư [1] .

Thói quen ăn uống và tình trạng thể chất ảnh hưởng mạnh mẽ đến nguy cơ béo phì [2]. Việc phân loại mức độ béo phì dựa trên các yếu tố này không chỉ cung cấp cái nhìn tổng quát mà còn mở ra giải pháp kiểm soát hiệu quả [5].

Tầm quan trọng và ứng dụng thực tế

1. **Tầm soát nguy cơ:** Các hệ thống dựa trên học máy giúp nhận diện người có nguy cơ béo phì cao, hỗ trợ phòng ngừa [6].
2. **Cá nhân hóa trị liệu:** Xây dựng liệu trình dựa trên dữ liệu thói quen dinh dưỡng và lối sống cá nhân [3].
3. **Hỗ trợ ra quyết định:** Cung cấp công cụ cho chính phủ phát triển chính sách dinh dưỡng.

Dựa trên dataset "Estimation of obesity levels based on eating habits and physical condition" từ UCI Machine Learning Repository [4], dự án áp dụng các phương pháp sau để giải quyết bài toán phân loại mức độ béo phì:

1. **Logistic Regression**: Sử dụng để phân tích mối quan hệ tuyến tính giữa các biến đầu vào như thói quen ăn uống (tần suất ăn nhanh, tiêu thụ rau củ) và tình trạng thể chất (mức độ hoạt động thể dục) với mức độ béo phì.
2. **Random Forest**: Kết hợp nhiều cây quyết định để xử lý các đặc điểm phi tuyến tính trong dataset, giúp cải thiện hiệu suất và độ chính xác của phân loại.
3. **Support Vector Machine (SVM)**: Ứng dụng để phân loại với các kernel phi

tuyến nhằm tối ưu hóa độ chính xác trên các tập dữ liệu có ranh giới phức tạp.

2. CÔNG TRÌNH LIÊN QUAN

Nghiên cứu 1: "Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits" [7]

Dữ liệu sử dụng 2009 mẫu dữ liệu với 16 đặc trưng.

Các phương pháp được sử dụng trong nghiên cứu: Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC) để khắc phục vấn đề imbalance (mất cân bằng) lớp. Sử dụng Recursive Feature Elimination (RFE) để chọn các đặc trưng có liên quan nhất từ một tập dữ liệu nhất định, nhằm cải thiện hiệu suất của mô hình bằng cách giảm chiều của đặc trưng. Tối ưu siêu tham số (hyperparameter) sử dụng trong trong các mô hình bằng kỹ thuật Bayesian

Các thuật toán được sử dụng là Logistic Regression, Random Forest, và Support Vector Machine với các số liệu sử dụng để đánh giá hiệu suất của các mô hình: Accuracy, recall, precision, F1-score, AUC, precision–recall curve evaluation.

Bài nghiên cứu chỉ ra rằng mô hình Logistic Regression cho kết quả tốt nhất. Việc lựa chọn đặc trưng đã nâng cao hiệu quả của mô hình. Củng cố cho khả năng ứng dụng vào thực tế.

Nghiên cứu 2: "Estimation of obesity levels based on dietary habits and condition physical using computational intelligence" [8]

Dữ liệu sử dụng 2111 mẫu dữ liệu với 16 đặc trưng.

Các phương pháp được sử dụng trong nghiên cứu: sử dụng hàm MaxAbsScaler để cân bằng đặc trưng và tính toán với MeanImputerc để thay thế các giá trị bị thiếu bằng giá trị trung bình của đặc trưng. Sử dụng ma trận tương quan (Correlation matrix).

Các thuật toán được sử dụng là Light Gradient Boosting Machine classifier, Random Forest , Decision Tree , Extremely Randomized Trees và Logistic Regression với các số liệu sử dụng để đánh giá hiệu suất của các mô hình: Accuracy, recall, precision, F1-score, AUC.

Bài nghiên cứu chỉ ra rằng mô hình phân loại LightGBM có giá trị AUC có trọng số (0,9990) cho kết quả tốt nhất. Các kết quả được trình bày có thể hữu ích để phân tích tính phù hợp của các phương pháp dựa trên tính toán thông minh để nghiên cứu các bệnh khác nhau, phát hiện chúng một cách đầy đủ và giảm thiểu tác động đến xã hội.

