UIT
Trường Đại học Công nghệ Thông tin

Khoa Khoa học và Kỹ thuật Thông tin

BÁO CÁO ĐỒ ÁN

XÂY DỰNG MÔ HÌNH ĐỀ XUẤT MÓN ĂN LÀNH MẠNH

MÔN HỌC: TIỀN XỬ LÝ VÀ XÂY DỰNG BỘ DỮ LIỆU - DS108.O21

GVHD: TS. Nguyễn Gia Tuấn Anh
GV HDTH: Trần Quốc Khánh

Nhóm sinh viên thực hiện: Nhóm 10
22520224 - Nguyễn Thành Đạt
22520131 - Lê Xuân Bình
22520269 - Nguyễn Duy Đức
22521526 - Phạm Ngọc Trí

MỤC LỤC

1. GIỚI THIỆU

TỔNG QUAN

Đề xuất các món ăn lành mạnh dựa trên **các chỉ số dinh dưỡng** và **nguyên liệu**

- **Input:** Nguyên liệu mong muốn, các chỉ số dinh dưỡng
- **Output:** Các công thức món ăn lành mạnh thoả mãn yêu cầu người dùng

PHẠM VI NGHIÊN CỨU

Các công thức món ăn lành mạnh đã được định lượng hóa

MỤC TIÊU ĐỀ TÀI

Xây dựng mô hình đề xuất món ăn

ỨNG DỤNG ĐỀ TÀI

Hệ thống gợi ý chế độ ăn dinh dưỡng

1. GIỚI THIỆU

ĐỘNG LỰC CỦA NGHIÊN CỨU

Các nghiên cứu trước đó:

- A **Collaborative Filtering Approach** for Food Recommendation (T. Nguyen, P. Tran, and M. Le)
- **Deep Learning** for Nutritional Analysis and Healthy Food Recommendation (J. Smith, A. Johnson, and R. Brown)
- **Big Data-based** Food Recommendation System using Social Media and Cooking Websites (S. Kim, H. Lee, and Y. Cho)
- Personalized Food Recommendation System based on **Health Profiles** (Y. Zhang, L. Wang, and D. Li)

=> Content-based recommend system using machine learning on small data

2. THU THẬP DỮ LIỆU

Công cụ thu thập dữ liệu:

