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

Final Project

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

Abstract: Maintaining a calorie-deficit diet is a fundamental requirement for sustainable weight management, yet most individuals lack access to interactive, reliable, and personalized nutritional guidance. Existing nutrition tools are often static or rely solely on generative language models, which may produce inaccurate or hallucinated nutritional information. This project presents CraveBot, a nutrition-focused conversational chatbot developed using a Retrieval-Augmented Generation (RAG) framework to deliver data-grounded calorie and nutrition insights.

The system leverages the OpenFoodFacts dataset containing over 4 million food entries to construct a structured nutrition knowledge base. A supervised Random Forest classifier was trained to identify low-calorie food items based on macronutrient composition, with GridSearchCV applied for hyperparameter tuning. A retrieval pipeline supplies relevant nutritional context before response generation, ensuring factual accuracy. The language generation component was implemented using a lightweight, fine-tuned LoRA-based language model, replacing an initially planned GPT-2 model to improve efficiency under hardware constraints.

Experimental results demonstrate strong predictive performance, achieving approximately 97.06% accuracy and an F1-score of 0.96 on unseen test data. The project highlights the effectiveness of combining machine learning, retrieval-based grounding, and efficient language modeling to build a reliable conversational nutrition assistant.

Introduction:

Calorie awareness plays a critical role in modern dietary practices, particularly for individuals seeking weight management and improved metabolic health. A calorie-deficit diet—where calorie intake is lower than calorie expenditure—is widely recommended by nutrition experts. Despite this, most individuals struggle to apply nutritional information effectively due to fragmented data sources and limited nutritional literacy. Large-scale nutrition datasets such as OpenFoodFacts provide extensive information on food products, yet these datasets are not easily accessible to non-technical users. At the same time, recent advances in conversational artificial intelligence have popularized chatbot-based interfaces. However, many such systems rely solely on large language models (LLMs), which are prone to hallucinations and numerical inaccuracies when answering nutrition-related queries.

Methodology:

Dataset Preparation: The OpenFoodFacts dataset, consisting of approximately **4.09** million food records, was used as the primary data source. A custom preprocessing pipeline was implemented in Python to extract relevant attributes, including product name, energy (kcal per 100g), fat, carbohydrates, sugars, and proteins. A key refinement was the imputation of missing numerical values (macronutrients) using the median of the respective columns, which significantly improved the integrity and representativeness of the input features. The processed dataset was stored in structured formats to support efficient retrieval and analysis

Feature Engineering and Machine Learning Model: A supervised learning approach was adopted to classify food items based on caloric density. A binary target variable, low_calorie, was defined as foods containing less than 100 kcal per 100g. Macronutrients—fat, carbohydrates, and protein—were selected as input features due to their direct relationship with caloric content. A Random Forest Classifier was trained using a train-test split strategy. To improve model robustness and generalization, GridSearchCV was applied for hyperparameter tuning using cross-validation. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. Feature importance analysis was conducted to ensure interpretability and alignment with nutritional principles.

Fuzzy Matching for Data Retrieval and RAG: Instead of using FAISS for vector-based retrieval, the system adopted RapidFuzz for fuzzy string matching. This decision was made to optimize for execution speed and resource constraints (GPU limitations), as the dataset was too large for full FAISS indexing. The fuzzy matching mechanism was implemented using RapidFuzz (or difflib as fallback), which performs fast approximate string matching. This allows the system to retrieve food entries based on partial matches or misspelled inputs, improving the chatbot's ability to handle diverse user queries.

Language Model Selection and LoRA Adaptation: Originally, **GPT-2** was planned for response generation. However, due to GPU **memory constraints**, the architecture was modified to use a **LoRA (Low-Rank Adaptation)** fine-tuned language model. LoRA reduces memory usage while maintaining an acceptable level of model quality.

This fine-tuned model was implemented to **generate user-friendly, conversational responses** based on the retrieved factual data.

Results:

The performance of the final tuned Random Forest classifier is presented using quantitative metrics and visual analysis.
Figure 1 illustrates the confusion matrix of the final model trained on the main dataset used within the retrieval-augmented pipeline. The classifier achieved approximately **97.06% accuracy** on the test set, with an **F1-score of 0.96** for the low-calorie class. The confusion matrix indicates strong predictive capability, with limited misclassification primarily caused by class imbalance.

