BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

BÁO CÁO TỔNG KẾT

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN

NGHIÊN CỨU VÀ XÂY DỰNG WEBSITE THEO DÕI THỰC PHẨM ĂN HẰNG NGÀY

Mã số đề tài: 071

TPHCM, 03/2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

BÁO CÁO TỔNG KẾT

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN

NGHIÊN CỨU VÀ XÂY DỰNG WEBSITE THEO DÕI THỰC PHẨM ĂN HẰNG NGÀY

Mã số đề tài: 071

Chủ nhiệm đề tài: Nguyễn Lư Hồng Phương

Khoa: Công nghệ thông tin

Các thành viên: Nguyễn Thị Mai

Người hướng dẫn: TS. Trương Hoàng Vinh

TPHCM, 03/2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ
THÀNH PHỐ HỒ CHÍ MINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

THÔNG TIN KẾT QUẢ NGHIÊN CỨU CỦA ĐỀ TÀI

1. **Thông tin chung:**
   - Tên đề tài: Nghiên cứu và xây dựng website theo dõi thực phẩm ăn hằng ngày.
   - Mã số đề tài: 071
   - Sinh viên chủ nhiệm đề tài: Nguyễn Lư Hồng Phương
   - Khoa: Công nghệ thông tin
   - Giảng viên hướng dẫn: TS. Trương Hoàng Vinh
2. **Mục tiêu đề tài:**

   Nghiên cứu nhằm xây dựng một hệ thống cung cấp thông tin chi tiết và đáng tin cậy về hàm lượng calo trong thực phẩm, dựa trên các nguồn dữ liệu chính thức và khoa học. Website không chỉ giúp người dùng tra cứu nhanh chóng lượng calo của từng loại thực phẩm mà còn hỗ trợ phân tích chế độ dinh dưỡng cá nhân, đưa ra các gợi ý cải thiện thực đơn phù hợp với mục tiêu sức khỏe như giảm cân, duy trì cân nặng hoặc tăng cường thể lực. Bên cạnh đó, hệ thống tính toán đưa ra đề xuất những điều chỉnh hợp lý nhằm tối ưu hóa chế độ dinh dưỡng. Đồng thời, người dùng cũng có thể nhận được các khuyến nghị về thời gian ăn uống hợp lý, cân bằng giữa các nhóm chất (đạm, béo, carbohydrate, vitamin, khoáng chất) dựa trên nhu cầu cá nhân.
3. **Kết quả nghiên cứu:**

   Hệ thống theo dõi thực phẩm ăn hằng ngày đã được xây dựng và triển khai với ba tính năng chính gồm:
   - Gợi ý bữa ăn cá nhân hóa: Hệ thống sử dụng thuật toán K-Means để phân nhóm theo mức độ dinh dưỡng của thực phẩm, để đưa ra đề xuất ba bữa ăn chính hợp lý cho người dùng, dựa trên các chỉ số và mục tiêu dinh dưỡng cá nhân.
   - Thiết kế bữa ăn theo nhu cầu: Hệ thống cho phép người dùng tự tạo thực đơn dựa trên sở thích và nhu cầu dinh dưỡng, vận dụng thuật toán Decision Tree để phân loại dinh dưỡng giúp người dùng lựa chọn bữa ăn cân đối và phù hợp với mục tiêu sức khỏe.
   - Tra cứu thông tin thực phẩm: Cung cấp giá trị dinh dưỡng chi tiết về 533 loại thực phẩm Việt Nam, giúp người dùng dễ dàng kiểm soát lượng calo và thành phần dinh dưỡng.
4. **Đóng góp về mặt kinh tế - xã hội, giáo dục và đào tạo, an ninh, quốc phòng và khả năng áp dụng của đề tài:**

   Hệ thống nhằm cung cấp thông tin dinh dưỡng chính xác, đáng tin cậy cho người dùng, hỗ trợ họ trong việc duy trì lối sống lành mạnh và cân đối. Bằng cách quản lý lượng calo và thành phần dinh dưỡng một cách hiệu quả, hệ thống giúp người dùng dễ dàng theo dõi, điều chỉnh chế độ ăn uống phù hợp với nhu cầu cá nhân.

   Một điểm đặc biệt của hệ thống là sử dụng bộ dữ liệu thực phẩm dành riêng cho người Việt Nam, được xây dựng dựa trên các nguồn đáng tin cậy, bao gồm Viện Nghiên cứu Dinh dưỡng TP. Hồ Chí Minh. Bộ dữ liệu này tổng hợp đa dạng các loại thực phẩm phổ biến trong bữa ăn hàng ngày của người Việt, từ các món ăn truyền thống đến thực phẩm công nghiệp hiện đại. Điều này giúp người dùng tra cứu linh hoạt giá trị dinh dưỡng của từng món ăn, từ đó đưa ra các lựa chọn thực phẩm phù hợp với mục tiêu sức khỏe của mình.

   Ngoài ra, hệ thống còn hướng đến cân bằng dinh dưỡng tổng thể, đảm bảo cung cấp đầy đủ các nhóm chất cần thiết như protein, chất béo, carbohydrate, vitamin và khoáng chất. Nhờ tích hợp trí tuệ nhân tạo (AI) và Machine Learning, hệ thống có khả năng phân tích thói quen ăn uống của người dùng, đưa ra các đề xuất tối ưu hóa chế độ ăn hằng ngày, phù hợp với mục tiêu cá nhân như giảm cân, tăng cơ hay duy trì sức khỏe.

Với các tính năng trên, hệ thống không chỉ hỗ trợ cá nhân hóa chế độ dinh dưỡng, mà còn góp phần nâng cao nhận thức về dinh dưỡng khoa học, giúp người Việt hình thành thói quen ăn uống lành mạnh và cải thiện chất lượng cuộc sống một cách bền vững.

MỤC LỤC

Danh sách hình vẽ

Danh sách các bảng

Danh sách từ viết tắt

AI	Artificial Intelligence.
BIA	Machine Learning.
BMI	Body Mass Index.
BMR	Machine Learning.
KNN	K-Nearest Neighbor.
ML	Machine Learning.
TDEE	Total Daily Energy Expenditure.

Danh sách từ tiếng anh

Bernouli Naıve Bayes	Dữ liệu nhị phân.
Classification	Phân loại.
Clustering	Phân cụm.
Decision Tree Algorithm	Thuật toán Cây quyết định.
Gaussian Naıve Bayes	Dữ liệu liên tục.
K-means clustering	Phân cụm K-means.
K-nearest Neighbours	KNN - K lân cận gần nhất.
Machine Learning	Máy học.
Multinomial Naıve Bayes	Dữ liệu rời rạc.
Regression	Hồi quy.
Reinforcement	Học củng cố.
Supervised Learning	Học có giám sát.
Total Daily Energy Expenditure	Tổng năng lượng tiêu hao hàng ngày.
Unsupervised	Học không giám sát.

Chương 1

Tổng quan về đề tài

1.1 Lý do chọn đề tài

Hiện nay, khi chất lượng cuộc sống ngày càng cải thiện, đồng thời con người phải đối mặt với lịch trình bận rộn, việc đảm bảo một chế độ ăn uống cân bằng và khoa học là điều cần thiết. Tuy nhiên, các phương pháp tra cứu thủ công về thành phần dinh dưỡng trong thực phẩm như vitamin A, vitamin B, protein, chất xơ... trong thực phẩm thường phức tạp, mất nhiều thời gian và không thuận tiện cho người dùng phổ thông. Điều này dẫn đến tình trạng nhiều người không nắm rõ được lượng dinh dưỡng họ tiêu thụ mỗi ngày, gây ảnh hưởng đến sức khỏe tổng thể. Do đó, nghiên cứu và xây dựng website theo dõi thực phẩm ăn hằng ngày, áp dụng Machine Learning gợi ý bữa ăn hợp lý cho người dùng, kết hợp bộ tra cứu, và thiết kế bữa ăn theo nhu cầu người dùng là một giải pháp cần thiết. Hệ thống này không chỉ giúp tiết kiệm thời gian mà còn cung cấp thông tin dinh dưỡng chính xác từ các Chuyên gia Dinh dưỡng thuộc Viện Nghiên cứu Dinh dưỡng TP. Hồ Chí Minh, hỗ trợ đưa ra gợi ý nhằm cải thiện chế độ dinh dưỡng hiện tại của người dùng dựa trên tính toán chỉ số khối cơ thể (BMI) và Tổng năng lượng chuyển hóa (TDEE) của người dùng .

1.2 Câu hỏi nghiên cứu

1.3 Mục tiêu nghiên cứu

Nghiên cứu nhằm xây dựng một hệ thống cung cấp thông tin chi tiết và đáng tin cậy về hàm lượng calo trong thực phẩm, dựa trên các nguồn dữ liệu chính thức và khoa học. Website không chỉ giúp người dùng tra cứu nhanh chóng lượng calo của từng loại thực phẩm mà còn hỗ trợ phân tích chế độ dinh dưỡng cá nhân, đưa ra các gợi ý cải thiện thực đơn phù hợp với mục tiêu sức khỏe như giảm cân, duy trì cân nặng hoặc tăng cường thể lực. Bên cạnh đó, hệ thống tính toán đưa ra đề xuất những điều chỉnh hợp lý nhằm tối ưu hóa chế độ dinh dưỡng. Đồng thời, người dùng cũng có thể nhận được các khuyến nghị về thời gian ăn uống hợp lý, cân bằng giữa các nhóm chất (đạm, béo, carbohydrate, vitamin, khoáng chất) dựa trên nhu cầu cá nhân.

1.4 Đối tượng nghiên cứu

Thành phần dinh dưỡng và lượng calo của các thực phẩm phổ biến của người Việt Nam là đối tượng quan trọng được xoáy sâu trong nghiên cứu. Dữ liệu được lấy từ Viện dinh dưỡng quốc gia và đã được xử lý, chọn lọc và phân loại thực phẩm theo nhóm: Thực phẩm tự nhiên (rau, củ, trái cây, thịt, cá, sữa), thực phẩm chế biến sẵn, đồ uống, thức ăn nhanh và một số nhóm thực phẩm phổ biến khác.

Dựa trên hành vi và nhu cầu của người dùng là những người quan tâm đến sức khỏe và dinh dưỡng như người muốn giảm cân, duy trì vóc dáng, người tập luyện thể thao cần theo dõi lượng calo nạp vào để tối ưu hiệu suất hay người đang mắc các bệnh lý cần kiểm soát một số thành phần có trong thực phẩm (tiểu đường, tim mạch, dị ứng, béo phì, huyết áp cao, v.v.) để xây dựng gợi ý thích hợp.

1.5 Phạm vi nghiên cứu

Hệ thống theo dõi thực phẩm bao gồm ba thành phần chính: tra cứu bảng thành phần dinh dưỡng Việt Nam, gợi ý bữa ăn cho người dùng, và thiết kế bữa ăn dựa trên mô hình Machine Learning (ML).

Trước tiên, hệ thống sử dụng dữ liệu từ Viện Nghiên cứu Dinh dưỡng TP. Hồ Chí Minh, trong đó nhóm nghiên cứu đã chọn lọc các thành phần dinh dưỡng phổ biến và quan trọng nhất để người dùng dễ tra cứu. Tiếp theo, dựa trên các thông tin cá nhân của người dùng như chiều cao, cân nặng, giới tính, hệ thống sẽ tính toán và đưa ra gợi ý bữa ăn phù hợp, đảm bảo đáp ứng nhu cầu dinh dưỡng trong phạm vi một ngày.

Cuối cùng, hệ thống tích hợp Machine Learning để phân loại thực phẩm theo ba mức độ dinh dưỡng chính, bao gồm: lượng calo cao, lượng calo trung bình và lượng calo thấp.

1.6 Sơ lược về phương pháp nghiên cứu

Xuất phát từ ý tưởng cải thiện sức khỏe người dùng, nhóm nghiên cứu đã phát triển một website theo dõi thực phẩm ăn hằng ngày. Dự án sử dụng bộ dữ liệu trích xuất từ sách "Bảng Thành Phần Dinh Dưỡng Việt Nam", do Bộ Y Tế - Viện Dinh Dưỡng cung cấp, và tập trung xây dựng mô hình Machine Learning nhằm gợi ý bữa ăn dinh dưỡng phù hợp dựa trên đặc điểm thể trạng của người dùng. Hệ thống cũng hỗ trợ theo dõi lượng calo tiêu thụ dựa trên chỉ số BMI [1].

Dự án áp dụng phương pháp K-Means Clustering để phân cụm các bữa ăn phù hợp, kết hợp với mô hình phân loại thực phẩm theo giá trị dinh dưỡng. Ngoài ra, hệ thống tích hợp bộ công cụ tìm kiếm quen thuộc cùng với chức năng thiết kế bữa ăn theo nhu cầu cá nhân, giúp người dùng dễ dàng theo dõi và điều chỉnh chế độ ăn uống.

Website được xây dựng trên nền tảng Streamlit, giúp triển khai nhanh chóng và tạo ra giao diện thân thiện, dễ sử dụng cho người dùng.

1.7 Ý nghĩa và đóng góp của nghiên cứu

Hệ thống nhằm cung cấp thông tin dinh dưỡng chính xác, đáng tin cậy cho người dùng, hỗ trợ họ trong việc duy trì lối sống lành mạnh và cân đối. Bằng cách quản lý lượng calo và thành phần dinh dưỡng một cách hiệu quả, hệ thống giúp người dùng dễ dàng theo dõi, điều chỉnh chế độ ăn uống phù hợp với nhu cầu cá nhân.

Một điểm đặc biệt của hệ thống là sử dụng bộ dữ liệu thực phẩm dành riêng cho người Việt Nam, được xây dựng dựa trên các nguồn đáng tin cậy, bao gồm Viện Nghiên cứu Dinh dưỡng TP. Hồ Chí Minh. Bộ dữ liệu này tổng hợp đa dạng các loại thực phẩm phổ biến trong bữa ăn hàng ngày của người Việt, từ các món ăn truyền thống đến thực phẩm công nghiệp hiện đại. Điều này giúp người dùng tra cứu linh hoạt giá trị dinh dưỡng của từng món ăn, từ đó đưa ra các lựa chọn thực phẩm phù hợp với mục tiêu sức khỏe của mình.

Ngoài ra, hệ thống còn hướng đến cân bằng dinh dưỡng tổng thể, đảm bảo cung cấp đầy đủ các nhóm chất cần thiết như protein, chất béo, carbohydrate, vitamin và khoáng chất. Nhờ tích hợp trí tuệ nhân tạo (AI) và Machine Learning, hệ thống có khả năng phân tích thói quen ăn uống của người dùng, đưa ra các đề xuất tối ưu hóa chế độ ăn hằng ngày, phù hợp với mục tiêu cá nhân như giảm cân, tăng cơ hay duy trì sức khỏe.

Với các tính năng trên, hệ thống không chỉ hỗ trợ cá nhân hóa chế độ dinh dưỡng, mà còn góp phần nâng cao nhận thức về dinh dưỡng khoa học, giúp người Việt hình thành thói quen ăn uống lành mạnh và cải thiện chất lượng cuộc sống một cách bền vững.

Chương 2

Cơ sở lý thuyết

Chương này mô tả tổng quan tình hình nghiên cứu trong và ngoài nước, khái niệm nghiên cứu và lý thuyết nền tảng.