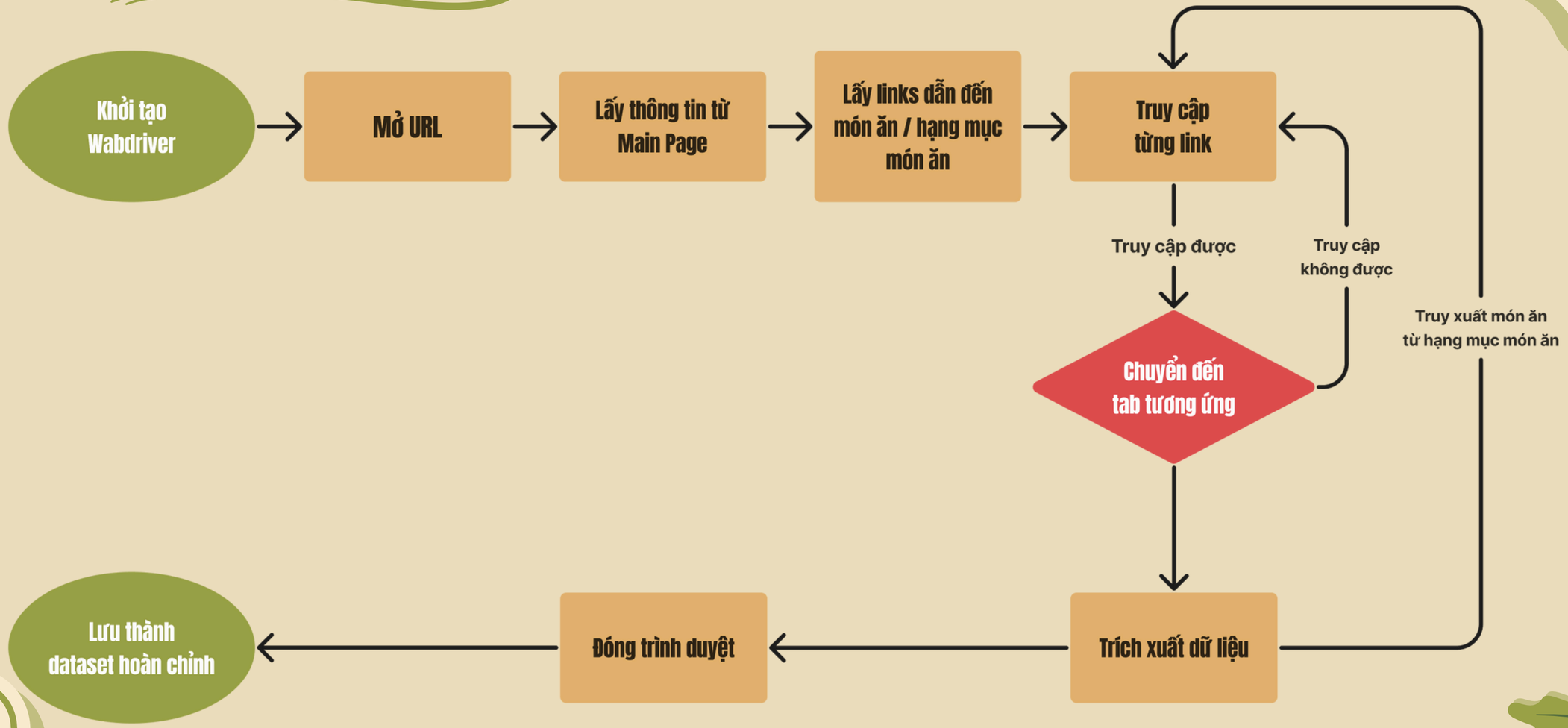
- Selenium

Trang web crawl dữ liệu:

- Eating Well
- Food

Food: Một website với hơn 500 nghìn công thức.
EatingWell: Năm 2021 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.

