Department of Electrical and Computer Engineering
North South University (NSU)

CSE 445: Machine Learning
Section 6

Project

Instructor: Dr. Mohammad Mahmudul Alam Semester: Fall 2025

Title:	CraveBot: A RAG Chatbot for Calorie Deficit Meal Guidance
Name:	Arafat Zaman Ratul
NSU ID:	2311539042

Project Update

Title: CraveBot: A RAG Chatbot for Calorie Deficit Meal Guidance

Abstract:
CraveBot is a nutrition-focused conversational chatbot that utilizes a Retrieval-Augmented Generation (RAG) architecture. It aims to help users receive personalized calorie-deficient meal suggestions and nutrition insights through data-driven responses. The project leverages the OpenFoodFacts dataset (4.09 million entries), which provides extensive nutritional information for food products.
The primary objectives are to develop a clean, structured nutrition knowledge base, train a supervised classifier to identify calorie-wise food items, and prepare a retrieval and generation pipeline that integrates a semantic retriever with an LLM.
To achieve this, I utilized Python, Pandas, scikit-learn, with plans to integrate Gradio SentenceTransformers, FAISS, and GPT-2 for a full RAG implementation. The ultimate goal is to develop a functional nutrition assistant capable of accurately and interactively answering food and calorie-related queries.

Methodology:

- **Data Preparation:**
  - Extracted and flattened OpenFoodFacts entries using a custom Python function to capture product_name, energy_kcal_100g, fat_100g, carbohydrates_100g, sugars_100g, and proteins_100g.
  - Cleaned missing data and saved the dataset as both CSV and Parquet for efficiency.
- **Feature Engineering & Modeling:**
  - Created a binary classification target: low_calorie = energy_kcal_100g < 100.
  - Selected three main predictors (fat, carbohydrates, proteins).
  - Trained a RandomForestClassifier baseline, followed by a compact version using two features (fat, carbohydrates).
  - Evaluated models using accuracy, precision, recall, F1-score, and confusion matrix.
  - Implemented feature importance analysis for interpretability.
- **Tools and Environment:**
  - **Tools:** Python, scikit-learn, Pandas, Matplotlib, Gradio, SentenceTransformers (planned), FAISS (planned).
  - **Environment:** Google Colab (T4 GPU) with progress tracking, saved logs, and model persistence using joblib.

Initial Results:
Baseline Random Forest (3 Features):

- Test Accuracy: 96.19%
- Precision (Low Calorie): 0.94
- Recall (Low Calorie): 0.85
- F1-Score (Low Calorie): 0.89
- The confusion matrix indicates a strong overall performance, with moderate false negatives resulting from class imbalance.

Compact Model (2 Features):

- Test Accuracy: 93.69%
- Increased false negatives indicate that proteins provide additional predictive strength.

Insights:

- Feature importance confirmed that fat and carbohydrates dominate the classification of calories.
- The model behaves consistently with nutritional logic.
- High accuracy achieved despite the imbalance between high- and low-calorie classes.

Challenges Faced:

- Long data processing time for 4 M entries
- Missing values required pragmatic handling (fillna(0))
- Class imbalance affected recall, making accuracy alone a misleading measure.
- Next-phase RAG integration (FAISS + GPT-2) requires higher computational resources
- A significant time requirement is associated with hyperparameter tuning for the model.

Next Steps:

- Finalize and persist the trained Random Forest models using the joblib library.
- Perform hyperparameter tuning via GridSearchCV/RandomizedSearchCV or Optuna with cross-validation and PR-AUC tracking.
- Build a semantic retriever using SentenceTransformers and FAISS to enable search-based context retrieval.
- Integrate an LLM (GPT-2 or API model) to enable question answering with factual grounding and finalize with Gradio demo.