2.1 Tổng quan tình hình nghiên cứu ngoài nước

Đời sống ngày càng hiện đại, chính vì thế mà mọi người trên ngày càng quan tâm tới sức khoẻ họ nhiều hơn, đặc biệt là chế độ ăn uống hằng ngày. Do đó, mà nhiều nghiên cứu trên thế giới đã và đang được thực hiện nhằm phát triển các hệ thống hỗ trợ người dùng cải thiện hơn về chế độ ăn của họ. Một số nghiên cứu được biết đến như bài báo “Food Recommendation System Using Clustering Analysis for Diabetic Patients”[2]. Hệ thống đưa ra gợi ý thực phẩm dành cho các bệnh nhân tiểu đường ở Thái Lan, sử dụng phương pháp Clustering để phân tích dữ liệu và nghiên cứu dựa trên yếu tố dinh dưỡng, chỉ số đường huyết của bệnh nhân.

Ngoài ra, một nghiên cứu khác đáng quan tâm như “Nutrition-Related Mobile Application for Daily Dietary Self-Monitoring” [3]. Nghiên cứu này tập trung phát triển ứng dụng di động giúp người dùng theo dõi chế độ ăn hằng ngày bằng phương pháp đánh giá tài liệu định tính (Quality Literature Review). Bên cạnh đó các nghiên cứu khác như “Personalizedself-monitoring of energy balance through integration in a web-application of dietary, anthropometric, and physicalactivity data” [4], “Personalized Diet Recommendation System Using Machine Learning” [5] và nhiều nghiên cứu khác.

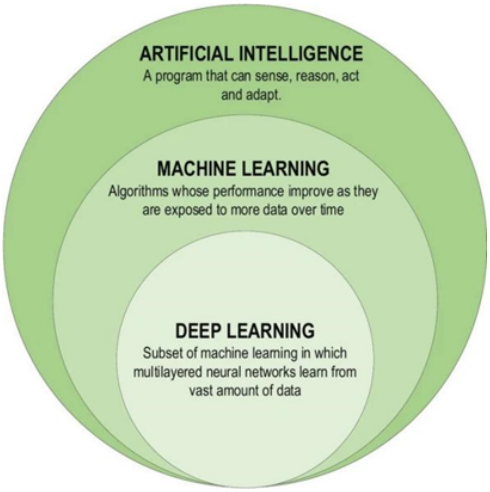
2.2 Tổng quan tình hình nghiên cứu trong nước

Tại Việt Nam, các nghiên cứu về sức khỏe người tiêu dùng cũng nhận được sự quan tâm đông đảo từ các chuyên gia. Một phần mềm ứng dụng di động nổi bật như Nuti Expert do viện nghiên cứu dinh dưỡng TP. Hồ Chí Minh sáng lập, giúp người Việt Nam dễ dàng tiếp cận với sổ tay sức khỏe hơn. Hoặc bài nghiên cứu "Development of the Vietnamese Healthy Eating Index "[6] đánh giá thói quen ăn uống của người Việt Nam nhằm nâng cao ý thức về ăn uống phù hợp và một số nghiên cứu khác.

2.3 Khái niệm nghiên cứu

Phần này trình bày các khái niệm cơ bản về Machine Learning và các thuật toán được áp dụng trong nghiên cứu và xây dựng website theo dõi thực phẩm ăn hằng ngày.

2.3.1 Machine Learning là gì?



Hình 2.3.1: Mô hình Deep Learning

Machine Learning (ML) là một nhánh của Trí tuệ nhân tạo dựa trên việc có thể giúp máy tính và hệ thống mô phỏng theo cách con người học tập, thực hiện các nhiệm vụ một cách tự động, đồng thời cải thiện hiệu suất và độ chính xác thông qua kinh nghiệm và việc tiếp xúc với nhiều dữ liệu hơn [7].
Machine Learning đuợc chia thành nhiều loại dựa trên cách mô hình học từ dữ liệu và phổ biến nhất là 3 loại chính, cụ thể:

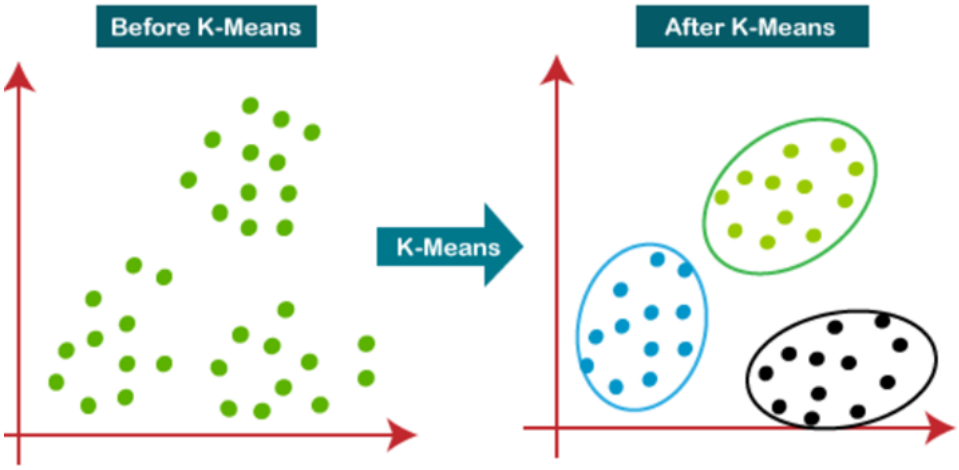
- Supervised Learning: Là môt dạng trong mô hình học máy nhằm thu nhận thông tin về mối quan hệ giữa đầu vào và đầu ra của một hệ thống dựa trên một tập hợp các mẫu huấn luyện có cặp đầu vào - đầu ra được cho trước [8]
- Unsupervised Learning: Trái ngược với Supervised Learning, mô hình không có nhãn được cho trước, thay vào đó hệ thống tự tìm kiếm các mẫu, cấu trúc hoặc mối quan hệ tiềm ẩn trong dữ liệu.
- Reinforcement Learning: Là một loại quy trình học máy tập trung đưa ra quyết định bởi các tác nhân tự động mà không cần sự chỉ dẫn trực tiếp từ người sử dụng [9].

2.3.2 Một số thuật toán Machine Learning tiêu biểu

2.3.2.1. K-Means Clustering

K-Means là một loại thuật toán gom cụm được đề xuất bởi J.B. MacQueen, thuật toán này thuộc loại Unsupervised Learning và thường được sử dụng trong khai phá dự liệu hay nhận dạng mẫu dự liệu [10].

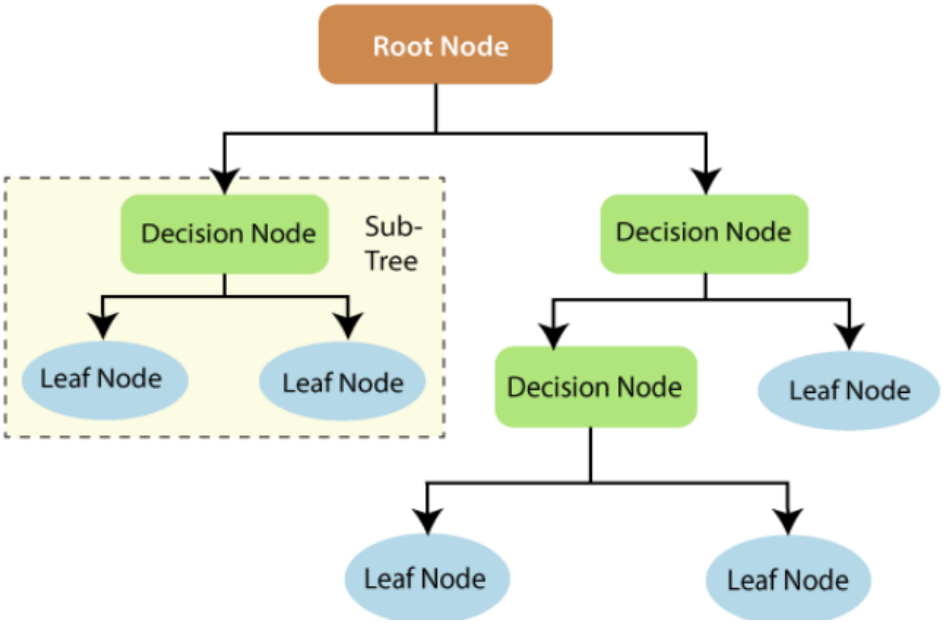
Thuật toán K-Means được biểu diễn ở hình 2.3.2 bắt đầu với việc chọn ra ngẫu nhiên điểm trung tâm hay còn gọi là centroid trong mỗi cụm K được truyền, và mỗi điểm dữ liệu sẽ được gán cho centroid gần nhất tạo thành một cụm. Sau khi thành một cụm, các centroid được cập nhập bằng cách tìm vị trí trung bình của các điểm trong mỗi cụm. Quá trình này lặp cho đến khi centroid không còn thay đổi về giá trị điểm. Và mục tiêu của bài toán K-Means là chia các đối tượng được phân phối vào K cụm, sao cho các đối tượng điểm dữ liệu có sự tương đồng cao và sự tương đồng thấp giữa các cụm.



Hình 2.3.2.1: Mô hình biểu diễn phân cụm bằng thuật toán K-Means

2.3.2.2. Decision Tree Algorithm

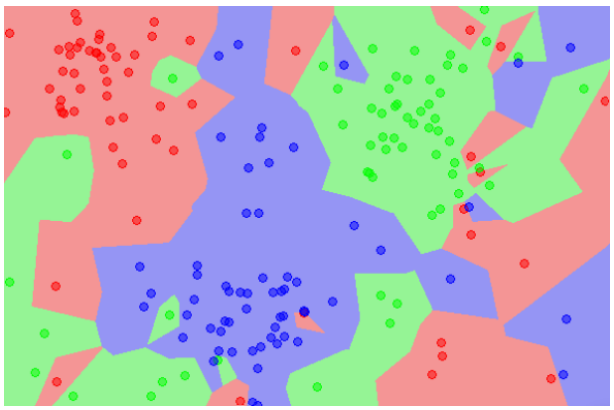
Decision Tree (Cây quyết định) là một thuật toán học máy có giám sát, được sử dụng rộng rãi trong các bài toán phân loại (Classification) [11] và hồi quy (Regression) [12]. Việc chia dữ liệu thành các nhánh với mục đích chia để trị (divide and conquer) với mỗi nút trong cây là một câu hỏi hoặc điều kiện để phân tách dữ liệu, quá trình chia nhánh diễn ra tiếp tục cho đến khi dữ liệu trong một nhánh thuộc về cùng một nhóm hoặc đạt đến điều kiện dừng. Từ đó tạo ra một cấu trúc cây giúp đưa ra quyết định một cách trực quan và dễ hiểu, phù hợp với dữ liệu có đặt trừng rời rạc và liên tục [13].



Hình 2.3.2.2: Hình minh họa cấu trúc của một Decision Tree

2.3.2.3. K-nearest Neighbours

K-Nearest Neighbors (KNN) là một thuật toán học máy có giám sát dễ triển khai, sử dụng khoảng cách giữa các điểm dữ liệu để phân loại hoặc dự đoán giá trị mới dựa trên lớp chiếm đa số của K láng giềng gần nhất của nó. Khi có một điểm dữ liệu mới, thuật toán sẽ tìm K điểm dữ liệu gần nhất trong tập huấn luyện (K là một siêu tham số được lập trình viên chọn) và gán nhãn cho điểm mới dựa trên số lượng nhãn của các điểm lân cận đó.



Hình 2.3.2.3: Mô hình minh họa một bài toán phân loại sử dùng K-nearest Neighbors (KNN).

Hình 2.3.2.3, là một ví dụ cụ thể về bài toán sử dụng thuật toán K-nearest Neighbors (KNN). Giả sử điểm mới nằm ở khu vực giữa vùng xanh dương và xanh lá. Chúng ta chọn K = 5 (xem xét 5 láng giềng gần nhất). Sau đó thực hiện tính khoảng cách từ điểm mới đến tất cả các điểm trong tập dữ liệu chọn 5 điểm gần nhất, giả sử kết quả là: 3 điểm xanh dương và 2 điểm xanh lá => Nhãn được dự đoán: xanh dương (vì đa số trong 5 láng giềng gần nhất là xanh dương)

2.3.2.4. Naive Baives

Thuật toán Naive Bayes là một thuật toán phân loại dựa trên xác suất, được phát triển dựa trên định lý Bayes, được đặt theo tên của nhà toán học Thomas Bayes (1702-1761).

- Công thức Bayes: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Trong đó:
  - P(AB) là xác suất xảy ra sự kiện A khi đã biết sự kiện B.
  - P(BA) là xác suất xảy ra sự kiện B khi đã biết sự kiện A.
  - P(A) và P(B) lần lượt là xác suất tiên nghiệm của A và B.

Thuật toán này được gọi là "Naive"(ngây thơ) vì giả định rằng các đặc trưng của dữ liệu độc lập với nhau, mặc dù trong thực tế, điều này không phải lúc nào cũng đúng. Thuật toán này có nhiều biến thể khác nhau, bao gồm Gaussian Naïve Bayes (cho dữ liệu liên tục), Multinomial Naïve Bayes (cho dữ liệu rời rạc như phân loại văn bản) và Bernoulli Naïve Bayes (cho dữ liệu nhị phân) [14]. Thuật toán giả định về tính độc lập của các đặc trưng là đặc điểm nổi bật có tốc độ xử lý nhanh và khả năng hoạt động tốt ngay cả với lượng dữ liệu nhỏ. Tuy có thể không hoàn toàn chính xác, Naïve Bayes vẫn hoạt động rất tốt trong thực tế, đặc biệt với các bài toán phân loại có số lượng dữ liệu lớn [15].

2.3.3 Chỉ số dinh dưỡng cá nhân

2.3.3.1. BMR (tỷ lệ trao đổi chất cơ bản) dành cho người châu Á, với công thức

Chỉ số BMR (Basal Metabolic Rate) được đề xuất bởi hai nhà khoa học James Arthur Harris và Francis Gano Benedict vào đầu thế kỷ 20. Là lượng calo tiêu thụ tối thiểu cần thiết cho cơ thể trong một ngày như hô hấp, tuần hoàn máu, duy trì nhiệt độ cơ thể. Các yếu tố ảnh hưởng đến chỉ số BMR bao gồm: Khối lượng cơ bắp, kích thước cơ thể, tuổi tác, nội tiết tố, giới tính, di truyền, hoạt động thể chất, chế độ ăn uống. Trong đó, có hai cách tính BMR là phương pháp tính BMR theo cơ chế điện trở (BIA) dựa vào khối lượng các nhóm: cơ, mỡ, nước, xương,... và Phương pháp tính BMR theo Harris- Benedict (cân nặng: kg, chiều cao: cm và tuổi: năm):

- **Công thức Harris- Benedict:**
  - **Nữ giới:** $BMR = 655 + [9.6 \times \text{Cân nặng (kg)}] + [1.8 \times \text{chiều cao(cm)}] - (4.7 \times \text{số tuổi})$.
  - **Nam giới:** $BMR = 66 + [13.7 \times \text{Cân nặng (kg)}] + [5 \times \text{chiều cao(cm)}] - (6.8 \times \text{số tuổi})$.