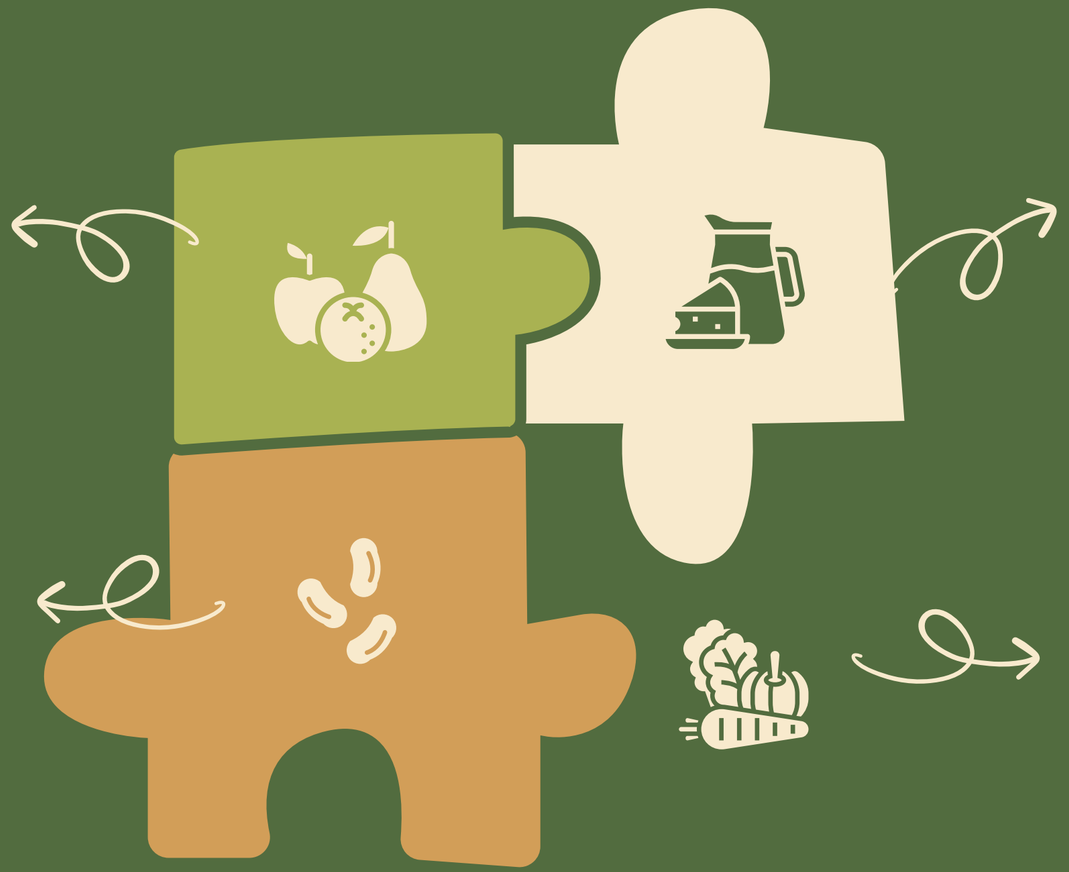
2. THU THẬP DỮ LIỆU



3. TIỀN XỬ LÝ DỮ LIỆU

TỔNG QUAN VỀ DỮ LIỆU THU THẬP

EDA VÀ PREPROCESSING CƠ BẢN



KẾT QUẢ THU ĐƯỢC

KIỂM TRA CHẤT LƯỢNG DỮ LIỆU DỰA TRÊN 5 TIÊU CHÍ VÀ ÁP DỤNG GREAT EXPECTATIONS

TỔNG QUAN VỀ DỮ LIỆU

FOOD

Dataset có những thuộc tính sau:

- Title, Link, Description
- Image_URL, Image_Path,
- Time, Author,
- Ingredient, Direction,
- Calories, Calories From Fat, Total Fat, Saturated Fat, Cholesterol, Sodium, Total Carbohydrate, Dietary Fiber, Sugars, Protein
- Diet_label

Tổng kết quả thu được là 466 dòng và 14 cột

Food.

TỔNG QUAN VỀ DỮ LIỆU

EATING WELL

Dataset có những thuộc tính sau:

- Name
- Description
- Time
- Ingredients
- Directions
- Nutrition

Tổng kết quả thu được là 2604 dòng và 7 cột

TỔNG QUAN VỀ DỮ LIỆU

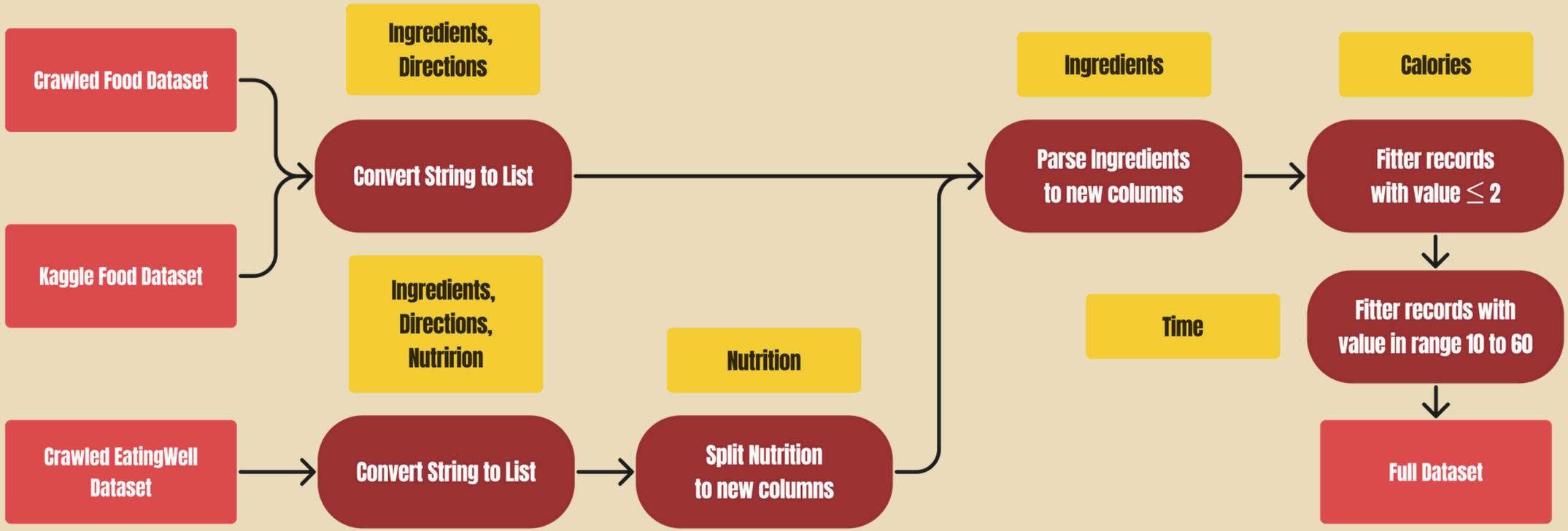
KAGGLE FOOD

Dataset có những thuộc tính sau:

- RecipeID, Name, AuthorID, AuthorName,
- PreptTime, CookTime, TotalTime,
- DatePublish, Images, RecipeCategory, Keywords
- RecipeIngredientQuantities, RecipeIngredientParts,
- AggregatedRating, ReviewCount,
