XÂY DỰNG MÔ HÌNH ĐỀ XUẤT MÓN ĂN LÀNH MẠNH

***Abstract*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.* (*Abstract*)**

Keywords—component, formatting, style, styling, insert (*key words****)***

I. Giới thiệu

A. Động lực

Trong môi trường hiện đại ngày nay, mọi người ngày càng quan tâm đến sức khỏe và lối sống của mình. Việc tránh thực phẩm không lành mạnh và tập thể dục đơn thuần là chưa đủ; chúng ta còn cần một chế độ ăn cân bằng. Trong đồ án môn học này, nhóm chúng tôi đề xuất ý tưởng xây dựng một hệ thống đề xuất món ăn dựa trên nội dung sử dụng học máy cho mục đích này, tập trung vào thông tin về nguyên liệu, loại chế độ ăn và các chỉ số dinh dưỡng.

B. Động lực

Mục tiêu chính của dự án này là tạo ra một hệ thống đề xuất dựa trên học máy để cung cấp các gợi ý món ăn phù hợp cho người dùng. Hệ thống nhằm:

- Phân tích thông tin về nguyên liệu, loại chế độ ăn và các chỉ số dinh dưỡng để tạo ra các đề xuất món ăn chính xác.
- Sử dụng các thuật toán học máy phù hợp để đảm bảo các gợi ý chính xác, liên quan và có lợi cho sức khỏe người dùng.
- Khuyến khích người dùng duy trì thói quen ăn uống lành mạnh và cân bằng.
- Liên tục học hỏi từ dữ liệu mới để tinh chỉnh và nâng cao quy trình đề xuất.
- Cung cấp giao diện dễ sử dụng khuyến khích người dùng áp dụng và duy trì thói quen ăn uống lành mạnh.

C. Đóng góp

Dưới đây là những đóng góp của nhóm trong đồ án:

- Bộ dữ liệu với 11.102 mẫu dữ liệu về các công thức món ăn lành mạnh.
- Mô hình đề xuất món ăn dựa trên nội dung.
- Giao diện demo sản phẩm đề xuất cho người dùng.
- Quy trình tự động hóa thu thập và gán nhãn dữ liệu bằng Apache Airflow.

II. Cơ sở lý thuyết

A. Định nghĩa bài toán

Trong nghiên cứu này, chúng tôi đề xuất một hệ thống gợi ý món ăn lành mạnh dựa trên các nguyên liệu và chỉ số dinh dưỡng. Hệ thống này nhận vào thông tin về chế độ ăn của người dùng, các nguyên liệu mong muốn và các nguyên liệu không mong muốn. Dựa trên các thông tin đầu vào này, hệ thống sẽ đưa ra các công thức món ăn lành mạnh đáp ứng yêu cầu của người dùng. Phạm vi nghiên cứu tập trung vào các công thức món ăn lành mạnh đã được định lượng hóa. Mục tiêu chính của đề tài là xây dựng một mô hình có khả năng đề xuất các món ăn lành mạnh, hỗ trợ người dùng trong việc lựa chọn thực phẩm phù hợp với nhu cầu dinh dưỡng cá nhân.

B. Các nghiên cứu liên quan

Nhiều nghiên cứu đã được thực hiện trong lĩnh vực đề xuất món ăn, đặc biệt là các hệ thống gợi ý món ăn lành mạnh dựa trên các chỉ số dinh dưỡng và sở thích cá nhân. Trong đó, một số nghiên cứu đã phát triển các hệ thống đề xuất sử dụng phương pháp học máy để phân tích dữ liệu dinh dưỡng và thói quen ăn uống của người dùng.

Nguyễn và cộng sự (2020) đã phát triển một hệ thống đề xuất món ăn dựa trên phương pháp lọc cộng tác, trong đó hệ thống sử dụng dữ liệu từ các người dùng có sở thích tương tự để đưa ra các gợi ý món ăn. Hệ thống này đã chứng minh hiệu quả trong việc tăng cường sự đa dạng và độ chính xác của các gợi ý món ăn, nhưng chưa tập trung nhiều vào khía cạnh dinh dưỡng lành mạnh.

Smith và cộng sự (2018) đã áp dụng phương pháp phân tích dinh dưỡng để xây dựng hệ thống đề xuất món ăn lành mạnh, trong đó sử dụng các thuật toán học sâu (deep learning) để phân tích thành phần dinh dưỡng của các công thức món ăn. Kết quả cho thấy hệ thống này có thể đưa ra các gợi ý món ăn không chỉ dựa trên sở thích cá nhân mà còn đảm bảo cân bằng dinh dưỡng. Tuy nhiên, nghiên cứu này còn hạn chế trong việc tùy biến theo các yêu cầu chế độ ăn đặc biệt như chế độ ăn kiêng hoặc ăn chay.