Nghiên cứu 3: "Machine learning Techniques to Predict Overweight or Obesity" [9]

Dữ liệu được thu thập qua cuộc khảo sát với 16 câu hỏi được đưa ra tạo thành 16 đặc trưng sử dụng trong bài nghiên cứu.

Các phương pháp được sử dụng trong nghiên cứu: cân bằng tập dữ liệu, chuyển đổi dữ liệu với các dữ liệu không phải số với kỹ thuật mã hóa one-hot, mã hóa ký tự và mã hóa label (nhãn). Chuẩn hóa MIN-MAX chia tỉ lệ trong phạm vi [0 - 1] hoặc [0.1 - 1.0]

Các thuật toán được sử dụng là Random Forest , Decision Tree, Support Vector Machines, K-Nearest Neighbors, Gaussian Naive Bayes, Multilayer Perceptron, Gradient Boosting và Extreme Gradient Boosting.

Sử dụng K-fold Cross Validation để đánh giá mô hình học máy, tính toán confusion matrix với accuracy, precision, recall, f1-score, tối ưu siêu tham số cho mô hình.

Bài nghiên cứu cho thấy hiệu suất của mô hình Random Forest là hiệu quả nhất. Nghiên cứu cho rằng các mô hình được phát triển đã chứng minh được học máy là một công cụ mạnh mẽ có thể được sử dụng trong lĩnh vực y tế để đưa ra quyết định điều trị kịp thời cho những người có nguy cơ béo phì.

3. PHÁT BIỂU BÀI TOÁN

3.1. Bài toán

Trong nghiên cứu này, tập trung giải quyết bài toán phân loại mức độ béo phì dựa trên các dữ liệu về thói quen ăn uống và tình trạng thể chất. Đây là một bài toán phân loại thuộc nhóm học có giám sát (Supervised Learning),

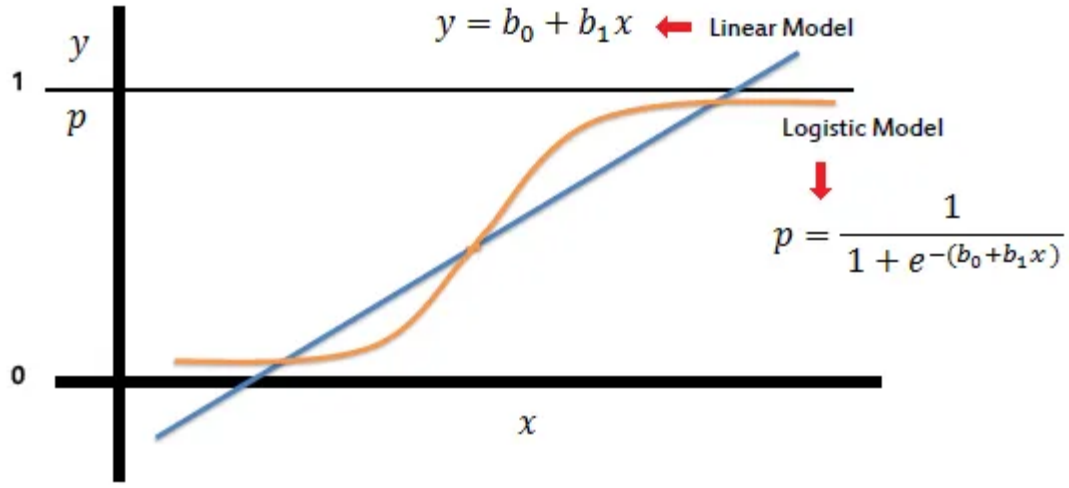
Inputs: Bộ dữ liệu bao gồm 16 thuộc tính đặc trưng liên quan đến thói quen ăn uống và tình trạng thể chất gồm: Gender, Age, Height, Weight, Family history with overweight, FAVC (Frequent Consumption of High Caloric Food), FCVC (Frequency of Consumption of Vegetables), NCP (Number of Main Meals), CAEC (Consumption of Food Between Meals), SMOKE, CH2O (Daily Water Consumption), SCC (Calories Monitoring), FAF (Frequency of Physical Activity), TUE (Time Using Technology Devices), CALC (Consumption of Alcohol), MTRANS (Transportation Method).