- Calories, FatContent, SaturatedFatContent, CholesterolContent, SodiumContent, CarbohydrateContent, FiberContent, SugarContent, ProteinContent, RecipeServings, RecipeYield,
- RecipeInstructions

Tổng kết quả thu được là 14800 dòng và 27 cột

QUY TRÌNH TIỀN XỬ LÝ DỮ LIỆU



EDA VÀ PREPROCESSING

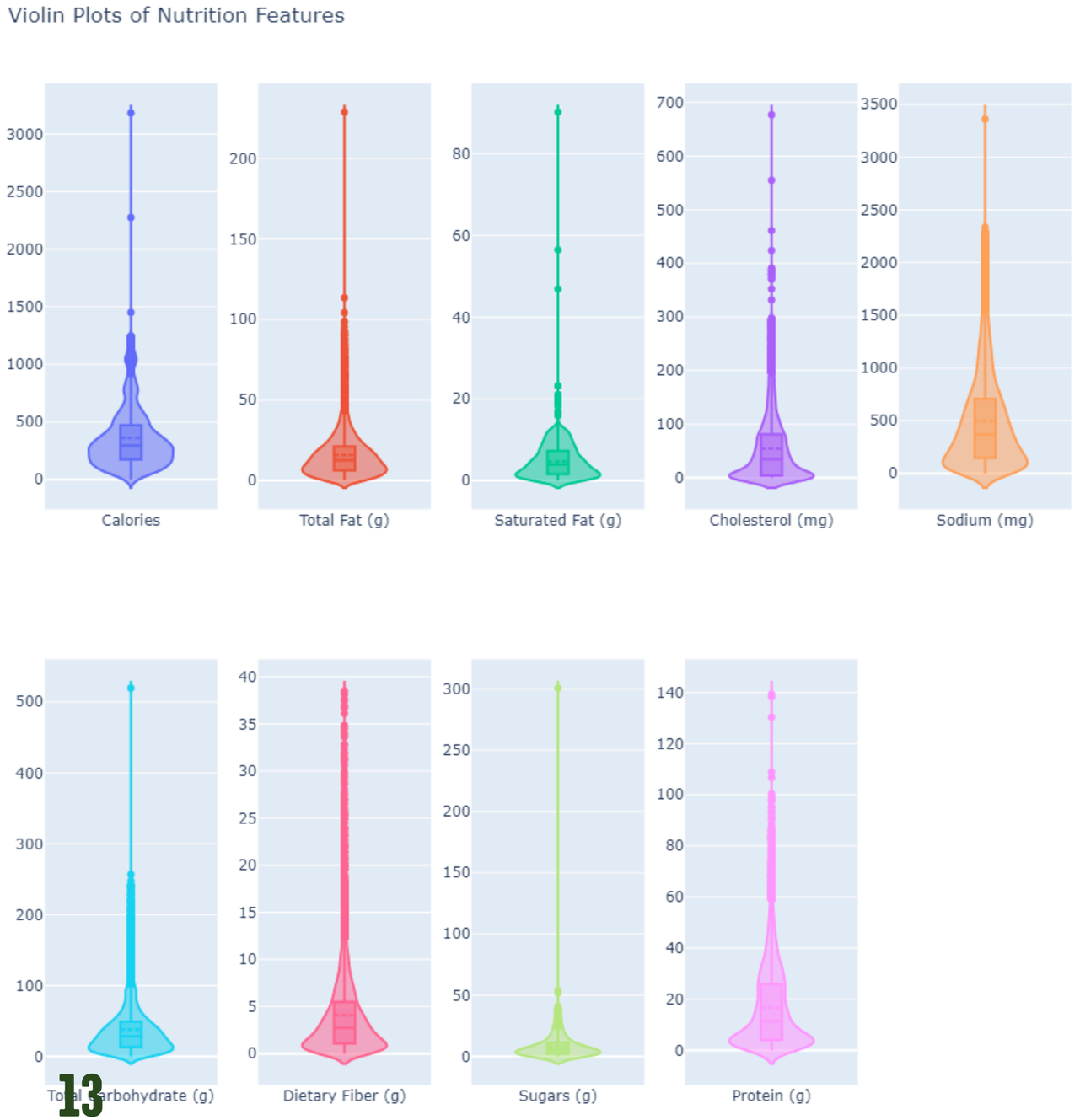
Từ các bước dữ liệu thu thập được sau đó tiến hành gộp lại thành 1 file dataset chung