2.3.3.2. Tính BMI (Chỉ số khối lượng cơ thể người).

BMI (Body Mass Index) là chỉ số này do nhà khoa học người Bỉ Adolphe Quetelet đề xuất vào năm 1832. Được sử dụng để đánh giá cơ bản mức độ gầy hay béo của một người dựa trên cân nặng và chiều cao. Công thức tính BMI:

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

Dựa vào thang bảng phân loại của Hiệp hội đái tháo đường các nước châu Á (IDI WPRO) [18] [16] dành cho người châu Á như sau:

Phân loại	WHO BMI (kg/m²)	IDI & WPRO BMI (kg/m²)
Gầy độ 3	< 16	< 16
Gầy độ 2	16 - 16.9	16 - 16.9
Gầy độ 1	17 - 18.4	17 - 18.4
Bình thường	**18.5 - 24.9**	**18.5 - 22.9**
Tiền béo phì (Thừa cân)	25 - 29.9	23 - 24.9
Béo phì độ I	30 - 34.9	25 - 29.9
Béo phì độ II	35 - 39.9	≥ 30
Béo phì độ III	≥ 40	

Bảng 2.1: Đánh giá và phân loại tình trạng dinh dưỡng dựa vào chỉ số BMI

2.3.3.3. Tính TDEE (Tổng năng lượng tiêu thụ mỗi ngày)

TDEE (Total Daily Energy Expenditure) là tổng năng lượng tiêu hao hàng ngày của một người, là lượng năng lượng được cơ thể con người đốt cháy. Bao gồm tất cả lượng calo đốt cháy trong một ngày từ các hoạt động sinh hoạt, làm việc, tập luyện và trao đổi chất cơ bản. Công thức tính TDEE:

$$TDEE = BRM \times R.$$

Trong đó, R phụ thuộc vào tần suất vận động của mỗi người.
Từ kết quả TDEE và bảng 2.2 từ Tổ chức Y tế Thế giới (WHO) [17]

Mức độ vận động	Mô tả	Hệ số
Ít vận động	Người ít hoặc không tham gia hoạt động thể chất	1.2
Vận động nhẹ	Vận động thể chất, tập thể dục 1-3 ngày/tuần.	1.375
Vận động vừa phải	Vận động thể chất, tập thể dục 3-5 ngày/tuần.	1.55
Vận động nhiều	Vận động thể chất, tập thể dục 6-7 ngày/tuần.	1.725
Vận động rất nhiều	Vận động, tập thể dục hơn 90 phút mỗi ngày hoặc làm công việc nặng.	1.9

Bảng 2.2: Bảng phân loại mức độ vận động và hệ số tương ứng

2.4 Lý thuyết nền tảng

2.4.1 K-means clustering

K-Means là thuật toán phân cụm phổ biến, giúp chia tập dữ liệu thành K cụm dựa trên sự tương đồng giữa các điểm dữ liệu. Ý tưởng chính là tìm K centroid (tâm cụm) sao cho tổng khoảng cách từ các điểm dữ liệu đến tâm cụm của chúng là nhỏ nhất.
Cho tập dữ liêu $X = \{x_1, x_2, \ldots, x_n\}$ gồm n điểm dữ liệu trong không gian d chiều, thuật toán K-Means sẽ nhóm chúng vào K cụm $C_1, C_2, ..., C_k$
Hàm mất mát:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

Trong đó, $||x_i - \mu_k||^2$ là khoảng cách Euclidean bình phương giữa điểm x_i và tâm cụm μ_k. C_k là tập hợp các điểm thuộc cụm thứ k. μ_k là tâm của cụm C_k được tính bằng trung bình cộng của tất cả các điểm trong cụm:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

2.4.2 Decision Tree Algorithm

Trong Decision Tree Algorithm mỗi nhánh được chọn sao cho tối ưu về mặt thông tin, tức là giảm sự hỗn loạn (impurity) trong dữ liệu.

- Entropy - Đo lường mức độ hỗn loạn của một tập dữ liệu:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

  - Trong đó: S là tập dữ liệu đang xét, c là số lớp (class), p_i là xác suất của lớp thứ i

- Gini Impurity - Đo lường xác suất một điểm bị phân loại sai:

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2$$

2.4.3 K-nearest Neighbours

Nguyên tắc chính của KNN là xác định đầu ra của một điểm dữ liệu mới dựa trên các điểm lân cận gần nhất trong tập huấn luyện.

2.4.3.1 Khoảng cách giữa các điểm dữ liệu

KNN dựa vào khoảng cách giữa điểm cần dự đoán và các điểm lân cận. Một số cách đo khoảng cách phổ biến:

- Khoảng cách Euclidean:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

- Khoảng cách Manhattan

$$d(x_i, x_j) = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$

- Khoảng cách Minkowski (Tổng quát hóa Euclidean Manhattan)

$$d(x_i, x_j) = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}$$

2.4.3.2 Dự đoán với thuật toán KNN

- Phân loại (Classification)

Nhãn của điểm cần dự đoán x' được xác định bằng nhãn của K K điểm lân cận có khoảng cách nhỏ nhất.Dùng majority voting:

$$y' = \arg\max_{c} \sum_{i \in N_k(x')} 1(y_i = c)$$

Trong đó, $N_k(x')$ là tập K điểm lân cận của x'. $1(y_i = c)$ là hàm chỉ báo, bằng 1 nếu y_i thuộc lớp c, ngược lại bằng 0. y' là nhãn được dự đoán.

- Hồi quy (Regression)

Giá trị dự đoán là trung bình của K K điểm lân cận:

$$\hat{y} = \frac{1}{K} \sum_{i \in N_k(x')} y_i$$

2.4.4 Naive Baives

Thuật toán Naive Bayes là một thuật toán phân loại dựa trên xác suất, được phát triển dựa trên định lý Bayes, được đặt theo tên của nhà toán học Thomas Bayes (1702-1761).

- Công thức Bayes: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Trong đó:
  - P(AB) là xác suất xảy ra sự kiện A khi đã biết sự kiện B.
  - P(BA) là xác suất xảy ra sự kiện B khi đã biết sự kiện A.
  - P(A) và P(B) lần lượt là xác suất tiên nghiệm của A và B.

Với mỗi kiểu dữ liệu khác nhau sẻ có biến thể khác nhau. Trong đó có ba loại biến thể chính:

- Gaussian Naïve Bayes (GNB): Dùng cho dữ liệu liên tục có phân phối chuẩn:
  - Công thức Xác suất điều kiện: $P(X_j|C_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_j-\mu)^2}{2\sigma^2}\right)$
  - Trong đó:
    - μ là giá trị trung bình của đặc trưng X_j trong lớp C_i.
    - σ^2 là phương sai của đặc trưng X_j trong lớp C_i.
- Multinomial Naïve Bayes (MNB): Dùng cho dữ liệu rời rạc, đặc biệt phù hợp với bài toán phân loại văn bản.
  - Công thức Xác suất điều kiện: $P(X|C_i) = \frac{N!}{X_1!X_2!...X_n!} \prod_{j=1}^{n} P(X_j|C_i)^{X_j}$
  - Trong đó:
    - X_j là số lần xuất hiện của đặc trưng j.
    - $P(X_j|C_i)$ là xác suất của đặc trưng j trong lớp C_i.
- Bernoulli Naïve Bayes: Dùng khi đặc trưng là nhị phân (0 hoặc 1).
  - Công thức Xác suất điều kiện: $P(X|C_i) = \prod_{j=1}^{n} P(X_j|C_i)^{X_j}(1 - P(X_j|C_i))^{(1-X_j)}$

Chương 3

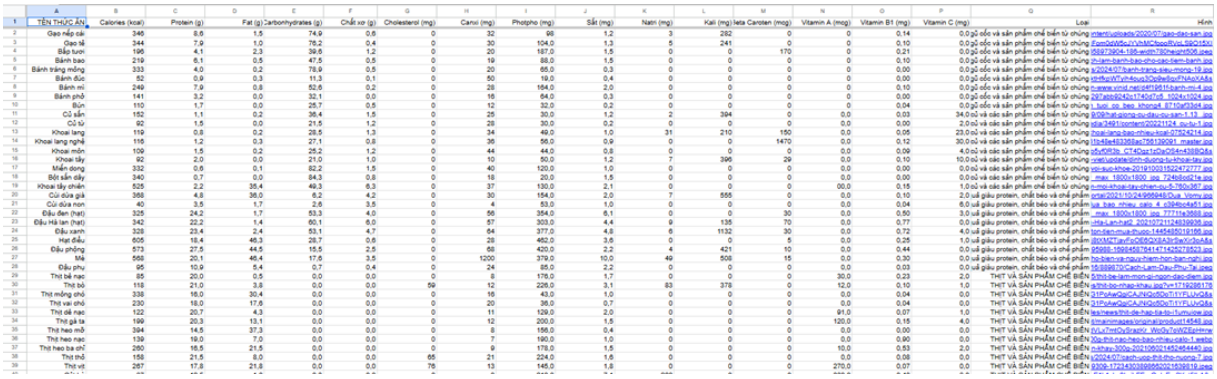
Phương pháp nghiên cứu

Chương này mô tả quy trình nghiên cứu và phương pháp nghiên cứu.

3.1 Quy trình nghiên cứu

3.1.1 Chuẩn bị dữ liệu

Nhóm đã thu thập bộ dữ liệu dinh dưỡng [20] Hình 3.1.1. Bộ dữ liêu được trích xuất và thu thập từ tài liệu “Bảng thành phần dinh dưỡng Việt Nam"[21] và bảng của tài liệu "Thành phần dinh dưỡng một số thức ăn nhanh"[22]. Dữ liệu có 533 dòng, 18 cột và mỗi dòng đại diện cho một loại thực phẩm Việt Nam.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	TÊN THỰC ĂN	Calories (kcal)	Protein (g)	Fat (g)	Carbonhydrates (g)	Chất xơ (g)	Cholesterol (mg)	Canxi (mg)	Photpho (mg)	Sắt (mg)	Natri (mg)	Kali (mg)	leta Caroten (mcg)	Vitamin A (mcg)	Vitamin B1 (mg)	Vitamin C (mg)	Loại	Hình
2	Gạo nếp cái	346	8,6	1,5	74,9	0,6	0	32	98	1,2	3	282	0	0	0,14		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
3	Gạo tẻ	344	7,9	1,0	76,2	0,4	0	30	104,0	1,3	5	241	0	0	0,10		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
4	Bắp tươi	196	4,1	2,3	39,6	1,2	0	20	187,0	1,5	0	0	170	0	0,21		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
5	Bánh bao	219	6,1	0,5	47,5	0,5	0	19	88,0	1,5	0	0	0	0	0,00		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
6	Bánh tráng mỏng	333	4,0	0,2	78,9	0,5	0	20	65,0	0,3	0	0	0	0	0,00		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
7	Bánh đúc	52	0,9	0,3	11,3	0,1	0	50	19,0	0,4	0	0	0	0	0,00		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
8	Bánh mì	249	7,9	0,8	52,6	0,2	0	28	164,0	2,0	0	0	0	0	0,10		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
9	Bánh phở	141	3,2	0,0	32,1	0,0	0	18	64,0	0,3	0	0	0	0	0,00		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
10	Bún	110	1,7	0,0	25,7	0,5	0	12	32,0	0,2	0	0	0	0	0,04		0,0 ngũ cốc và sản phẩm chế biến từ chúng	
11	Củ sắn	152	1,1	0,2	36,4	1,5	0	25	30,0	1,2	2	394	0	0,0	0,03		34,0 củ và các sản phẩm chế biến từ chúng	
12	Củ từ	92	1,5	0,0	21,5	1,2	0	28	30,0	0,2	0	0	0	0,0	0,00		2,0 củ và các sản phẩm chế biến từ chúng	
13	Khoai lang	119	0,8	0,2	28,5	1,3	0	34	49,0	1,0	31	210	150	0,0	0,05		23,0 củ và các sản phẩm chế biến từ chúng	
14	Khoai lang nghệ	116	1,2	0,3	27,1	0,8	0	36	56,0	0,9	0	0	1470	0,0	0,12		30,0 củ và các sản phẩm chế biến từ chúng	
15	Khoai môn	109	1,5	0,2	25,2	1,2	0	44	44,0	0,8	0	0	0	0,0	0,09		4,0 củ và các sản phẩm chế biến từ chúng	
16	Khoai tây	92	2,0	0,0	21,0	1,0	0	10	50,0	1,2	7	396	29	0,0	0,10		10,0 củ và các sản phẩm chế biến từ chúng	
17	Miến dong	332	0,6	0,1	82,2	1,5	0	40	120,0	1,0	0	0	0	0,0	0,00		0,0 củ và các sản phẩm chế biến từ chúng	
18	Bột sắn dây	340	0,7	0,0	84,3	0,8	0	18	20,0	1,5	0	0	0	0,0	0,00		0,0 củ và các sản phẩm chế biến từ chúng	
19	Khoai tây chiên	525	2,2	35,4	49,3	6,3	0	37	130,0	2,1	0	0	0	00,0	0,15		1,0 củ và các sản phẩm chế biến từ chúng	
20	Củ dừa già	368	4,8	36,0	6,2	4,2	0	30	154,0	2,0	7	555	0	0,0	0,10		2,0 uả giàu protein, chất béo và chế phẩm	
21	Củ dừa non	40	3,5	1,7	2,6	3,5	0	4	53,0	1,0	0	0	0	0,0	0,04		6,0 uả giàu protein, chất béo và chế phẩm	
22	Đậu đen (hạt)	325	24,2	1,7	53,3	4,0	0	56	354,0	6,1	0	0	30	0,0	0,50		3,0 uả giàu protein, chất béo và chế phẩm	
23	Đậu Hà lan (hạt)	342	22,2	1,4	60,1	6,0	0	57	303,0	4,4	9	135	70	0,0	0,77		0,0 uả giàu protein, chất béo và chế phẩm	
24	Đậu xanh	328	23,4	2,4	53,1	4,7	0	64	377,0	4,8	6	1132	30	0,0	0,72		4,0 uả giàu protein, chất béo và chế phẩm	
25	Hạt điều	605	18,4	46,3	28,7	0,6	0	28	462,0	3,6	0	0	5	0,0	0,25		1,0 uả giàu protein, chất béo và chế phẩm	
26	Đậu phộng	573	27,5	44,5	15,5	2,5	0	68	420,0	2,2	4	421	10	0,0	0,44		0,0 uả giàu protein, chất béo và chế phẩm	
27	Mè	568	20,1	46,4	17,6	3,5	0	1200	379,0	10,0	49	508	15	0,0	0,30		0,0 uả giàu protein, chất béo và chế phẩm	
28	Đậu phụ	95	10,9	5,4	0,7	0,4	0	24	85,0	2,2	0	0	0	0,0	0,03		0,0 uả giàu protein, chất béo và chế phẩm	
29	Thịt bê nạc	85	20,0	0,5	0,0	0,0	0	8	178,0	1,7	0	0	0	30,0	0,23	2,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
30	Thịt bò	118	21,0	3,8	0,0	0,0	59	12	226,0	3,1	83	378	0	12,0	0,10	1,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
31	Thịt mỡng chó	338	16,0	30,4	0,0	0,0	0	16	43,0	1,0	0	0	0	0,0	0,04	0,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
32	Thịt vai chó	230	16,0	17,6	0,0	0,0	0	20	36,0	0,7	0	0	0	0,0	0,04	0,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
33	Thịt dê nạc	122	20,7	4,3	0,0	0,0	0	11	129,0	2,0	0	0	0	91,0	0,07	1,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
34	Thịt gà ta	199	20,3	13,1	0,0	0,0	0	12	200,0	1,5	0	0	0	120,0	0,15	4,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
35	Thịt heo mỡ	394	14,5	37,3	0,0	0,0	0	8	156,0	0,4	0	0	0	0,0	0,00	0,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
36	Thịt heo nạc	139	19,0	7,0	0,0	0,0	0	7	190,0	1,0	0	0	0	0,0	0,90	0,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
37	Thịt heo ba chỉ	260	16,5	21,5	0,0	0,0	0	9	178,0	1,5	0	0	0	10,0	0,53	2,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
38	Thịt thỏ	156	21,5	8,0	0,0	0,0	65	21	224,0	1,6	0	0	0	0,0	0,08	0,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	
39	Thịt vịt	267	17,8	21,8	0,0	0,0	76	13	145,0	1,8	0	0	0	270,0	0,07	0,0	THỊT VÀ SẢN PHẨM CHẾ BIẾN	