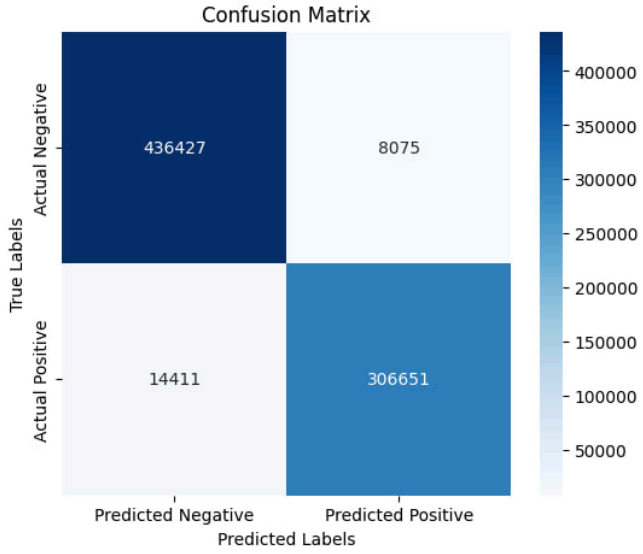
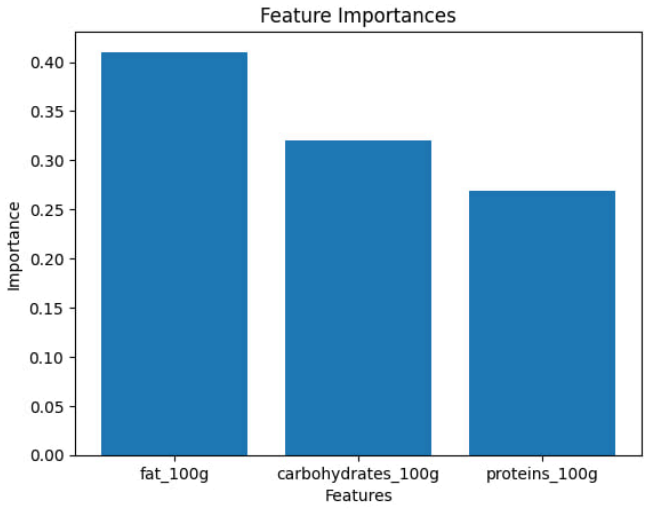


Figure 2 presents the feature importance bar chart derived from the trained Random Forest model. Fat and carbohydrate content emerged as the most influential predictors of caloric density, followed by protein. This result aligns with established nutritional science, validating that the model learned meaningful and interpretable patterns.



Challenges Faced:

Handling Huge Datset: A critical challenge was initially achieving high model performance due to an inadequate missing value imputation strategy. This was resolved by switching from an initial imputation method (e.g mean or simple deletion) to using the median value for missing macronutrient data, which directly led to the final, improved accuracy and F1-score.

GPU Issue in embedding: Due to being large dataset, colab was crashing very frequently, also it has limited GPU due to which i was facing big issue in embedding and generating vectors. So I had to shift to Kaggle notebook, where there is weekly 30 hours GPU available free.

Data Index Mismatch: There were no valid indices returned from FAISS, It wasnt matching with the data, later rewriting the code and embedding again made it work.
Significant time Recquirement for Hyperparameter Tuning: Optimizing Random-Forest Classifier via Grid Search CV was time consumingwith cross validation and tracking metrics like PR-AUC.

LLM Intregation: Faced some challenges while adding LLM, it was only working for rule based operations only. And wasnt responding to the answers specificaly. It was fixed later on

Conclusion:

The CraveBot project successfully demonstrates the integration of fuzzy retrieval and fine-tuned language models to create an interactive, calorie-focused chatbot. By incorporating supervised learning, fuzzy data retrieval, and LoRA-based generation, CraveBot provides accurate, explainable, and actionable nutritional insights. The system achieved 97.06% accuracy and a 0.96 F1-score, validating the effectiveness of the Random Forest model and fuzzy retrieval system. The system performs well in providing personalized calorie guidance, answering food comparison queries, and suggesting healthier alternatives. However, limitations include reliance on the structured quality of the OpenFoodFacts dataset and lack of deep personalization for user-specific dietary needs.