Các thuộc tính có được sau cùng là:

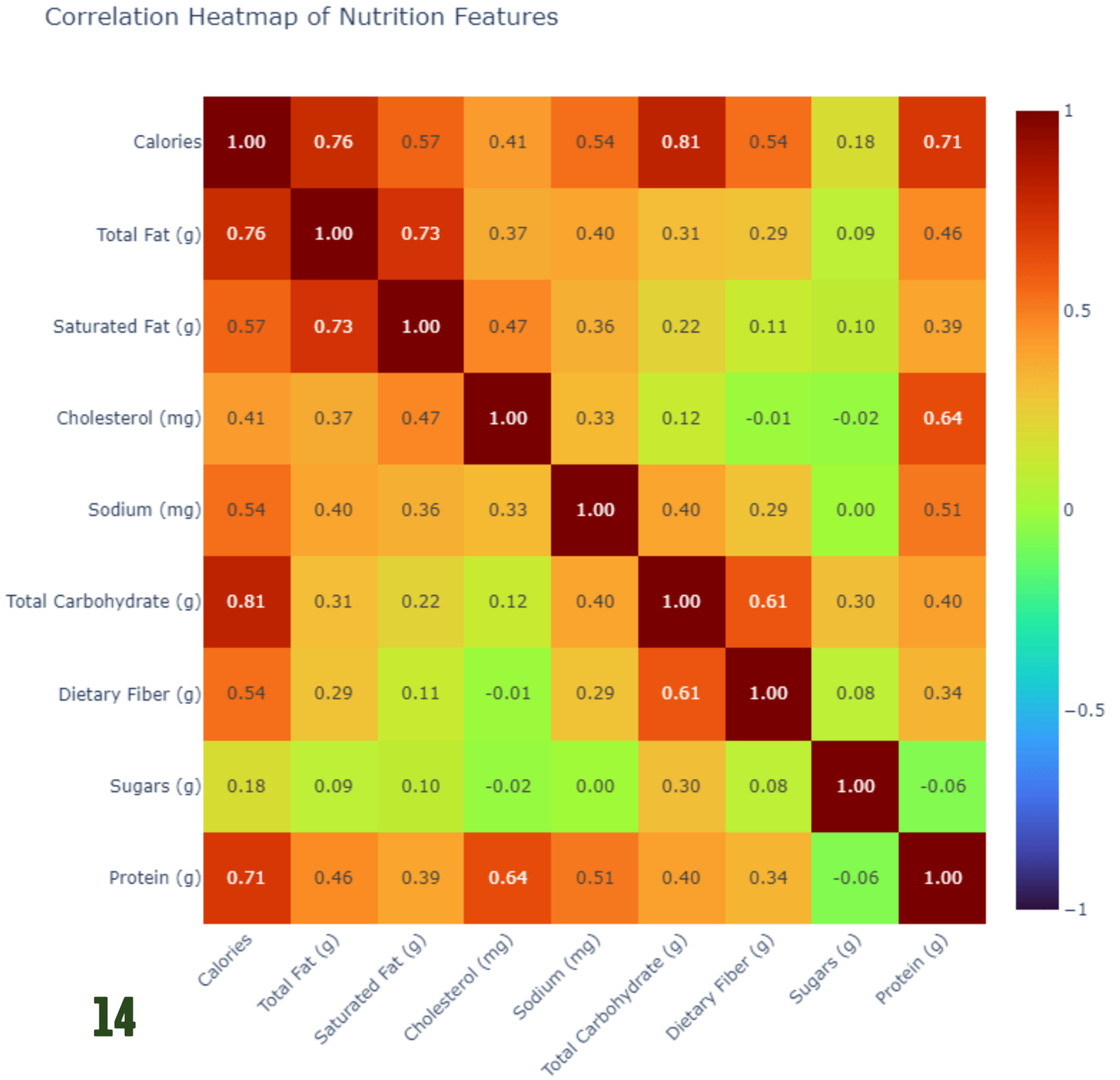
- Title, Time (mins), Direction, Ingredient,
- Calories, Total Fat (g), Saturated Fat (g), Cholesterol (mg), Sodium (mg), Total Cacbohydrate (g), Dietary Fiber (g), Sugars (g), Protein (g)
- Diet Label, Ingredient_units