Hình 3.1.1: Bảng dữ liệu thực phẩm

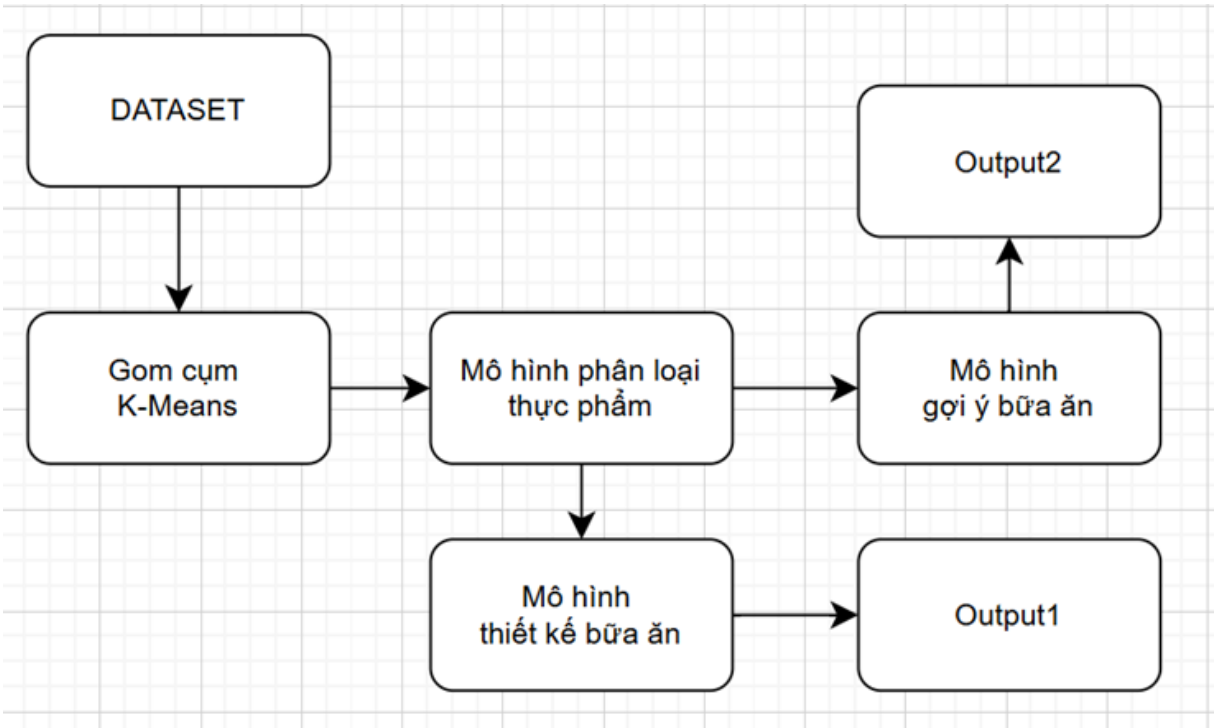
3.1.2 Use Case

Trong hệ thống này với hai tác nhân chính là người dùng (User) và mô hình máy học (ML model), mỗi bên đều đảm nhận những chức năng quan trọng. Người dùng có thể xem bữa ăn được thiết kế theo nhu cầu dinh dưỡng, truy cập gợi ý bữa ăn và tra cứu thông tin thực phẩm. Trong khi đó, mô hình máy học phân tích dữ liệu, dự đoán nhóm dinh dưỡng của thực phẩm và đưa ra gợi ý bữa ăn phù hợp với từng cá nhân.

Hình 3.1.2: Hình thiết kế use case của nghiên cứu

3.1.3 Kiến trúc hệ thống

Tổng quan về cách tổ chức và xây dựng website theo dõi thực phẩm ăn hằng ngày, cho thấy các thành phần tương tác lẫn nhau giữa các luồng dữ liệu và chức năng. Hình 3.1.3.1 biểu diễn kiến trúc hệ thống, trong đó luồng dữ liệu bắt đầu từ Dataset – bộ dữ liệu thực phẩm được thu thập. Tiếp theo, dữ liệu được phân cụm bằng phương pháp K-Means để chia nhóm dinh dưỡng. Sau đó, quá trình phân loại thức ăn được thực hiện bằng mô hình phân loại Machine Learning. Nếu luồng dữ liệu đi mô hình thiết kế bữa ăn sẽ cho ra đầu ra 1, ngược lại nếu luồng dữ liệu vào mô hình gợi ý bữa ăn sẽ cho ra đầu ra 2.



Hình 3.1.3.1: Kiến trúc hệ thống tổng quát

Đầu tiên là quy trình hệ thống thực hiện gợi ý bữa ăn dựa trên thể trạng người dùng.

1. Người dùng vào trang website gợi ý bữa ăn hằng ngày và nhập các trường thông tin cá nhân như độ tuổi hiện tại, giới tính, chiều cao, cân nặng, mục tiêu chế độ ăn, và mức độ vận động hiện tại.

2. Hệ thống tính toán các chỉ số dinh dưỡng cá nhân
3. Hệ thống ghi nhận và các thông tin được xử lý qua mô hình Machine Learning theo trình tự sau:
   3.1 K-Means gom cụm thực phẩm dựa trên các đặc trưng của dinh dưỡng của bộ dữ liệu thực phẩm Việt Nam.

   3.2 Dùng mô hình phân loại thực phẩm bằng Machine Learning để phân loại thực phẩm từ nhãn dữ liệu đã được phân cụm trước đó.
4. Sau khi phân tích dữ liệu của người dùng, hệ thống tính các chỉ số BMR và TDEE cần mỗi ngày, tình trạng cơ thể BMI, để cảnh báo mức độ ăn uống của người dùng và lượng calor có thể tiêu thụ được ở ba bữa ăn chính cụ thể là bữa sáng, bữa trưa, và buổi chiều.
   - Công thức chia lượng calo cho 3 bữa [23]:

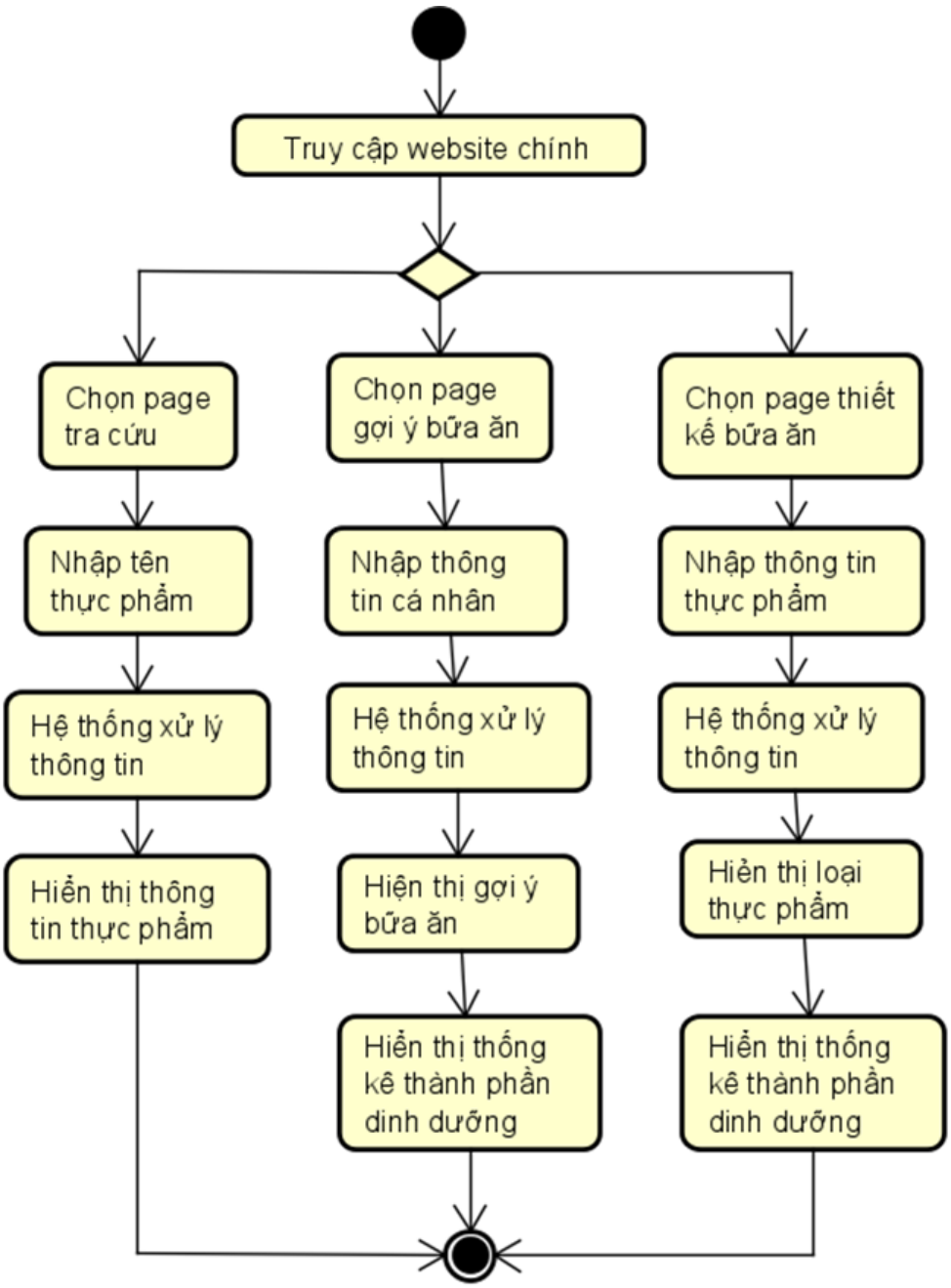
     Lượng calo buổi ăn sáng = Tổng calo hằng ngày x 0,3
     Lượng calo buổi ăn trưa = Tổng calo hằng ngày x 0,4
     Lượng calo buổi ăn chiều = Tổng calo hằng ngày x 0,3
   - Chia nhóm thức ăn theo mục tiêu người dùng:
     - Mục tiêu ‘Tăng cân’: Nhóm Calo cao + Nhóm Calo trung bình.
     - Mục tiêu ‘Duy trì’: Nhóm Calo cao + Nhóm Calo trung bình + Nhóm Calo thấp.
     - Mục tiêu ‘Giảm cân’: Nhóm Calo thấp.
5. Hệ thống dựa trên lượng calor được quy định ở ba buổi ăn chính và mục tiêu chế độ ăn người dùng, sau đó hiện thị gợi ý ba bữa ăn cho người dùng trong một ngày, và cho người dùng thấy bảng so sánh thống kê giữa các giá trị dinh dưỡng của mỗi phần ăn.
6. Hệ thống hoàn thiện gợi ý bữa ăn.

Tiếp theo là quy trình thiết kế bữa ăn theo nhu cầu dinh dưỡng của người dùng.

1. Người dùng vào trang thiết kế bữa ăn theo nhu cầu dinh dưỡng.
2. Người dùng thêm món ăn và nhập các trường thông tin về các giá trị dinh dưỡng món ăn mà được người dùng sử dụng.
3. Hệ thống ghi nhận và thông tin được xử lý qua Machine Learning để phân loại thực phẩm theo 3 mức độ dinh dưỡng chính cụ thể là lượng calo thấp, lượng calo vừa và lượng calo cao.
4. Sau đó hệ thống hiển thị các thông tin tổng hợp dinh dưỡng các món ăn từ dữ liệu đầu vào, các giá trị thống kê thành phần dinh dưỡng giúp người dùng cân nhắc về xây dựng bữa ăn của mình.

Cuối cùng là quy trình tra cứu thực phẩm

1. Người dùng vào trang tra cứu thực phẩm.
2. Người dùng nhập tên thực phẩm hoặc loại thực phẩm cần tra cứu
3. Hệ thống ghi nhận và trả truy vấn thực phẩm cần tìm.



Hình 3.1.3.2: Sơ đồ luồng hoạt động của hệ thống.

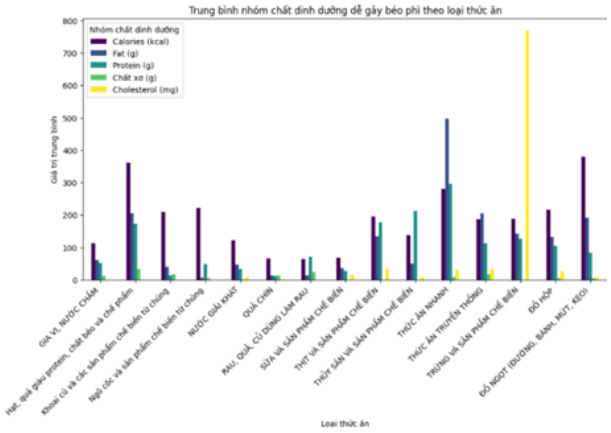
3.2 Phương pháp nghiên cứu

3.2.1 Phân tích dữ liệu

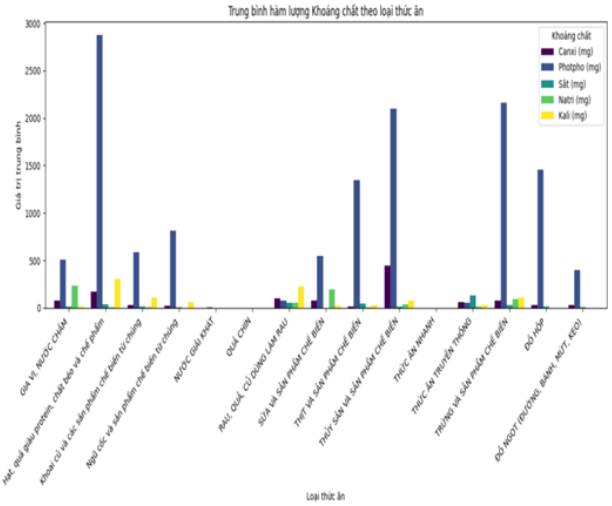
Từ bộ dữ liệu thực phẩm thu thập được, nhóm đã lựa chọn các đặc trưng quan trọng bao gồm: Calories (kcal), Protein (g), Fat (g), Carbohydrates (g), Chất xơ (g), Cholesterol (mg), Canxi (mg), Photpho (mg), Sắt (mg), Natri (mg), Kali (mg), Beta Caroten (mcg), Vitamin A (mcg), Vitamin B1 (mg) và Vitamin C (mg). Các đặc trưng này được phân tích dựa trên các chỉ số dinh dưỡng quan trọng nhằm làm rõ sự khác biệt về các thành phần dinh dưỡng có thể tác động đến sức khỏe người dùng. Từ những giá trị dinh dưỡng này nhóm đã phân tích thành ba nhóm dinh dưỡng chính:

- Nhóm giàu năng lượng (Calories, chất béo, protein, chất xơ, cholesterol) ở hình 3.2.1.1 thường tập trung ở thực phẩm chế biến sẵn, thức ăn nhanh, thịt và các sản phẩm từ thịt, đồ ngọt có thể gây nguy cơ béo phì nếu tiêu thụ nhiều.
- Nhóm giàu khoáng chất (Canxi, Photpho, Sắt, Kali, Natri) ở hình **??**.2 tập trung ở hạt, quả giàu protein, thủy sản, thịt và ngũ cốc giúp hỗ trợ sự phát triển hệ cơ răng xương và tuần hoàn máu.