Outputs: Nhãn đầu ra là một trong các mức độ béo phì, được phân loại thành 7 nhóm: Insufficient Weight (Thiếu cân), Normal Weight (Cân nặng bình thường), Overweight Level I (Thừa cân cấp 1), Overweight Level II (Thừa cân cấp 2), Obesity Type I (Béo phì loại 1), Obesity Type II (Béo phì loại 2), Obesity Type III (Béo phì loại 3).

3.2. Thuật toán

3.2.1. Logistic Regression

Hồi quy Logistic là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm. Hồi quy Logistic làm việc dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành xác suất thuộc về một trong hai lớp nhị phân. Tuy nhiên, với một chút mở rộng và chất xám, logistic regression có thể dễ dàng được sử dụng cho vấn đề phân loại nhiều lớp.

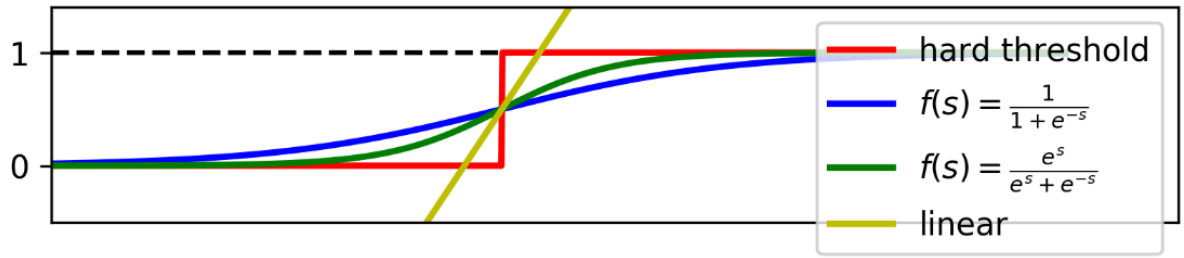


Logistic Regression sử dụng hàm phi tuyến để xác định xác suất của hai lớp 0 và 1.

Đầu ra dự đoán của logistic regression thường được viết chung dưới dạng:

$$f(\mathbf{x}) = \theta(\mathbf{w}^T \mathbf{x})$$

Trong đó θ được gọi là logistic function. Một số activation cho mô hình tuyến tính được cho trong hình dưới đây



Các activation function khác nhau.

Hồi quy Logistic hoạt động dựa trên hàm Sigmoid, được biểu diễn như sau:

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

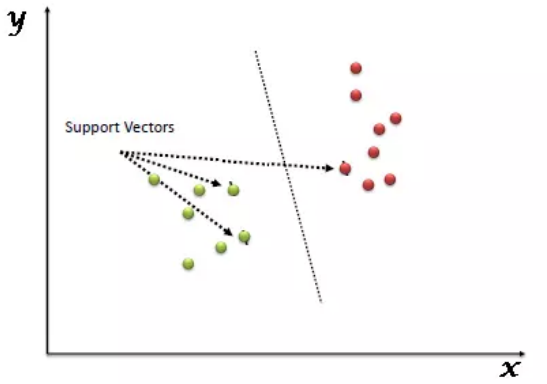
Hàm Sigmoid nhận đầu vào là một giá trị z bất kỳ, và trả về đầu ra là một giá trị xác suất nằm trong khoảng [0,1]. Khi áp dụng vào mô hình Hồi quy Logistic với đầu vào là ma trận dữ liệu X và trọng số w, ta có z = Xw.

Việc huấn luyện của mô hình là tìm ra bộ trọng số w sao cho đầu ra dự đoán của hàm Sigmoid gần với kết quả thực tế nhất. Để làm được điều này, ta sử dụng hàm mất mát (Loss Function) để đánh giá hiệu năng của mô hình. Mô hình càng tốt khi hàm mất mát càng nhỏ.