Tổng kết quả thu được là 10269 dòng và 16 cột

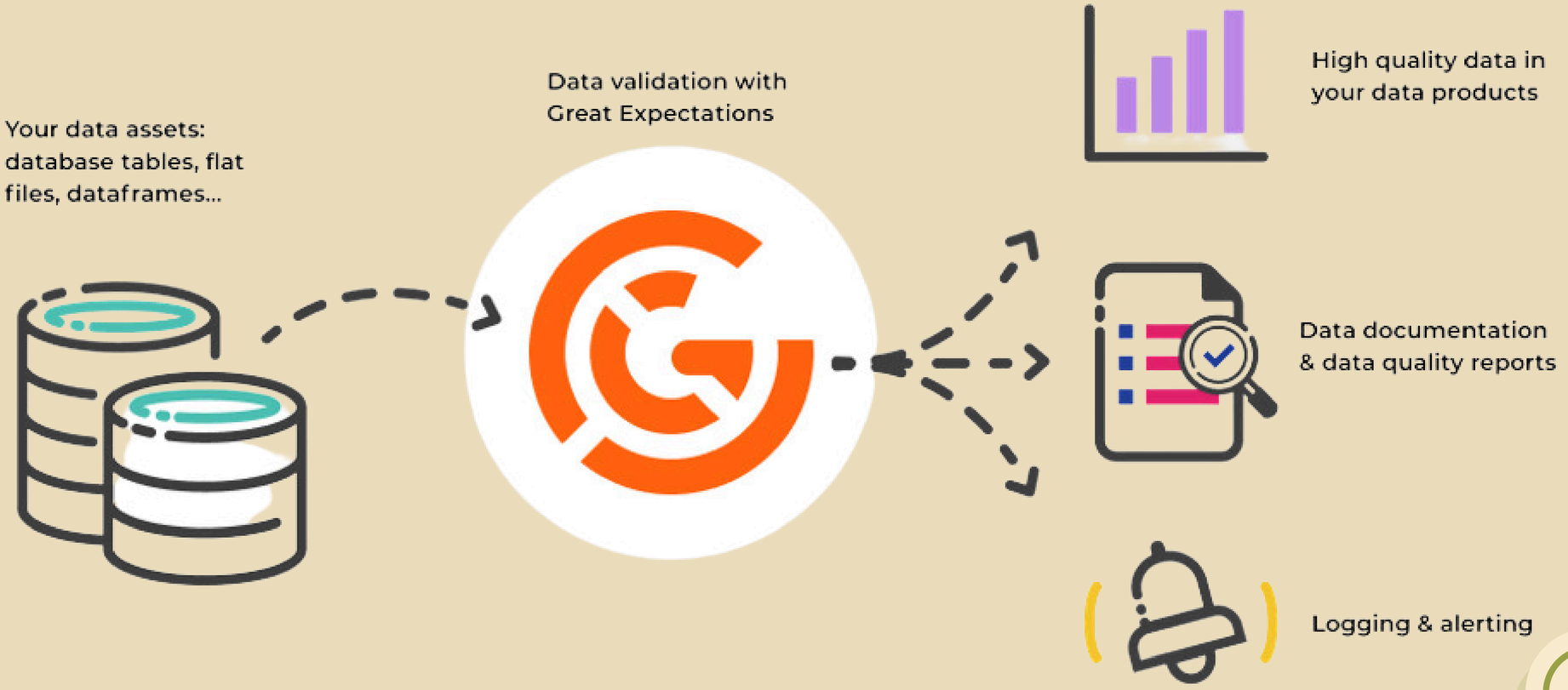
PHÂN TÍCH BỘ DỮ LIỆU



PHÂN TÍCH BỘ DỮ LIỆU



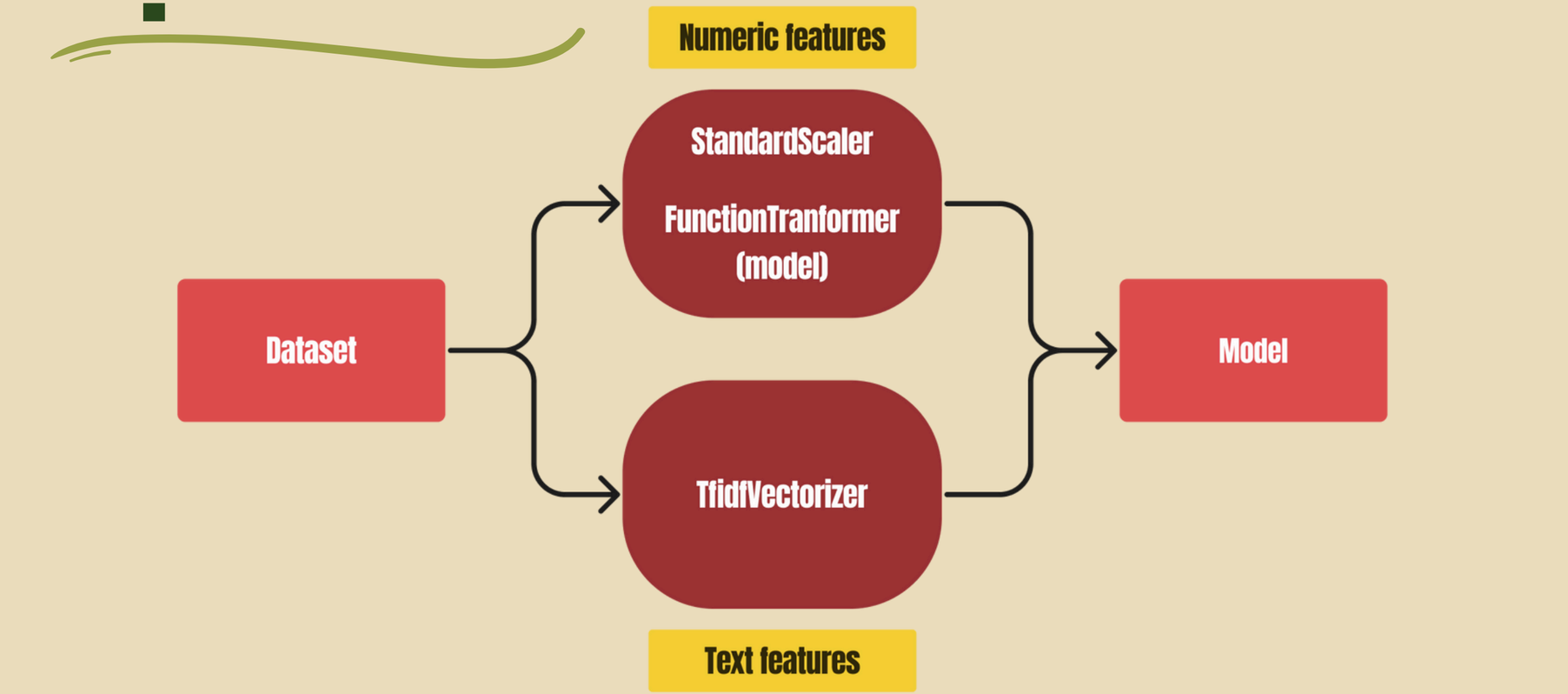
ĐÁNH GIÁ DATA QUALITY DỰA TRÊN 5 TIÊU CHÍ