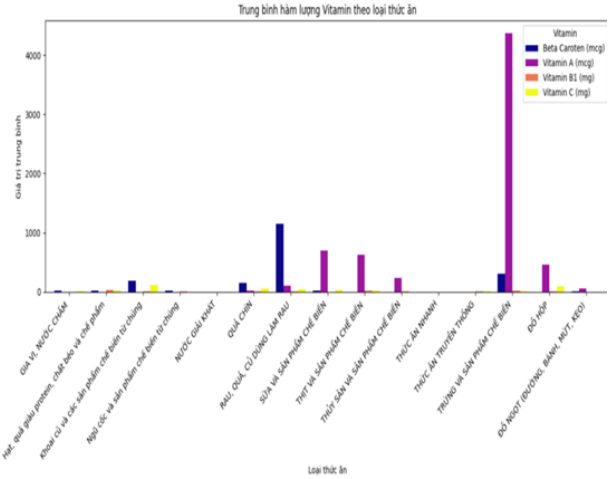
- Nhóm giàu Vitamin (Beta Caroten, Vitamin A, Vitamin B2, Vitamin C) ở hình 3.2.1.3 thường tập trung chủ yếu ở nhóm rau củ quả, sữa và sản phẩm chế biến, trứng và sản phẩm chế biến, quả chín.



Hình 3.2.1.1: Nhóm giàu chất dinh dưỡng



Hình 3.2.1.2: Nhóm khoáng chất



Hình 3.2.1.3: Nhóm Vitamin

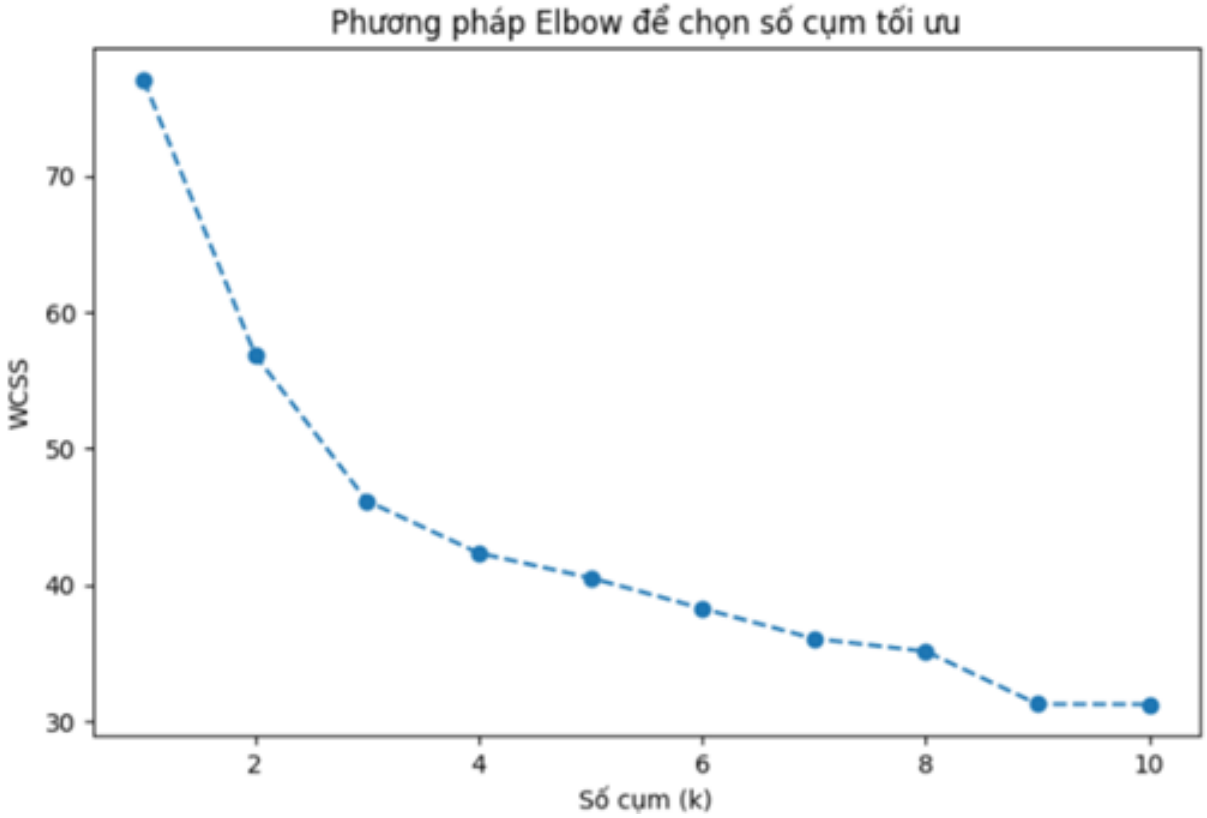
3.2.2 Các phương pháp đã sử dụng.

Dưới đây là phần trình bày về các phương pháp đã sử dụng để gom cụm và phân loại thực phẩm từ đó để có kết quả chọn mô hình tốt nhất.

3.2.2.2 Gom cụm K-Means

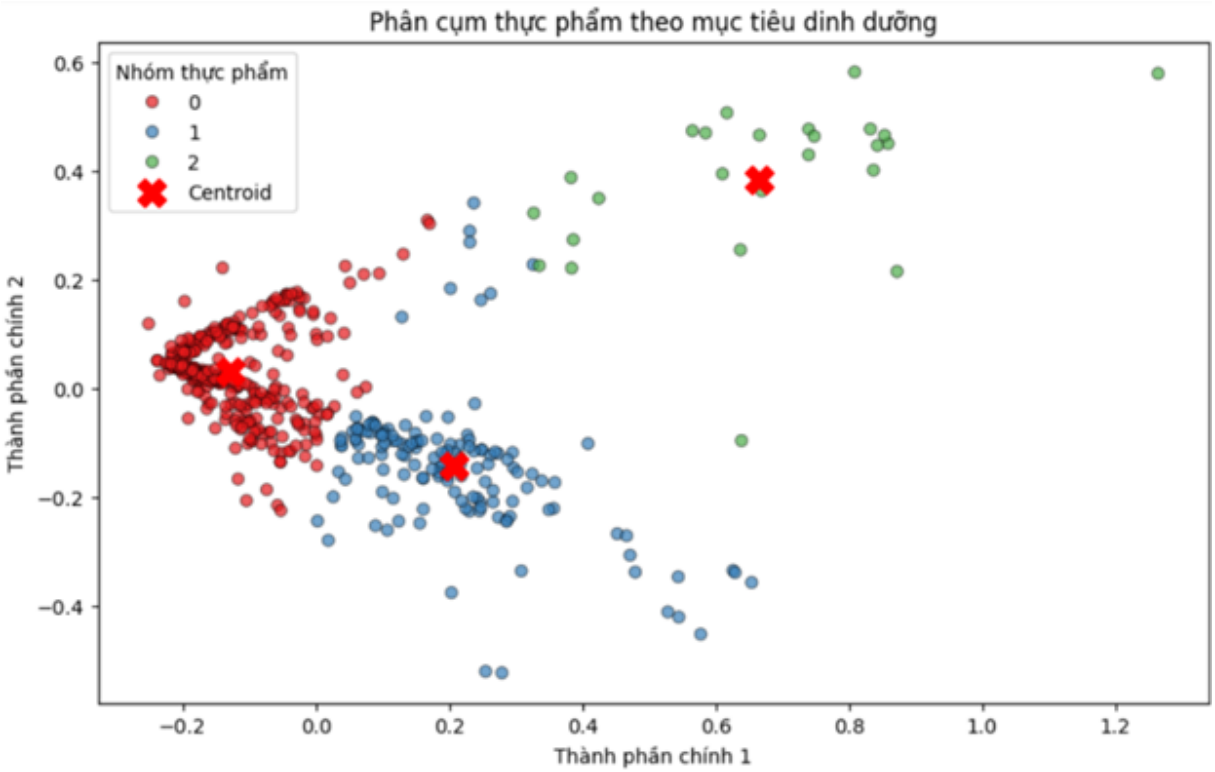
Sau khi tiền xử lý dữ liệu và phân tích bộ dữ liệu, nhóm thực hiện chuẩn hóa dữ liệu bằng phương pháp MinMaxScaler để chuyển đổi các đặc trưng dinh dưỡng về cùng phạm vi từ 0 đến 1 để đảm bảo các biến có đơn vị

đo lường khác nhau không ảnh hưởng đến kết quả phân cụm. Nhóm tiến hành thực hiện chọn số cụm tối ưu bằng phương pháp Elbow, được biết đến là một kỹ thuật xác định số cụm tối ưu trong K-Means bằng việc xem xét tính tổng bình phương khoảng cách trong cụm Within-Cluster Sum of Squares (WCSS). Thực hiện phương pháp Elbow bằng cách chạy thuật toán K-Means với các gía trị k từ 1 đến 10 trên bộ dữ liệu thực phẩm đã được chuẩn bị trước. Tiếp theo, ta tính WCSS cho mỗi giá trị k và vẽ đồ thị 3.2.2 biểu diễn mối quan hệ giữa k và WCSS bằng thư viện matplotlib. Quan sát đồ thị, ta thấy từ k = 1 đến k = 2, WCSS giảm dần đều, nhưng tại k = 3 có sự thay đổi rõ rệt, sau đó tiếp tục giảm ổn định từ k = 4 đến k = 10. Điều này cho thấy k = 3 có thể là số cụm tối ưu.



Hình 3.2.2.2: Phương pháp Elbow để chọn số cụm tối ưu

Nhóm tiến hành phân cụm bộ dữ liệu thực phẩm bằng phương pháp K-Means với đầu vào là số cụm k=3 và bộ dữ liệu thực phẩm, quá trình phân cụm được thực hiện qua trình tự sau: Khởi tạo bằng cách chọn ngẫu nhiên K điểm trong tập dữ liệu để làm tâm cụm ban đầu, sau đó mỗi điểm dữ liệu được gán cụm có tâm gần nhất bằng phương pháp tính khoảng cách Euclidean, tiếp tục tính toán loại tâm cụm mới bằng cách tính trung bình cộng các điểm trong cụm. Quá trình này được lặp lại cho đến khi các giá trị của tâm hay còn gọi là centroid không đổi. Vì dữ liệu thực phẩm có nhiều thuộc tính khác nhau, nhóm áp dụng Phân tích Thành phần Chính (PCA) để giảm số chiều xuống còn 2, giúp trực quan hóa kết quả phân cụm một cách rõ ràng hơn. Biểu đồ 3.2.2 cho thấy sự phân tách thành ba cụm dữ liệu với các đặc trưng rõ ràng.



Hình 3.2.2.2: Phân cụm thực phẩm theo mục tiêu dinh dưỡng

3.2.2.3 Một số phương pháp phân loại.

Sau khi thiết lập nhãn cho từng cụm tương ứng với các nhãn Calo cao, Calo trung bình, Calo cao và nhóm tiến hành áp dùng một số thuật toán phân loại nhằm xây dựng mô hình mô hình dự đoán loại nhóm dinh dưỡng của thực phẩm để thiết kế bữa ăn phù hợp theo nhu cầu người dùng, dựa trên các đặc trưng dinh dưỡng. Dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80:20, để đảm bảo mô hình có thể học tốt. Dưới đây là ba phương pháp phân loại khá phổ biến mà nhóm đã áp dụng.

K-Nearest Neighbors (KNN)

Nhóm áp dụng phương pháp K-Fold Cross Validation để tối ưu hóa mô hình và tránh hiện tượng overfitting, là hiện tượng khi dữ liệu nhiễu hoặc dữ liệu bất thường trong tập huấn luyện đều được chọn để học và đưa ra quy luật của mô hình, dẫn đến việc mô hình có thể hoạt động rất tốt trên tập huấn luyện nhưng lại có độ chính xác kém khi dự đoán trên tập dữ liệu mới.

Để tối ưu hóa mô hình K-Nearest Neighbors (KNN), nhóm đã áp dụng K-Fold Cross Validation với K=10, nghĩa là chia tập dữ liệu thành 10 phần bằng nhau. Mỗi lần kiểm tra, một phần sẽ được sử dụng làm tập kiểm tra, trong khi 9 phần còn lại sẽ được sử dụng để huấn luyện mô hình. Quá trình này được lặp lại 10 lần, với mỗi phần dữ liệu một lần làm tập kiểm tra. Sau khi tất cả các lần kiểm tra hoàn thành, kết quả sẽ được tổng hợp để tính toán độ chính xác trung bình của mô hình.

Trong quá trình thực hiện K-Fold Cross Validation với KNN, nhóm đã thử nghiệm với nhiều giá trị K khác nhau, đại diện cho số lượng láng giềng gần nhất mà thuật toán sử dụng để đưa ra dự đoán cho mỗi điểm dữ liệu. Cụ thể, nhóm thử nghiệm các giá trị K = 3, K = 5 và K = 7 để xác định giá trị K tối ưu giúp mô hình đưa ra dự đoán chính xác nhất.

Naive Bayes

Nhóm sử dụng phương pháp phân loại bằng Naive Bayes, là một phương pháp phân loại dựa trên xác suất, giả định rằng các đặc trưng là độc lập với nhau cho mỗi lớp. Nhóm áp dụng Naive Baives với phân phối chuẩn (Gaussian) để phân loại thực phẩm dựa trên đặc trưng dinh dưỡng. Gaussian Naive Bayes giả định rằng các đặc trưng trong bộ dữ liệu tuân theo phân phối chuẩn, tức là mỗi đặc trưng có thể được mô hình hóa bằng một phân phối chuẩn với trung bình và độ lệch chuẩn riêng biệt cho mỗi lớp. Trong trường hợp này, các lớp được xác định là "Calo cao", "Calo trung bình", và "Calo thấp". Phương pháp tính toán xác suất dựa trên công thức [24]:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

Trong đó :

- $P(C_i|X)$ là xác suất có điều kiện của lớp C_i (Calo cao, Calo trung bình, hoặc Calo thấp) cho một điểm dữ liệu X (một thực phẩm với các đặc trưng dinh dưỡng).
- $P(X|C_i)$ là xác suất của các đặc trưng X trong lớp C_i, được tính bằng phân phối chuẩn (Gaussian).
- $P(Ci)$ là xác suất tiên nghiệm của lớp C_i, được tính từ tần suất xuất hiện của lớp đó trong dữ liệu huấn luyện.
- $P(X)$ là xác suất của các đặc trưng X, dùng để chuẩn hóa các xác suất và không ảnh hưởng đến quyết định cuối cùng.