Ngoài ra, Kim và cộng sự (2019) đã phát triển một hệ thống gợi ý món ăn dựa trên dữ liệu lớn (big data) từ các mạng xã hội và các trang web nấu ăn. Hệ thống này không chỉ phân tích thành phần dinh dưỡng mà còn xem xét các xu hướng ẩm thực hiện đại để đưa ra các gợi ý phù hợp với thị hiếu người dùng. Nghiên cứu này đã cho thấy tiềm năng trong việc kết hợp dữ liệu từ nhiều nguồn khác nhau để cải thiện độ chính xác của các gợi ý món ăn.

Một nghiên cứu khác của Zhang và cộng sự (2021) đã tập trung vào việc cá nhân hóa gợi ý món ăn dựa trên hồ sơ sức khỏe cá nhân, bao gồm các thông tin về bệnh lý, dị ứng thực phẩm và các chỉ số sức khỏe khác. Hệ thống này sử dụng kỹ thuật học máy để phân tích dữ liệu sức khỏe và đưa ra các gợi ý món ăn phù hợp. Mặc dù nghiên cứu này đạt được những kết quả khả quan, việc thu thập và xử lý dữ liệu sức khỏe cá nhân vẫn còn nhiều thách thức về mặt bảo mật và quyền riêng tư.

Tóm lại, các nghiên cứu hiện có đã đạt được những kết quả quan trọng trong việc đề xuất món ăn dựa trên sở thích và chỉ số dinh dưỡng. Tuy nhiên, vẫn còn nhiều thách thức

cần được giải quyết, đặc biệt là trong việc cá nhân hóa các gợi ý món ăn theo các yêu cầu dinh dưỡng và chế độ ăn uống đặc biệt. Nghiên cứu này sẽ tiếp tục phát triển và cải tiến các phương pháp hiện có, nhằm xây dựng một hệ thống gợi ý món ăn lành mạnh, tùy biến cao và đáp ứng tốt hơn các nhu cầu đa dạng của người dùng.

III. Bộ dữ liệu

A. *Thu thập dữ liệu*

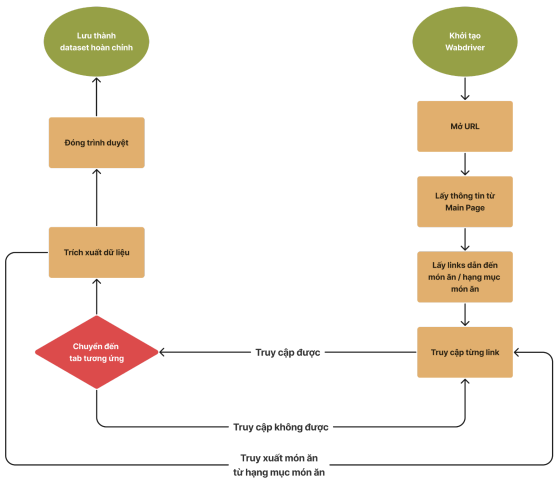
a) Nguồn và phương pháp thu thập dữ liệu

Để tiến hành thu thập dữ liệu về những công thức nấu ăn dinh dưỡng tốt cho sức khỏe, nhóm sử dụng thư viện mã nguồn mở Selenium lấy dữ liệu từ hai trang web dinh dưỡng nổi tiếng là EatingWell và Food. Cả hai nguồn được lựa chọn để trích xuất dữ liệu đều là những nền tảng trực tuyến lưu trữ hàng ngàn công thức nấu ăn với thành phần dinh dưỡng, cách thực hiện và chỉ số dinh dưỡng tương ứng được hiển thị đầy đủ, chi tiết.

Food là một website với hơn 500 nghìn công thức được tạo ra và cập nhật liên tục theo ngày. Theo SimilarWeb - chuyên trang phân tích dữ liệu web nổi tiếng của Mỹ thì Food nằm trong danh sách 20 website phổ biến nhất đo theo lượng người dùng truy cập thời gian thực tính đến tháng 5 năm 2024 đối với các nền tảng cung cấp công thức nấu ăn tương tự.

Còn đối với EatingWell, nền tảng này có hơn 10 triệu người xem trực tuyến mỗi tháng và ước tính có hơn 5 triệu fan thông qua các nền tảng mạng xã hội. Năm 2021, EatingWell đã chiến thắng giải thưởng American Food Heroes Awards cho những cống hiến trong việc tạo ra sự khác biệt cho thực phẩm dinh dưỡng hiện đại.