Hàm mất mát (Loss Function) là một hàm số được sử dụng để đo lường mức độ lỗi mà mô hình của chúng ta tạo ra khi dự đoán các kết quả từ dữ liệu đầu vào. Trong bài toán Hồi quy Logistic, chúng ta sử dụng hàm mất mát Cross-Entropy (còn gọi là Log Loss) để đánh giá hiệu năng của mô hình.

3.2.2. SVM (Support Vector Machine).

SVM (Support Vector Machine) là 1 thuật toán học máy thuộc nhóm Supervised Learning (học có giám sát) được sử dụng trong các bài toán phân lớp dữ liệu (classification) hay hồi qui (Regression). Trong thuật toán này, chúng ta vẽ đồ thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "siêu phẳng" (*hyper-plane*) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.



Margin là khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng với các phân lớp.SVM luôn cố gắng cực đại hóa margin này, từ đó thu được một siêu phẳng tạo khoảng cách xa nhất với các mẫu dữ liệu. Nhờ vậy, SVM có thể giảm thiểu việc phân lớp sai (misclassification) đối với điểm dữ liệu mới đưa vào.

Các bước thực hiện:

1. **Nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau**: Vẽ đồ thị dữ liệu là các điểm trong n chiều với giá trị của mỗi tính năng là một phần liên kết.
2. **Tìm siêu phẳng (hyper-plane) phân chia các lớp**: Siêu phẳng là đường thẳng phân chia các lớp ra thành hai phần riêng biệt.
3. **Tối đa hóa margin**: Margin là khoảng cách giữa siêu phẳng đến hai điểm dữ liệu gần nhất tương ứng với các phân lớp. SVM cố gắng cực đại hóa margin để giảm thiểu việc phân lớp sai.

3.2.3. Random Forest.

Rừng ngẫu nhiên hay Cây quyết định ngẫu nhiên là một nhóm gồm các cây quyết định hợp tác làm việc cùng nhau để cung cấp một đầu ra duy nhất.Thuật toán Random Forest là một kỹ thuật học cây mạnh mẽ trong Học máy. Nó hoạt động bằng cách tạo ra một số Cây Quyết Định (Decision Trees) trong giai đoạn huấn luyện. Mỗi cây được xây dựng bằng cách sử dụng một tập hợp con ngẫu nhiên của tập dữ liệu để đo lường một tập hợp con ngẫu nhiên của các đặc trưng trong mỗi phân vùng. Sự ngẫu nhiên này tạo ra sự biến đổi giữa các cây riêng lẻ, giảm nguy cơ quá khớp (overfitting) và cải thiện hiệu suất dự đoán tổng thể.

Trong quá trình dự đoán, thuật toán tổng hợp kết quả của tất cả các cây, bằng cách bỏ phiếu (đối với các tác vụ phân loại) hoặc tính trung bình (đối với các tác vụ hồi quy). Quá trình ra quyết định hợp tác này, được hỗ trợ bởi nhiều cây với các thông tin chi tiết của chúng, mang lại kết quả ổn định và chính xác. Random forests được sử dụng rộng rãi cho các chức năng phân loại và hồi quy, nổi tiếng với khả năng xử lý dữ liệu phức tạp, giảm hiện tượng quá khớp (overfitting) và cung cấp các dự báo đáng tin cậy trong nhiều môi trường khác nhau.

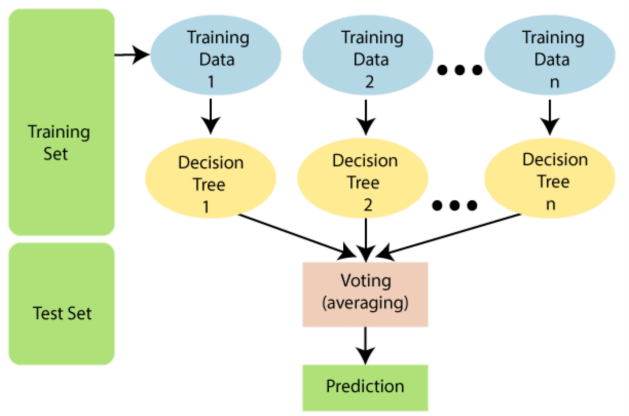
Các bước thực hiện của thuật toán:

B1 - Đầu tiên, hãy bắt đầu với việc chọn các mẫu ngẫu nhiên từ một tập dữ liệu nhất định.