Để tính P(X|Ci), Naive Bayes giả sử rằng mỗi đặc trưng Xi là độc lập với các đặc trưng khác trong cùng một lớp. Do đó, ta có công thức dưới đây:

$$P(X|C_i) = P(X_1|C_i) \cdot P(X_2|C_i) \cdot \cdots \cdot P(X_n|C_i)$$

Các đặc trưng dinh dưỡng của thực phẩm (ví dụ: calo, protein, carbohydrate) được ký hiệu là X_i, và xác suất $P(X_i|C_i)$ được tính từ phân phối chuẩn với trung bình và độ lệch chuẩn của đặc trưng trong lớp C_i. Lớp C_i có xác suất cao nhất $P(Ci|X)$ sẽ được chọn làm dự đoán cho thực phẩm, giúp phân loại thực phẩm vào các nhóm calo phù hợp dựa trên dữ liệu huấn luyện.

Decision Tree

Nhóm sử dụng phương pháp phân loại bằng Cây Quyết Định (Decision Tree), là một thuật toán phân loại dựa trên việc xây dựng cây phân loại thông qua việc chia nhỏ không gian đặc trưng của dữ liệu. Mỗi nút trong cây đại diện cho một câu hỏi kiểm tra đặc trưng dinh dưỡng của thực phẩm (ví dụ: calo, protein, carbohydrate), và mỗi nhánh sẽ biểu thị kết quả của câu hỏi đó. Quyết định cuối cùng được đưa ra tại các lá của cây, mỗi lá sẽ tương ứng với một nhãn phân loại như "Calo cao", "Calo trung bình"hoặc "Calo thấp".

Trong quá trình huấn luyện cây quyết định, nhóm đã điều chỉnh các tham số quan trọng để tối ưu hóa mô hình, trong đó một tham số quan trọng là max_depth. Tham số này xác định độ sâu tối đa của cây, ảnh hưởng trực tiếp đến khả năng phân loại của mô hình. Nếu độ sâu quá nhỏ, mô hình có thể không học đầy đủ thông tin từ dữ liệu, gây ra hiện tượng underfitting. Ngược lại, nếu độ sâu quá lớn, mô hình có thể học quá kỹ các chi tiết không quan trọng từ dữ liệu huấn luyện, dẫn đến hiện tượng overfitting.

Nhóm thử nghiệm với các giá trị max_depth khác nhau từ 1 đến 10 để tìm ra giá trị tối ưu. Quá trình huấn luyện và đánh giá giúp nhóm chọn được giá trị max_depth tối ưu cho mô hình Decision Tree, từ đó cải thiện khả năng phân loại thực phẩm vào các nhóm calo phù hợp, dựa trên các đặc trưng dinh dưỡng đã được học từ dữ liệu huấn luyện.

3.2.3 Triển khai mô hình trên Streamlit

Streamlit được ra mắt lần đầu vào năm 2019, là một thư viện Python mã nguồn mở được phát triển bởi Adrien Treuille, cùng với Thiago Teixeira và Amanda Kelly. Họ là những chuyên gia về trí tuệ nhân tạo và khoa học dữ liệu, từng làm việc tại Google, Carnegie Mellon, và các công ty công nghệ lớn khác. Giúp phát triển nhanh các ứng dụng web tương tác, đặc biệt phù hợp với các ứng dụng liên quan đến khoa học dữ liệu, machine learning, và hiển thị dữ liệu. Điểm mạnh của Streamlit là đơn giản, dễ sử dụng, không cần kiến thức sâu về lập trình web như HTML, CSS hay JavaScript. [19].

Mô hình hoạt động của Streamlit khác biệt hoàn toàn so với các web framework truyền thống. Thay vì sử dụng mô hình MVC (Model-View-Controller) phức tạp, Streamlit cho phép các nhà phát triển tạo ứng dụng web đơn giản bằng cách viết code Python thuần túy. Khi một script Streamlit được thực thi, nó chạy từ trên xuống dưới, xây dựng giao diện người dùng theo trình tự mà các thành phần UI được khai báo trong code. Điểm đặc biệt là mỗi khi người dùng tương tác với một thành phần (như button, slider, hoặc input), toàn bộ script sẽ được chạy lại từ đầu. Điều này tạo ra một mô hình phản ứng đơn giản nhưng hiệu quả, trong đó giao diện người dùng luôn phản ánh trạng thái hiện tại của dữ liệu và logic ứng dụng. Để tối ưu hiệu suất, Streamlit cung cấp hệ thống caching thông minh giúp tránh việc thực hiện lại các tính toán tốn kém mỗi khi script chạy lại. Nền tảng hoạt động trên kiến trúc client-server, với script Python chạy trên server và giao diện người dùng được hiển thị trên trình duyệt web. Các thành phần UI được tạo ra từ code Python được chuyển đổi thành các định dạng protobuf, sau đó được gửi đến trình duyệt và hiển thị dưới dạng các thành phần web tương tác. Mô hình đơn giản này cho phép các nhà khoa học dữ liệu và các nhà phát triển nhanh chóng tạo ra các ứng dụng tương tác mà không cần kiến thức chuyên sâu về phát triển web.

Chương 4

Kết quả nghiên cứu

Chương này bao gồm các kết quả thu được từ quá trình nghiên cứu

4.1 Các kết quả thu được từ quá trình nghiên cứu

4.1.1 Kết quả phương pháp phân loại

4.1.1.1 K-Nearest Neighbors

Nhóm áp dụng phương pháp K-Fold Cross Validation (K=10) với số lượng K láng giềng khác nhau (K=3,5,7) để tìm ra mô hình có độ chính xác cao nhất. Kết quả thu được từ quá trình đánh giá được thể hiện ở hình **??** cho thấy mô hình có K=3 đạt độ chính xác cao nhất với Accuracy = 0,9037, với các chỉ số Precision, Recall và F1-Score lần lượt là 0,9216, 0,8711 và 0,8867. Tuy nhiên, kết quả từ K=3 lại cho thấy sự phân bố không đồng đều giữa các

Model	Accuracy	Precision	Recall	F1 Score
KNN (K=3)	0.9037	0.9216	0.8711	0.8867
KNN (K=5)	0.8848	0.8619	0.8235	0.8329
KNN (K=7)	0.8754	0.8205	0.7817	0.7908

Bảng 4.1: Kết quả mô hình KNN với các giá trị K khác nhau

nhóm "Calo cao,""Calo thấp"và "Calo trung bình."Mặc dù độ chính xác (Accuracy) của mô hình đạt mức cao với 0,9037, nhưng Precision và Recall của các nhóm không hoàn toàn cân bằng. Cụ thể, với nhóm "Calo cao,"Precision đạt mức tối đa 1.00, nhưng Recall chỉ đạt 0.75, cho thấy mô hình có xu hướng phân loại chính xác nhóm này khi chúng xuất hiện, nhưng vẫn bỏ sót một số trường hợp. Với nhóm "Calo thấp,"Precision đạt 0.89 và Recall đạt 0.99, chỉ ra rằng mô hình phân loại nhóm này rất chính xác nhưng có thể phân loại sai một số ít trường hợp. Tuy nhiên, đối với nhóm "Calo trung bình,"dù Precision khá cao (0.92), nhưng Recall chỉ đạt 0.72, cho thấy mô hình vẫn chưa nhận diện đầy đủ các trường hợp trong nhóm này. Điều này phản ánh rằng mặc dù mô hình K=3 có độ chính xác cao, nhưng cần cải thiện khả năng nhận diện chính xác hơn cho nhóm "Calo trung bình."

Class	Precision	Recall	F1-Score	Support
Calo cao	1.00	0.75	0.86	4
Calo thấp	0.89	0.99	0.93	71
Calo trung bình	0.92	0.72	0.81	32

Bảng 4.2: Bảng kết quả phân loại của KNN với K=3

4.1.1.2 Naive Bayes

Kết quả từ mô hình Naive Bayes cho thấy độ chính xác (Accuracy) đạt 0,8598, với các chỉ số Precision, Recall và F1-Score lần lượt là 0,7184, 0,9009 và 0,7649. Dù mô hình này có độ chính xác vừa phải, tuy nhiên, có sự chênh lệch rõ rệt giữa các nhóm phân loại. Cụ thể, nhóm "Calo cao"có Precision thấp (0.40) nhưng Recall rất cao (1.00), cho thấy mô hình hoàn toàn nhận diện đúng nhóm "Calo cao"khi nó xuất hiện, nhưng lại có tỷ lệ sai cao đối với những mẫu không phải "Calo cao."Với nhóm "Calo thấp,"Precision đạt 0.98 và Recall đạt 0.86, mô hình này cho thấy khả năng phân loại chính xác nhóm "Calo thấp"rất cao, tuy nhiên, vẫn bỏ sót một số trường hợp. Đối với nhóm "Calo trung bình,"Precision là 0.77 và Recall là 0.84, phản ánh mô hình có sự cân bằng tốt giữa việc nhận

diện chính xác và khả năng nhận diện đầy đủ các mẫu trong nhóm này. Mặc dù mô hình Naive Bayes đạt độ chính xác cao, các nhóm "Calo cao"và "Calo thấp"vẫn có sự phân hóa đáng kể trong các chỉ số Precision và Recall, chỉ ra rằng mô hình cần cải thiện khả năng phân loại các nhóm này đồng đều hơn. F1-Score của nhóm "Calo thấp"đạt 0.92, cho thấy mô hình đã làm tốt trong việc phân loại nhóm này.

$$\text{Accuracy} = 0.8598, \quad \text{Precision} = 0.7184, \quad \text{Recall} = 0.9009, \quad \text{F1-Score} = 0.7649$$

Class	Precision	Recall	F1-Score
Calo cao	0.40	1.00	0.57
Calo thấp	0.98	0.86	0.92
Calo trung bình	0.77	0.84	0.81

Bảng 4.3: Báo cáo phân loại cho mô hình Naive Bayes

Macro Avg: Precision = 0.72, Recall = 0.90, F1-Score = 0.76

Weighted Avg: Precision = 0.90, Recall = 0.86, F1-Score = 0.87

4.1.1.3 Decision Tree

Kết quả thực hiện thuật toán Decision Tree với giá trị $MaxDepth$ từ 1 đến 10 cho thấy rằng tại $MaxDepth = 2$, mô hình đạt hiệu suất cao nhất với Accuracy 99.07% và các chỉ số khác như Precision, Recall, F1 Score cũng ở ngưỡng tương tự, điều này cho thấy mô hình có khả năng phân loại tốt nhất ở độ sâu này. Tuy nhiên từ $MaxDepth = 3$ trở đi, các chỉ số ổn định ở mức 98.13%, cho thấy mô hình không bị overfitting nhưng cũng không cải thiện thêm khi tặng độ sâu.

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree (Max Depth=1)	0.9439	0.9903	0.9439	0.9262
Decision Tree (Max Depth=2)	0.9907	0.9908	0.9907	0.9906
Decision Tree (Max Depth=3)	0.9626	0.9644	0.9626	0.9548
Decision Tree (Max Depth=4)	0.9813	0.9814	0.9813	0.9807
Decision Tree (Max Depth=5)	0.9813	0.9814	0.9813	0.9807
Decision Tree (Max Depth=6)	0.9813	0.9814	0.9813	0.9807
Decision Tree (Max Depth=7)	0.9813	0.9814	0.9813	0.9807
Decision Tree (Max Depth=8)	0.9813	0.9814	0.9813	0.9807
Decision Tree (Max Depth=9)	0.9813	0.9814	0.9813	0.9807
Decision Tree (Max Depth=10)	0.9813	0.9814	0.9813	0.9807

Bảng 4.4: Kết quả của mô hình Decision Tree với các giá trị $MaxDepth$ khác nhau

Tiếp tục với mô hình có Max Depth = 2, mô hình đạt precision rất cao (1.00 cho "Calo cao"và gần 1.00 cho "Calo thấp"và "Calo trung bình"), cho thấy mô hình rất chính xác trong việc dự đoán đúng các lớp.Tiếp theo, mô hình có Recall tốt đặc biệt đối với “Calo cao” và “Calo thấp”, cho thấy mô hình không bỏ sót đối tượng trong các lớp này.

Class	Precision	Recall	F1-Score	Support
Calo cao	1.00	1.00	1.00	4
Calo thấp	0.99	1.00	0.99	71
Calo trung bình	1.00	0.97	0.98	32

Bảng 4.5: Kết quả phân loại $MaxDepth = 2$

Báo cáo phân loại cho mô hình tốt nhất $MaxDepth = 2$:

- Accuracy: 0.9907
- Macro avg: Precision = 1.00, Recall = 0.99, F1-Score = 0.99
- Weighted avg: Precision = 0.99, Recall = 0.99, F1-Score = 0.99
- Total support: 107

4.1.2 Giao diện ứng dụng

Sau khi so sánh kết quả, nhóm đã chọn phương pháp gom cụm K-Means để xây dựng mô hình gợi ý bữa ăn theo yêu cầu người dùng và phương pháp phân loại Decision Tree để xây dựng mô hình phân loại thực phẩm, giúp thiết kế bữa ăn phù hợp với nhu cầu người dùng, và ứng dụng được triển khai bằng Streamlit để cung cấp giao diện người dùng trực quan và dễ sử dụng.

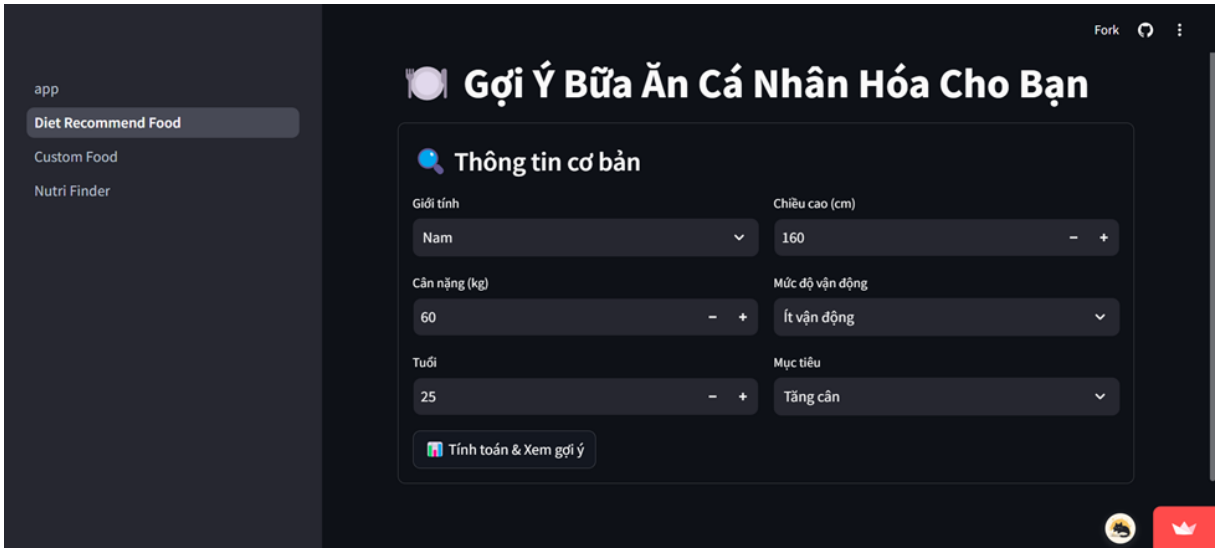
Giao diện trang chủ

Hình 4.1.2.1: Giao diện trang chủ

Trang chủ là nơi trung tâm chào đón người dùng khi họ truy cập lần đầu vào hệ thống. Giao diện thết kế đơn giản và dễ sử dụng, cho phép người dùng có thể truy cập các chức năng chính của hệ thống trực tiếp.