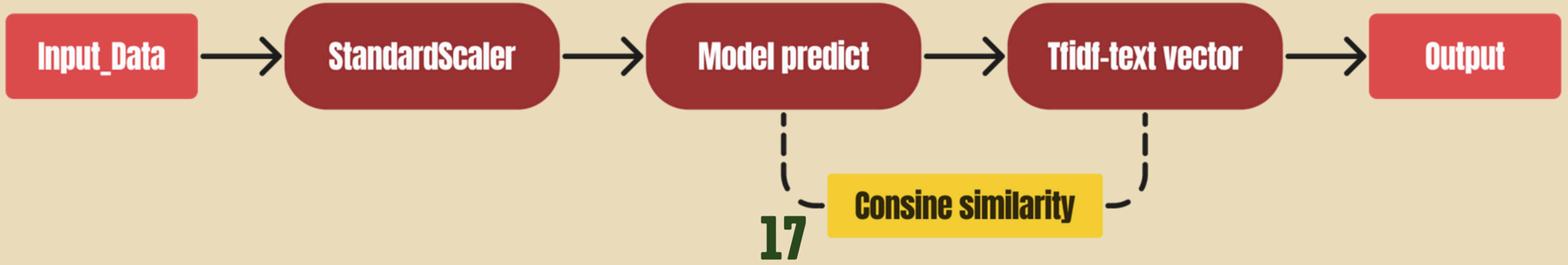


Đánh giá cho dataset của nhóm

- Accuracy:
  - các record có duplicate ?
  - các cột có giá trị null ?
- Completeness: có đầy đủ các feature mong muốn hay không ?
- Consistency: tính nhất quán về kiểu (int, float,...) dữ liệu ?
- Relevance: các chỉ số có nằm trong khoảng cụ thể nào không ?

4. XÂY DỰNG MÔ HÌNH





5. ĐÁNH GIÁ

Test Input:

Recipes ID 1

Output:

MODEL	Recipes ID	Numeric cosine similarity	Text cosine similarity
KNN	699	0.974178	0.13370248
GMM	Emty result	N/A	N/A
Radius Neighbors Classifier	6074 6733 3480 321 2031	0.924459 0.906683 0.894694 0.862496 0.861416	0.188852 0.000000 0.151270 0.166101 0.000000
K Means	27 51 57 60 76	N/A	N/A

6. TRIỂN KHAI ỨNG DỤNG

WEB APP

Công cụ xây dựng web app:

- Streamlit

Công cụ tạo tunnel giữa localhost và internet:

- Ngrok

Công cụ xây dựng API hiệu nâng cao hỗ trợ xây dựng web app:

- FastAPI

Streamlit

ngrok

FastAPI

7. KHÓ KHĂN

Chưa tổ chức được code có thể crawl tự động

8. MỤC TIÊU TRONG TƯƠNG LAI

Tự động hóa quy trình crawl và kiểm tra dữ liệu sử dụng Airflow

THANK YOU

AND

Q&A