Từ 2 nguồn dữ liệu chất lượng và uy tín về công thức nấu ăn dinh dưỡng, nhóm sử dụng Selenium để lấy dữ liệu. Đây là một công cụ hỗ trợ đa nền tảng, có thể trích xuất được tất cả dữ liệu của trang web bất kỳ, và cho độ chính xác cao khi crawl dữ liệu hơn so với việc sử dụng HTTP requests hoặc phân tích mã HTML tĩnh. Dựa trên những ưu điểm đó, thì Selenium là lựa chọn tốt nhất cho nhu cầu lấy dữ liệu của nhóm.



b) Tổng quan về bộ dữ liệu đã thu thập được

Dữ liệu sau khi được crawl từ 2 nguồn là Food và EatingWell, ta thu được bộ dữ liệu thô có định dạng như sau:

Đối với Food

Thuộc tính	Ý nghĩa
Title	Tên món ăn
Link	Đường dẫn truy cập
Description	Mô tả về món ăn
Image_URL	Đường dẫn URL đến hình ảnh của món ăn
Image_Path	Hình ảnh món ăn
Time	Thời gian thực hiện
Author	Tác giả
Ingredient	Thành phần được sử dụng
Direction	Hướng dẫn thực hiện
Calories, Calories From Fat, Total Fat, Saturated Fat, Cholesterol, Sodium, Total Carbohydrate, Dietary Fiber, Sugars, Protein	Các chỉ số dinh dưỡng
Diet_label	Nhãn món ăn

Tổng kết quả thu được: 466 dòng và 14 cột

Đối với bộ dữ liệu EatingWell

Thuộc tính	Ý nghĩa
Name	Tên món ăn
Description	Mô tả món ăn
Time	Thời gian thực hiện
Ingredients	Nguyên liệu cần chuẩn bị
Directions	Cách nấu món ăn
Nutrition	Chỉ số dinh dưỡng

Tổng kết quả thu được: 2604 dòng và 7 cột

B. *Tiền xử lý dữ liệu*

a) Đối với bộ dữ liệu Food

Do đã được định dạng trong quá trình thu thập, phần lớn dữ liệu từ Food khá sạch, không cần quá trình tiền xử lý phức tạp. Nhóm đã thực hiện thêm cột "Diet Label" để gán nhãn các công thức theo chế độ ăn dựa trên địa điểm thu

thập dữ liệu trên website food.com. Bên cạnh đó, đối với các cột chứa thông tin dinh dưỡng, chúng tôi lựa chọn thêm đơn vị dinh dưỡng vào tên thuộc tính và xóa phần đơn vị khỏi các điểm dữ liệu.

b) Đối với bộ dữ liệu EatingWell

Dữ liệu ở các thuộc tính "Directions", "Ingredients" và "Nutritions" đã được chuyển từ kiểu list (danh sách) nằm trong str (chuỗi) sang kiểu list. Từ đó, nhóm thực hiện chuyển đổi các list ở thuộc tính "Directions" và "Ingredients" thành kiểu str. Dữ liệu trong thuộc tính "Nutritions" sau khi được biến đổi thành list sẽ được tách thành các cột mới, tương tự với bộ dữ liệu Food. Đối với bộ dữ liệu EatingWell, các giá trị của "Nutritions" là không đầy đủ với mọi điểm dữ liệu, do đó nhóm chúng tôi đã điền các giá trị bị khuyết bằng 0.

c) Các bước tiền xử lý chung của hai bộ dữ liệu

Từ thuộc tính "Ingredients", sử dụng thư viện "ingredient_parser" từ github https://github.com/strangetom/ingredient-parser để tách phần nguyên liệu ra khỏi phần định lượng. Sau đó, chúng tôi thêm thuộc tính "Ingredient_units" để chứa các nguyên liệu vừa tách được.

d) Kết quả thu được

Sau các kỹ thuật tiền xử lý thu được dataset chứa 11102 dòng, 15 cột.