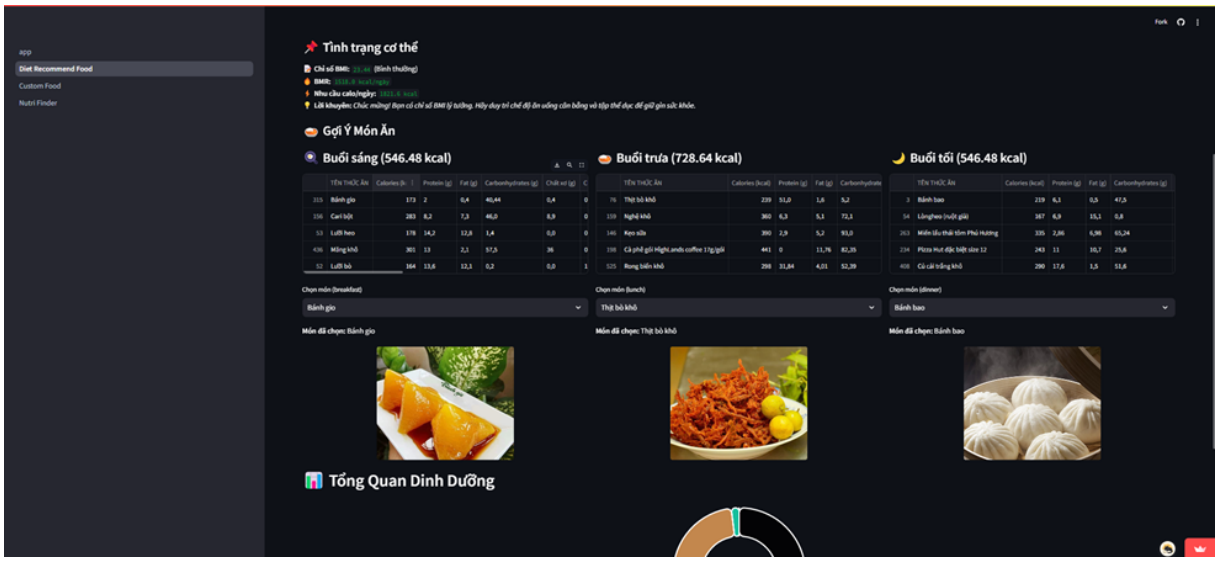
Giao diện gợi ý bữa ăn

Giao diện gợi ý bữa ăn được thiết kế tối giản, thân thiện và phù hợp với người dùng, giúp việc nhập thông tin cá nhân trở nên dễ dàng. Tuy nhiên, hiện tại chưa có cơ chế kiểm tra và ràng buộc dữ liệu để đảm bảo tính hợp lý trong thực tế (ví dụ: cân nặng 60kg nhưng tuổi chỉ là 1). Cần bổ sung các quy tắc xác thực để nâng cao độ chính xác và trải nghiệm người dùng.



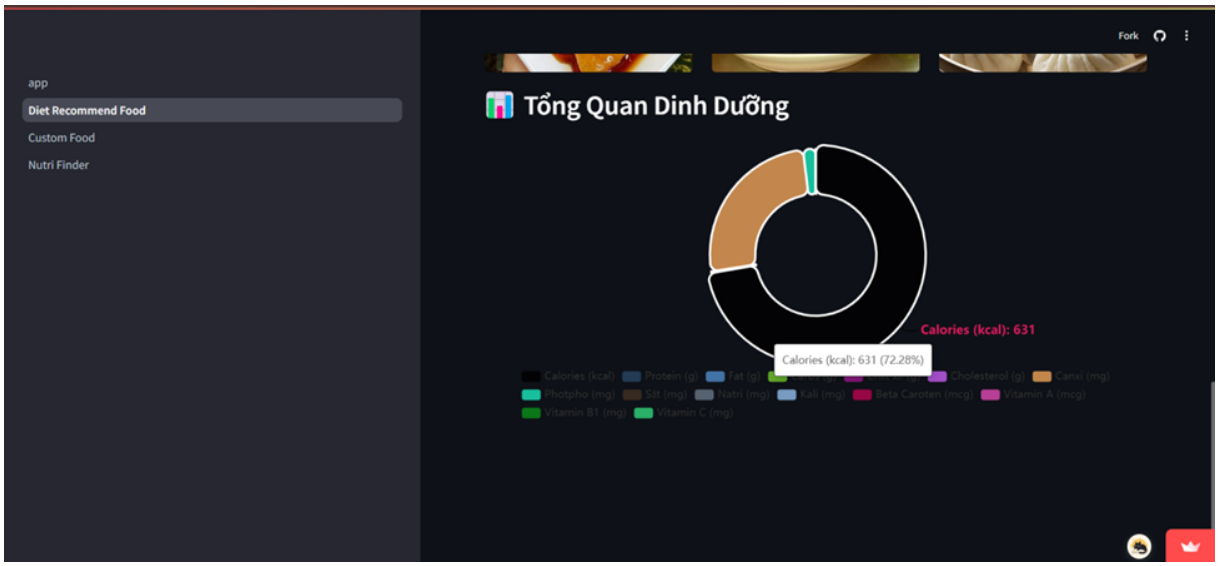
Hình 4.1.2.2: Giao diện nhập đầu vào gợi ý bữa ăn

Khi người dùng nhập thông tin cá nhân và nhấn vào nút "Tính toán Xem gợi ý", hệ thống sẽ tự động phân tích và hiển thị các chỉ số quan trọng về tình trạng cơ thể, bao gồm BMI, BMR, nhu cầu calo/ngày và lời khuyên dinh dưỡng. Ngoài ra, danh sách thực phẩm gợi ý cũng được đề xuất, giúp người dùng dễ dàng lựa chọn món ăn phù hợp. Giao diện được thiết kế trực quan, thân thiện và hỗ trợ người dùng theo dõi thông tin một cách hiệu quả.



Hình 4.1.2.3: Giao diện nạp dữ liệu gợi ý bữa ăn dựa vào thông tin input của người dùng

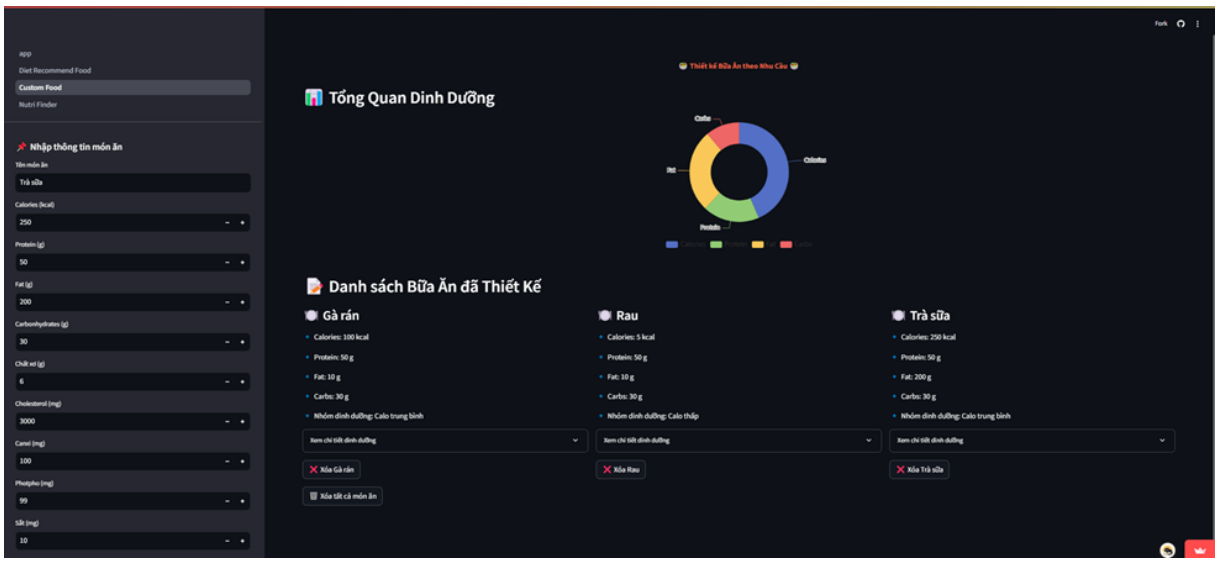
Ngay sau khi người dùng chọn các món ăn gợi ý, hệ thống sẽ tự động hiển thị biểu đồ tròn tổng quan về dinh dưỡng trong ngày. Điều này giúp người dùng dễ dàng theo dõi và điều chỉnh thực đơn để đảm bảo cân bằng dinh dưỡng hợp lý.



Hình 4.1.2.4: Giao diện hiển thị tổng quan dinh dưỡng của người dùng

Giao diện thiết kế bữa ăn

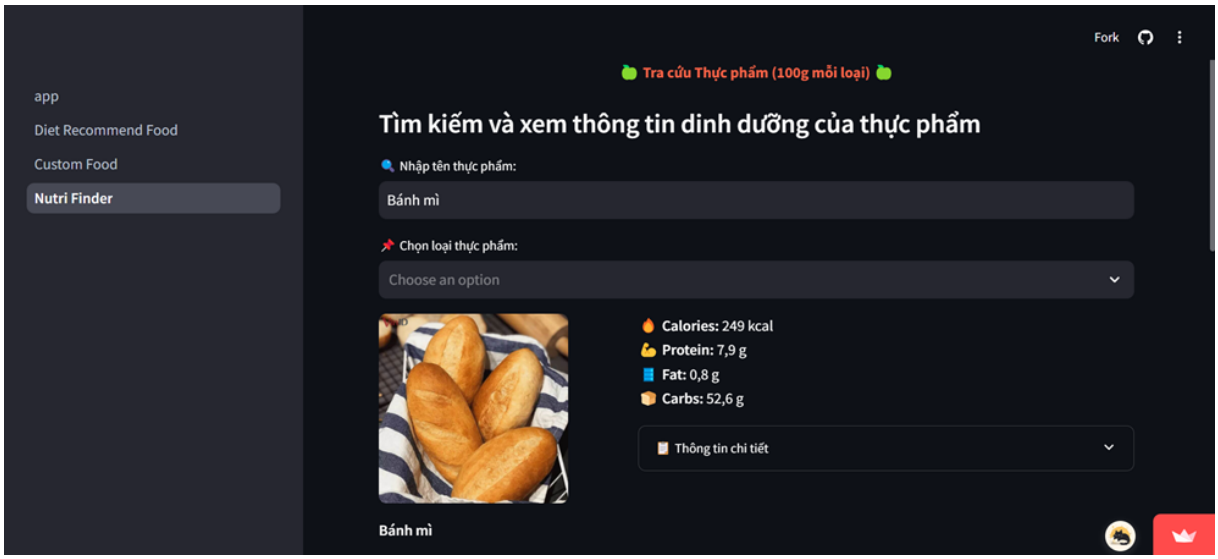
Giao diện thiết kế bữa ăn. người dùng nhập thủ công thông tin món ăn. Và hàm lượng các chất có trong thực phẩm. Chọn thêm thực phẩm. Ở đây hệ thống cho phép người dùng nhập nhiều thực phẩm. Hệ thống tổng hợp đưa ra kết quả thống kê dinh dưỡng trong ngày của của người dùng.



Hình 4.1.2.5: Giao diện Thiết kế bữa ăn.

Giao diện tra cứu thực phẩm

Giao diện tra cứu đơn giản, thân thiện với người dùng. Giúp người dễ dạng tìm thành phần dinh dưỡng của thực phâm.



Hình 4.1.2.6: Giao diện tra cứu thực phẩm

4.1.3 Thảo luận, so sánh kết quả nghiên cứu

Dưới đây là bảng kết quả so sánh của ba phương pháp phân loại được chọn có kết quả tốt nhất:

Thuật toán	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree (Max Depth = 2)	99	99	99	99
K-Nearest Neighbors (K=3)	88	86	82	83
Naive Bayes	85	71	90	76

Bảng 4.6: So sánh hiệu suất các thuật toán phân lớp

Nhận xét và phân tích:

- Decision Tree: Mô hình này đạt hiệu suất cao nhất với tất cả mô hình, với các chỉ số như độ chính xác (Accuracy), Precision, Recall và F1 Score đều đạt xấp xỉ 99%, do đó mô hình này có khả năng phân loại tốt và ổn định. Tuy nhiên, nó có thể dễ dàng bị overfitting nếu không điều chỉnh độ sâu của cây một cách hợp lý, đặc biệt khi dữ liệu có sự phức tạp cao

- K-Nearest Neighbors (KNN): KNN đạt độ chính xác 88% và các chỉ số Precision, Recall và F1 Score đều thấp hơn đáng kể so với Decision Tree. Mặc dù dễ sử dụng và không yêu cầu huấn luyện dữ liệu trước, nhưng KNN trở nên chậm và đòi hỏi tính toán phức tạp khi làm việc với dữ liệu lớn, điều này có thể làm giảm hiệu quả trong môi trường thực tế.
- Naive Bayes: Naive Bayes có độ chính xác thấp nhất với chỉ số Precision khá thấp (71%), mặc dù Recall đạt 90%, cho thấy khả năng nhận diện đối tượng tốt. Tuy nhiên, F1 Score chỉ đạt 76%, cho thấy sự không cân bằng giữa Precision và Recall. Naive Bayes giả định các đặc trưng là độc lập, điều này có thể không phù hợp với các bài toán có sự phụ thuộc giữa các đặc trưng.

Từ kết quả và nhận xét trên nhóm đã đưa ra các ưu và nhược điểm của ba phương pháp phân loại trên.

Thuật toán	**Accuracy (%)**	**Ưu điểm**	**Nhược điểm**
Decision Tree	99	Dễ hiểu, hiệu suất cao	Dễ bị overfitting, cần điều chỉnh độ sâu của cây phù hợp.
K-Nearest Neighbors	88	Dễ sử dụng, không cần huấn luyện dữ liệu trước.	Chậm với dữ liệu lớn, tính toán phức tạp.
Naive Bayes	85	Giả định độc lập giữa các đặc trưng	Không phù hợp với các bài toán có sự phụ thuộc giữa các đặc trưng.