B2- Tiếp theo, thuật toán này sẽ xây dựng một cây quyết định cho mọi mẫu. Sau đó, nó sẽ nhận được kết quả dự đoán từ mọi cây quyết định.

B3 - Ở bước này, sẽ thực hiện bình chọn cho mọi kết quả dự đoán.

B4 - Cuối cùng, chọn kết quả dự đoán được bình chọn nhiều nhất làm kết quả dự đoán cuối cùng.



Sơ đồ hoạt động của Random Forest

4. THỰC NGHIỆM.

4.1. Dữ liệu.

Trong nghiên cứu này, dữ liệu được thu thập từ một bộ dữ liệu bao gồm thông tin của 2111 cá

nhân từ ba quốc gia: Mexico, Peru, và Colombia. Bộ dữ liệu bao gồm 16 thuộc tính, thể hiện các đặc điểm nhân khẩu học, thói quen ăn uống, và hoạt động thể chất. Những thuộc tính này đóng vai trò quan trọng trong việc phân loại mức độ béo phì của mỗi cá nhân.

Quá trình xử lý dữ liệu được thực hiện qua nhiều bước. Đầu tiên, dữ liệu được tiền xử lý để đảm bảo chất lượng và tính nhất quán. Các giá trị bị thiếu hoặc không hợp lệ được kiểm tra và loại bỏ. Sau đó, các thuộc tính được chuẩn hóa để đảm bảo các giá trị có cùng thang đo, tránh ảnh hưởng tiêu cực đến hiệu suất của các mô hình phân loại. Các thuộc tính dạng phân loại được mã hóa thành giá trị số để phù hợp với các thuật toán học máy.

Bộ dữ liệu sau khi xử lý được sử dụng để huấn luyện các mô hình phân loại Logistic Regression, Random Forest, và SVM. Dữ liệu được chia thành hai tập: tập huấn luyện chiếm 80% và tập kiểm tra chiếm 20%. Phương pháp phân tầng được áp dụng để đảm bảo tỷ lệ các mức độ béo phì trong cả hai tập dữ liệu là đồng nhất.

Các đặc điểm nhân khẩu học, thói quen ăn uống, và hoạt động thể chất gồm:

Gender: Giới tính của cá nhân.

Age: Tuổi của cá nhân.

Height: Chiều cao.

Weight: Cân nặng.

Family history with overweight: Gia đình có tiền sử thừa cân hay không.

FAVC (Frequent Consumption of High Caloric Food): Thói quen tiêu thụ thực phẩm nhiều năng lượng.

FCVC (Frequency of Consumption of Vegetables): Tần suất ăn rau củ.

NCP (Number of Main Meals): Số lượng bữa ăn chính trong ngày.

CAEC (Consumption of Food Between Meals): Tần suất ăn vặt giữa các bữa ăn.

SMOKE: Thói quen hút thuốc.

CH2O (Daily Water Consumption): Lượng nước tiêu thụ hàng ngày.

SCC (Calories Monitoring): Theo dõi lượng calo.

FAF (Frequency of Physical Activity): Tần suất hoạt động thể chất hàng tuần.

TUE (Time Using Technology Devices): Thời gian sử dụng thiết bị công nghệ mỗi ngày.

CALC (Consumption of Alcohol): Tần suất tiêu thụ rượu bia.

MTRANS (Transportation Method): Phương tiện di chuyển chính.

4.2. Phương pháp.

Các bước thực hiện:

Import các thư viện cần thiết -> Lấy dataset từ UCI -> chia dữ liệu thành tập train và test -> Tiền xử lý dữ liệu (chuẩn hóa dữ liệu) -> Khởi tạo các mô hình -> Đánh giá hiệu suất của mô hình bằng các độ đo -> Hiển thị kết quả -> Kết thúc.

Chúng tôi sử dụng các độ đo đó để đánh giá hiệu quả là: Accuracy, Precison, Recall, F1.