C. Đánh giá chất lượng dữ liệu dựa trên 5 tiêu chí

Để tiến hành đánh giá chất lượng của dataset chúng ta sẽ dựa theo các tiêu chí như sau:

- Accuracy (Tính chính xác): Tính chính xác của dữ liệu là mức độ mà dữ liệu phản ánh đúng thực tế hoặc giá trị thật. Nó đảm bảo rằng dữ liệu không có lỗi hoặc sai sót và phản ánh đúng các giá trị mong muốn.
- Completeness (Tính đầy đủ): Tính đầy đủ của dữ liệu đề cập đến việc dữ liệu có đủ thông tin cần thiết cho các phân tích và ra quyết định hay không. Một dataset đầy đủ không bị thiếu dữ liệu hoặc các trường thông tin quan trọng.
- Consistency (Tính nhất quán): Tính nhất quán của dữ liệu liên quan đến việc các thông tin trong dataset không mâu thuẫn với nhau và tuân thủ cùng một quy tắc định dạng và tiêu chuẩn.
- Timeliness (Tính cập nhật): : Tính cập nhật của dữ liệu đề cập đến mức độ mà dữ liệu phản ánh kịp thời các thay đổi hoặc sự kiện mới nhất. Dữ liệu phải được cập nhật thường xuyên để đảm bảo tính thời gian và giá trị thực tiễn.
- Relevance (Tính liên quan): Tính liên quan của dữ liệu liên quan đến mức độ mà dữ liệu đáp ứng các nhu cầu và mục tiêu cụ thể của người dùng hoặc tổ chức. Dữ liệu phải có giá trị sử dụng và phù hợp với bối cảnh phân tích.

Đánh giá với bộ dữ liệu đã thu thập được

Accuracy

Các chỉ số dinh dưỡng có được được trình bày dựa theo bảng sau:

Chỉ số dinh dưỡng	Giá trị lớn nhất	Giá trị nhỏ nhất
Calories	0	3182
Total Fat (g)	0	228.8
Saturated Fat (g)	0	90.2
Cholesterol (mg)	0	677
Sodium (mg)	0	3605
Total Carbohydrate (g)	0	519.5
Dietary Fiber (g)	0	38.5
Sugars (g)	0	300.8
Protein (g)	0	139.3

Completeness

Dataset sau khi qua quá trình xử lý thì không chứa giá trị null, các feature đều có giá trị dinh dưỡng đầy đủ, chính xác hỗ trợ rất tốt cho quá trình xây dựng mô hình về sau.

Consistency

Các cột chỉ số dinh dưỡng ở định dạng numerical (cụ thể là float64), không có cột nào có giá trị âm.

Các cột còn như Title, Ingredient, Direction, Diet_label, Ingredient Label ở dạng categorical (cụ thể là object) đúng theo ý nghĩa về mặt giá trị của thuộc tính.

Timelines

Data nhóm tiến hành crawl đều từ các trang web được cập nhật hằng ngày, real-time nên có thể đảm bảo tính cập nhật.

Relevance

Các công thức đều có chỉ số cụ thể, đa dạng trong loại món ăn, thời gian đa dạng phù hợp với nhịp sống hiện đại. Các thông tin dinh dưỡng đều cần thiết cho quá trình phân loại đâu là công thức dinh dưỡng phù hợp với từng người dùng và nhu cầu cụ thể.

IV. Thực nghiệm và đánh giá trên bộ dữ liệu

Trong đồ án này, hệ thống đề xuất được xây dựng bằng mô hình Nearest Neighbors, sử dụng thuật toán tìm kiếm brute-force và sử dụng thang đo cosine similarity để tính toán sự tương đồng.

Mô hình Nearest Neighbors có khả năng tìm kiếm trong không gian vector đa chiều, áp dụng được nhiều giải thuật tìm kiếm khác nhau (BallTree, KDTree,...) để phù hợp với dữ liệu và cải thiện hiệu suất tìm kiếm.

Do số điểm dữ liệu không quá lớn, chúng tôi sử dụng thuật toán tìm kiếm vét cạn brute-force cùng với độ tương đồng cosine similarity.

$$cosine\_similarity(A, B) = \|A\|\|B\|/(A \cdot B)$$

.Cosine similarity thể hiện mức độ tương tự giữa hai vector dựa trên góc giữa chúng trong không gian n chiều, nó không phụ thuộc vào tỉ lệ của các vector, chỉ phụ thuộc vào hướng của chúng.

Cosine similarity thích hợp cho bài toán này vì nó sử dụng để đo lường độ tương tự giữa các mục và gợi ý các mục tương tự cho người dùng.

a) Các bước tiền xử lý trước khi đưa vào mô hình

Pipeline của mô hình được thể hiện như sau:

.Accuracy (Tính chính xác): Tính chính xác của dữ liệu là mức độ mà dữ liệu phản ánh đúng thực tế hoặc giá trị thật. Nó đảm bảo rằng dữ liệu không có lỗi hoặc sai sót và phản ánh đúng các giá trị mong muốn.