Bảng 4.7: So sánh thuật toán phân lớp Machine Learning

Chương 5

Kết luận và kiến nghị

5.1 Kết luận tổng quan về nghiên cứu, kết quả, quá trình nghiên cứu.

5.1.1 Tổng quan về nghiên cứu

Nghiên cứu tập trung vào việc xây dựng một hệ thống hỗ trợ người dùng theo dõi thực phẩm tiêu thụ hằng ngày, dựa trên dữ liệu thực phẩm Việt Nam. Hệ thống được thiết kế nhằm giúp người dùng lựa chọn bữa ăn phù hợp với thể trạng cá nhân, đồng thời hỗ trợ thiết kế bữa ăn theo nhu cầu, giúp họ xây dựng chế độ dinh dưỡng hợp lý. Ngoài ra, hệ thống còn cung cấp chức năng tra cứu thông tin chi tiết về giá trị dinh dưỡng của thực phẩm, giúp người dùng hiểu rõ hơn về thành phần dinh dưỡng trong mỗi bữa ăn.

5.1.2 Kết quả nghiên cứu

Hệ thống theo dõi thực phẩm ăn hằng ngày đã được xây dựng và triển khai với ba tính năng chính gồm:

- Gợi ý bữa ăn cá nhân hóa: Hệ thống sử dụng thuật toán K-Means để phân nhóm theo mức độ dinh dưỡng của thực phẩm, để đưa ra đề xuất ba bữa ăn chính hợp lý cho người dùng, dựa trên các chỉ số và mục tiêu dinh dưỡng cá nhân.
- Thiết kế bữa ăn theo nhu cầu: Hệ thống cho phép người dùng tự tạo thực đơn dựa trên sở thích và nhu cầu dinh dưỡng, vận dụng thuật toán Decision Tree để phân loại dinh dưỡng giúp người dùng lựa chọn bữa ăn cân đối và phù hợp với mục tiêu sức khỏe.
- Tra cứu thông tin thực phẩm: Cung cấp giá trị dinh dưỡng chi tiết về 533 loại thực phẩm Việt Nam, giúp người dùng dễ dàng kiểm soát lượng calo và thành phần dinh dưỡng.

5.1.3 Quá trình nghiên cứu

Quá trình nghiên cứu được thực hiện như sau:

- Thu thập và tiền xử lý dữ liệu: Nhóm đã thu thập gồm 533 loại thực phẩm Việt Nam và tiền xử lý, phân tích dữ liệu để đảm bảo tính chính xác và phù hợp với mô hình.
- Lựa chọn và áp dụng thuật toán: Nhóm đã thực hiện bốn thuật toán phổ biến, trong đó K-Means được sử dụng để phân nhóm thực phẩm theo mức độ dinh dưỡng, và các thuật toán như KNN, Naive Bayes, Decision Tree được áp dụng để phân loại thực phẩm, đánh giá mức độ phù hợp nhằm hỗ trợ thiết kế bữa ăn cá nhân hóa.
- Đánh giá và cải thiện mô hình: Nhóm đã kiểm tra các thuật toán, đánh giá hiệu suất để lựa chọn mô hình tối ưu. Kết quả cho thấy Decision Tree mang lại độ chính xác cao nhất trong việc phân loại thực phẩm, trong khi K-Means giúp nhóm thực phẩm hiệu quả.
- Xây dựng và triển khai hệ thống: Hệ thống được phát triển bằng Streamlit, giúp người dùng dễ tương tác.

5.2 Các đánh giá, kiến nghị, đề xuất, hàm ý quản trị để giải quyết vấn đề nghiên cứu.

5.2.1 Đánh giá

Nhìn chung, hệ thống đã đạt được một số mục tiêu đã đề ra trong việc hỗ trợ theo dõi thực phẩm tiêu thụ hằng ngày, tuy nhiên vẫn còn một số vấn đề cần khắc phục như:

- Tập dữ liệu giới hạn: Hệ thống mới chỉ sử dụng dữ liệu của 533 loại thực phẩm, chưa bao quát đầy đủ tất cả các loại thực phẩm phổ biến tại Việt Nam.
- Mô hình chưa cá nhân hóa hoàn toàn: Hệ thống chưa tích hợp các yếu tố như bệnh lý nền hoặc sở thích ăn uống cụ thể của từng cá nhân vào quá trình gợi ý bữa ăn.
- Thiếu chức năng quản lý thông tin người dùng: Hiện tại, hệ thống chưa triển khai cơ sở dữ liệu để lưu trữ thông tin người dùng, dẫn đến việc chưa thể theo dõi lịch sử ăn uống hoặc phân tích dữ liệu theo thời gian, do đó mà hệ thống chưa hỗ trợ tính năng tổng hợp dữ liệu theo ngày, tuần và nhắc nhở người dùng về mức calo tiêu thụ.
- Hạn chế về nền tảng Streamlit: Mặc dù Streamlit giúp triển khai hệ thống một cách nhanh chống nhưng vẫn còn đơn giản, chưa hỗ trợ các tính năng nâng cao.

5.2.2 Kiến nghị

- Mở rộng tập dữ liệu: Thu thập thêm dữ liệu về thực phẩm Việt Nam để cải thiện khả năng nhận diện và gợi ý bữa ăn.
- Cá nhân hóa mô hình: Tích hợp thêm thông tin về bệnh lý nền, khẩu vị, và chế độ ăn của từng cá nhân để tăng tính cá nhân hóa.
- Xây dựng cơ sở dữ liệu người dùng: Cho phép lưu trữ thông tin cá nhân, lịch sử ăn uống và cung cấp báo cáo dinh dưỡng theo ngày, tuần.
- Nâng cấp nền tảng công nghệ: Chuyển sang một framework mạnh hơn như Flask hoặc Spring Boot để hỗ trợ thêm các tính năng nâng cao như chatbot dinh dưỡng hoặc API gợi ý thực đơn tự động.

5.2.3 Đề xuất

- Hợp tác với chuyên gia dinh dưỡng: Nhằm xây dựng mô hình dinh dưỡng thực tế và dễ dàng tiếp cận người dùng hơn dựa trên sự tư vấn từ các chuyên gia.
- Cải thiện hơn về mô hình: Thử nghiệm và áp dụng các mô hình tiên tiến như Random Forest, Nearest Neighbor hoặc mô hình mạng nơ-ron sâu (Deep Learning) để nâng cao độ chính xác và tối ưu hóa quá trình đề xuất thực đơn.
- Phát triển thành phiên bản ứng dụng di động: Giúp người dùng tiện theo dõi và ghi nhận dữ liệu hằng ngày.

5.2.4 Hàm ý quản trị

- Đối với người dùng: Hỗ trợ người dùng nâng cao nhận thức về chế độ dinh dưỡng một cách khoa học.
- Đối với nhà phát triển: Liên tục cập nhập dữ liệu, cải thiện thuật toán để mở rộng tính năng tối ưu hóa trải nghiệm người dùng.
- Đối với các tổ chức y tế: Có thể sử dụng hệ thống như công cụ hỡ trợ trong việc tư vấn dinh dưỡng cho bệnh nhân.

5.3 Hạn chế của nghiên cứu

- Thuật toán K-Means chưa phải là lựa chọn tối ưu để phân cụm vì thuật toán này yêu cầu chỉ định số lượng cụm (k) trước khi phân tích, điều này có thể dẫn đến kết quả không chính xác nếu số lượng cụm được chọn không phù hợp với phân bố thực tế của dữ liệu. Ngoài ra, K-Means chỉ hoạt động tốt với dữ liệu có hình dạng cụm đều và có thể bị ảnh hưởng mạnh bởi các điểm ngoại lai (outliers) điều này có thể làm giảm chất lượng phân nhóm.

- Thuật toán Decision Tree dù có độ chính xác cao trong ba mô hình phân loại tuy nhiên vẫn dễ bị overfitting, và có thể không tận dụng hết được mối quan hệ phức tạp giữa các đặc trưng dinh dưỡng dẫn đến việc phân loại thực phẩm chưa được tối ưu, cần thử nghiệm trên các mô hình mạnh hơn như Random Forest, XGBoost, hoặc Neural Networks, để cải thiện hơn về khả năng dự đoán.
- Chưa tối ưu cho dữ liệu lớn: Dù hiện tại hệ thống chỉ sử dụng 533 loại thực phẩm, nhưng khi mở rộng với một lượng dữ liệu lớn hơn, các thuật toán như K-Means và Decision Tree có thể gặp phải vấn đề về hiệu suất và thời gian tính toán. Điều này có thể làm giảm trải nghiệm người dùng khi hệ thống phải xử lý số lượng thực phẩm lớn hoặc khi có nhiều yêu cầu đồng thời.

5.4 Đề xuất cho hướng nghiên cứu trong tương lai

- Ứng dụng nhận diện thực phẩm bằng hình ảnh: Tích hợp các công nghệ nhận diện hình ảnh như mạng nơ-ron tích chập (CNN) để người dùng có thể dễ dàng chụp ảnh thực phẩm và hệ thống sẽ tự động nhận diện và cung cấp thông tin dinh dưỡng tương ứng cho người dùng.
- Tích hợp trên thiết bị thông minh: Kết nối hệ thống với các thiết bị thông minh như đồng hồ thông minh để theo dõi tình trạng sức khỏe của người dùng. Dựa trên dữ liệu thu thập từ các thiết bị này, hệ thống có thể đưa ra các gợi ý bữa ăn phù hợp, giúp tối ưu hóa chế độ dinh dưỡng và đáp ứng nhu cầu sức khỏe cá nhân.

Tài liệu tham khảo

[1] Ananthajothi, K., Suganthi, M., Sujitha, P., & Visalatchi, R. (2023). "Diet Recommendation System Using ML". *Conference on Recent Trends in Data Science and its Applications*, 490-495. doi: rp-9788770040723.097

[2] Phanich, M., Pholkul, P., & Phimoltares, S. (2010). "Food Recommendation System Using Clustering Analysis for Diabetic Patients". *2010 International Conference on Information Science and Applications*. doi:10.1109/icisa.2010.5480416

[3] Maria Ulfa, Winny Setyonugroho, Tri Lestari, Esti Widiasih, and Anh Nguyen Quoc. (2022), "Nutrition-Related Mobile Application for Daily Dietary Self-Monitoring". *Journal of Nutrition and Metabolism*. doi:10.1155/2022/2476367.

[4] G. Bianchetti, A. Abeltino, C. Serantoni et al. (Apr. 2022), “Personalizedself-monitoring of energy balance through integration in aweb-application of dietary, anthropometric, and physicalactivity data,”. *Journal of Personalized Medicine*. vol. 12, no. 4,p. 568.

[5] Kumari, A. P. M. D. N. N., Satya, T. P., Manikanta, B., Chandana, A. P., & Aditya, Y. L. S. (2024). "Personalized diet recommendation system using machine learning. International". *Journal of Engineering Research & Technology*, 13(02). `https://www.ijert.org/personalized-diet-recommendation-system-using-machine-learning`

[6] Duong T. T. Van, Laura Trijsburg, Ha T. P. Do, Kayo Kurotani, Edith J. M. Feskens & Elise F. Talsma1 (2022). "Development of the Vietnamese Healthy Eating Index". *ournal of Nutritional Science*. vol. 11, e45, page 1 of 10

[7] IBM. (n.d.). "Machine learning". *IBM*. `https://www.ibm.com/think/topics/machine-learning`

[8] Zydney, J. M., Hai-Jew, S., Renninger, K. A., List, A., Hardy, I., Koerber, S., … Farrell, J. M. (2012). "Supervised Learning". *Encyclopedia of the Sciences of Learning*, 3243–3245. doi:10.1007/978-1-4419-1428-6_451 .

[9] IBM. (n.d.). "Reinforcement Learning". *IBM*. `https://www.ibm.com/think/topics/reinforcement-learning`.

[10] Li, Y., & Wu, H. (2012). "A clustering method based on K-means algorithm". *Physics Procedia*, 25, 1104–1109. doi: 10.1016/j.phpro.2012.03.206.

[11] Jijo, B. T., & Abdulazeez, A. M. (2021). "Classification Based on Decision Tree Algorithm for Machine Learning". *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, doi: 10.38094/jastt20165.

[12] Yang, L., Feng, L., Tian, L., & Dai, H. (2020). "Who Will Come: Predicting Freshman Registration Based on Decision Tree". *Computers, Materials & Continua*.Vol 65, Issue 2, p1825. doi: 10.32604/cmc.2020.010011.

[13] Lakshminarayanan, B.(2016). "Decision Trees and Forests: A Probabilistic Perspective". *Doctoral thesis , UCL (University College London).*

[14] Scikit-learn. (n.d.). "Naive Bayes". *Scikit-learn*. `https://scikit-learn.org/stable/modules/naive_bayes.html`

[15] Kazmierska, J., & Malicki, J. (2008). "Application of the Naïve Bayesian Classifier to optimize treatment decisions". *Radiotherapy and Oncology*, 86(2), 211–216. doi:10.1016/j.radonc.2007.10.019.

[16] WPRO/IDI (2000). The Asia-Pacific perspective: redefining obesity and its

[17] WHO expert consultation (2004). Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Public Health. 363. 9403. 157-163.

[18] Viện Dinh dưỡng TP.HCM. (2021). Theo dõi tình trạng dinh dưỡng bằng chỉ số BMI. Viện Dinh dưỡng TP.HCM. https://viendinhduongtphcm.org/vi/dinh-duong-co-ban/theo-doi-tinh-trang-dinh-duong-bang-chi-so-bmi.html.

[19] Streamlit. https://streamlit.io/docs/create-an-app

[20] Nguyễn, T. M., & Nguyễn, L. H. P. (2025). "Bộ dữ liệu thành phần dinh dưỡng". https://docs.google.com/spreadsheets/d/1mjMrweeCQVRrOiqQRkJGe66DWbzTKsJh/edit?usp=sharing&ouid=107947283027204754630&rtpof=true&sd=true

[21] Bộ Y tế Viện Dinh Dưỡng. (2007). "BẢNG THÀNH PHẦN THỰC PHẨM VIỆT NAM". *Nhà xuất bản y học*. https://www.fao.org/fileadmin/templates/food_composition/documents/pdf/VTN_FCT_2007.pdf.

[22] Lê, T. Đ. L. Lê, T. H. N. (.n.d.) "Thành phần dinh dưỡng một số thức ăn nhanh ". *Viện Dinh Dưỡng TP. Hồ Chí Minh*. https://viendinhduongtphcm.org/Media/Tai_lieu_chuyen_mon/tpdd_thucannhanh.pdf.

[23] TIÊU CHUẨN XÂY DỰNG THỰC ĐƠN VỀ DINH DƯỠNG ĐỐI VỚI BỮA ĂN HỌC ĐƯỜNG: https://syt.bacgiang.gov.vn/dinh-duong-hoc-uong/-/asset_publisher/6CWBO9WiZqsQ/content/tieu-chuan-xay-dung-thuc-on-ve-dinh-duong-oi-voi-bua-an-hoc-uong?inheritRedirect=false&utm_source=chatgpt.com

[24] Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). "Improved naive Bayes classification algorithm for traffic risk management". *EURASIP Journal on Advances in Signal Processing*, 2021(30).