4.3. Kết quả.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6974	0.6869	0.6974	0.6820
Random Forest	0.9527	0.9563	0.9527	0.9534
SVM	0.7447	0.7376	0.7447	0.7320

Kết quả đo lường hiệu suất phân loại

Logistic Regression không phù hợp với bộ dữ liệu này, có thể do bộ dữ liệu không tuyến tính, hoặc tồn tại mối quan hệ phức tạp giữa các đặc trưng mà Logistic Regression không thể nắm bắt.

Random Forest hoạt động rất tốt với dữ liệu này, nhờ khả năng xử lý dữ liệu phi tuyến tính, giải quyết vấn đề đa cộng tuyến và giảm overfitting.

SVM hoạt động khá tốt, nhưng chưa bằng Random Forest. Có thể cần điều chỉnh các siêu tham

số như kernel hoặc C để cải thiện hiệu suất.

5. KẾT LUẬN

Nghiên cứu này đã tập trung vào việc phân loại mức độ béo phì dựa trên thói quen ăn uống và tình trạng thể chất bằng cách sử dụng ba mô hình học máy phổ biến: Logistic Regression, Random Forest và SVM. Dữ liệu được sử dụng từ UCI Dataset (id=544), bao gồm thói quen ăn uống và hoạt động thể chất của các cá nhân. Quá trình nghiên cứu đã trải qua các bước tiền xử lý dữ liệu, bao gồm chuẩn hóa giá trị bằng MaxAbsScaler, xử lý dữ liệu bị thiếu bằng Mean Imputer, và phân tích tương quan để hiểu mối liên hệ giữa các đặc trưng.

Random Forest là mô hình phù hợp nhất trong nghiên cứu này với hiệu suất vượt trội trong việc phân loại mức độ béo phì. Điều này khẳng định vai trò quan trọng của việc lựa chọn mô hình học máy phù hợp với đặc trưng của dữ liệu. Các kết quả thu được không chỉ cung cấp thông tin quan trọng về cách phân loại mức độ béo phì mà còn mở ra tiềm năng cho các ứng dụng trong lĩnh vực y tế, giáo dục sức khỏe, và nâng cao chất lượng cuộc sống.

TÀI LIỆU THAM KHẢO

[1] Marie Ng, Fleming T, Robinson M, et al. "Global, regional, and national prevalence of overweight and obesity in children and adults, 1980–2013." *The Lancet* (2014).

[2] WHO. "Obesity and overweight." (2021).

[3] Rodriguez E, et al. "Machine Learning Techniques to Predict Overweight or Obesity." (2022).

[4] Cisneros M, et al. "Estimation of Obesity Levels Based on Eating Habits and Physical Condition." *UCI Repository* (2021).

[5] Martinez-Vizcaino V, et al. "Childhood Obesity Prevention Programs." *ScienceDirect* (2022).

[6] Mousavi M, et al. "Predictive Modeling for Obesity." (2020).

[7] Elias Rodríguez, Elen Rodríguez, Luiz Nascimento, Aneirson da Silva, Fernando Marins,

"Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits", Diagnostics, vol. 13, no. 18, 2023, pp. 2949.

[8] Elias Rodríguez, Elen Rodríguez, Luiz Nascimento, Aneirson da Silva, Fernando Marins, "Machine Learning Techniques to Predict Overweight or Obesity", CEUR Workshop Proceedings, 2021.

[9] Elias Rodríguez, Elen Rodríguez, Luiz Nascimento, Aneirson da Silva, Fernando Marins, "Estimation of Obesity Levels through Machine Learning Techniques", CEUR Workshop Proceedings, 2021.

PHÂN CHIA CÔNG VIỆC

	Võ Quốc Phong	Phạm Đức Đại
Công việc làm riêng	- Tóm tắt - Công trình liên quan - Thuật toán: Logistic Regression, Random Forest	- Bài toán - Giới thiệu - Thuật toán: SVM - Kết luận
Công việc làm chung	- Phần 4 thực nghiệm	

Link Colab

[Phan_loai_muc_do_beo_phi.ipynb - Colab]