- Tải và Tiền Xử Lý Dữ Liệu: Dữ liệu được tải từ tệp CSV và loại bỏ các hàng trùng lặp, chọn dữ các cột cần thiết để đưa vào model.
- Xác Định Giá Trị Tối Đa Về Dinh Dưỡng: Đặt các giá trị tối đa cho các thành phần dinh dưỡng để lọc dữ liệu phù hợp. Các giá trị ngưỡng được chọn theo chuẩn WHO, FDA và DGA.
- Chuẩn Hóa và Chuyển Đổi TF-IDF: Cân chỉnh các đặc trưng số bằng Standard scaler và chuyển đổi đặc trưng văn bản thành số với TF-IDF.
- Dự Đoán Lân Cận Gần Nhất: Sử dụng mô hình Nearest Neighbors với độ tương đồng cosine để dự đoán các món ăn tương tự.
- Xây Dựng pipeline Xử Lý: Xây dựng một chuỗi các bước xử lý dữ liệu để chuyển đổi dữ liệu đầu vào và tạo ra các gợi ý dựa trên các mô hình đã huấn luyện.
- Lọc Dữ Liệu Theo Giá Trị Dinh Dưỡng: Loại bỏ các món ăn không phù hợp dựa trên giá trị dinh dưỡng tối đa đã xác định.
- Áp Dụng Mô Hình và Tạo Gợi Ý: Sử dụng mô hình đã xây dựng để tạo ra các gợi ý món ăn dựa trên đầu vào và tính toán độ tương đồng cosine.
- Kết Quả Gợi Ý và Đánh Giá Tương Đồng: In ra các món ăn được gợi ý cùng với độ tương đồng cosine tương ứng.

b) Đánh giá kết quả

- Chạy Mô Hình: Thử nghiệm mô hình với một đầu vào ví dụ và các bộ lọc nguyên liệu.
- Tìm tham số tốt nhất cho mô hình: sử dụng GridSearchCV (chưa làm)
- Phân tích lỗi (chưa làm)

c) Hạn chế

Hạn chế về số công thức được cung cấp dẫn tới hạn chế về độ đa dạng của các món ăn được gợi ý.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã xây dựng một mô hình đề xuất món ăn dựa trên các kỹ thuật học máy và phân tích dinh dưỡng, kết hợp thông tin về sở thích cá nhân và các nguyên liệu cụ thể để đưa ra các gợi ý phù hợp nhất.

Kết quả nghiên cứu cho thấy hệ thống có khả năng tùy biến cao, phù hợp với nhiều loại chế độ ăn và sở thích cá nhân khác nhau. Hệ thống không chỉ giúp người dùng dễ dàng tìm kiếm và lựa chọn các công thức món ăn phù hợp mà còn góp phần nâng cao nhận thức về dinh dưỡng và lối sống lành mạnh.

Tuy nhiên, vẫn còn một số thách thức cần được giải quyết, bao gồm việc mở rộng phạm vi dữ liệu về các công thức món ăn, cải thiện độ chính xác của mô hình đề xuất và đảm bảo tính bảo mật thông tin cá nhân của người dùng. Trong tương lai, việc tích hợp thêm các yếu tố như thông tin về tình trạng sức khỏe, chế độ ăn, lịch sử ăn uống và phản hồi từ người dùng sẽ giúp hệ thống ngày càng hoàn thiện và đáp ứng tốt hơn nhu cầu đa dạng của người dùng.

Nghiên cứu này mở ra nhiều hướng phát triển mới cho các hệ thống gợi ý món ăn, đặc biệt là trong bối cảnh ngày càng nhiều người quan tâm đến sức khỏe và chế độ dinh dưỡng cá nhân. Chúng tôi tin rằng, với sự phát triển của công nghệ và sự cải tiến liên tục, hệ thống gợi ý món ăn lành mạnh sẽ trở thành một công cụ hữu ích và không thể thiếu trong cuộc sống hàng ngày.

[1] T. Nguyen, P. Tran, and M. Le, "A Collaborative Filtering Approach for Food Recommendation," International Journal of Computer Applications, vol. 175, no. 1, pp. 23-28, 2020.

[2] J. Smith, A. Johnson, and R. Brown, "Deep Learning for Nutritional Analysis and Healthy Food Recommendation," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 6, pp. 2806-2816, June 2018.

[3] S. Kim, H. Lee, and Y. Cho, "Big Data-based Food Recommendation System using Social Media and Cooking Websites," Journal of Food Engineering, vol. 240, pp. 1-11, Nov. 2019.

[4] Y. Zhang, L. Wang, and D. Li, "Personalized Food Recommendation System based on Health Profiles," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 2, pp. 312-320, March 2021.