INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Culturally Adaptive NLP-Based Diet Recommendation System (**HelaFoods**)

Dataset Created By

R.M.Sandaru Chanduap

Supervised by

Mr. Viraj Lakshitha

Abstract

This paper introduces a structured dataset of Sri Lankan foods designed to support research in food recommendation systems, dietary analytics, and machine learning applications. Despite the growth of food recommender systems in recent years, there is limited availability of high-quality, culturally diverse datasets that include foods from South Asian contexts such as Sri Lanka. The dataset presented here systematically organizes food items with clear identifiers, descriptions, cuisine labels, and dietary attributes. It was developed using a controlled generation method to ensure consistency, uniqueness, and readiness for machine learning tasks. Potential applications include classification modeling, clustering analysis, natural language processing of food descriptions, and personalized recommendation engines. By releasing this dataset on Harvard Dataverse, we aim to facilitate reproducible research and to fill a gap in existing food datasets.

1. Introduction

Food recommendation systems have become integral to digital culinary platforms, nutrition applications, and personalized dietary support tools. These systems rely on structured food data to understand user preferences, dietary constraints, and cultural contexts. However, most publicly available food datasets focus on Western cuisines or lack contextual information relevant to diverse populations. Sri Lankan cuisine, with its unique ingredients, preparation styles, and cultural associations, is underrepresented in the current landscape of food data.

The absence of well-organized Sri Lankan food data restricts research on culturally sensitive recommendation models, limits the scope of dietary studies within South Asia, and hampers the development of comprehensive global food information systems. To address this gap, we present a curated dataset of Sri Lankan food items, structured for machine learning and research use.

2. Background and Related Work

Food recommender systems typically leverage annotated datasets to learn associations between users and food items. Previous research has explored various recommendation methodologies, including collaborative filtering, content-based filtering, and hybrid approaches. These systems often depend on datasets containing nutritional profiles, user ratings, ingredient lists, and cuisine categorization.

Existing datasets such as Recipe1M and other large recipe collections provide rich data for model training but are predominantly focused on Western or broadly international cuisines. Moreover, many datasets lack explicit dietary tags or cultural context, which are necessary for localized recommendation tasks. The scarcity of Sri Lankan food data in these repositories highlights the need for specialized datasets that capture regional food characteristics.

3. Dataset Description

3.1 Overview

The dataset consists of **X** unique Sri Lankan food items stored in **CSV** format. Each entry represents a distinct food item with multiple attributes designed for use in machine learning and analytical tasks.

3.2 Attributes

Column Name	**Description**
food_id	Unique identifier for each food item
name	Official name of the food
description	Short textual description
cuisine	Cuisine label (e.g., “Sri Lankan”)
is_vegetarian	Binary indicator for vegetarian suitability
dietary_tags	Additional dietary information (e.g., halal, vegan)

3.3 Format

The dataset is formatted as a standard CSV file, ensuring compatibility with common data tools and machine learning frameworks. It can be imported into Python, R, MATLAB, or SQL databases with minimal preprocessing.

4. Dataset Construction Methodology

The dataset was assembled using a controlled process to ensure quality and consistency:

- Item identification: Food items were selected based on cultural relevance and representativeness of Sri Lankan cuisine.
- Attribute definition: Columns were defined to capture essential food characteristics relevant for recommendation and classification tasks.
- Data generation: Entries were generated systematically with unique identifiers and curated descriptions.
- Validation checks: Consistency checks were applied to avoid duplicates, ensure correct dietary labels, and maintain format integrity.

Although the dataset is not compiled from direct user logs or ratings, it is suitable for research purposes, especially for algorithm development and benchmarking.

5. Potential Applications

This dataset enables a range of research and practical applications:

- Food recommendation systems: Train and evaluate AI models that suggest food items based on dietary constraints and cultural preferences.
- Classification problems: Use the dataset for supervised learning tasks such as predicting dietary suitability.
- Clustering analysis: Explore grouping of food items by similarity in description or dietary features.
- Natural language processing: Analyze food descriptions for semantic patterns, embeddings, or recommendation metadata.

6. Limitations and Future Work

The dataset is curated rather than collected from real user behavior, which may limit its direct applicability to models requiring user preference data. Additionally, some dietary attributes (e.g., nutritional values) are not included in the initial version. Future work could expand the dataset to include ingredient lists, nutritional profiles, and user rating information.

7. Ethical Considerations

This dataset is provided for research and educational purposes only. It does not include personal or sensitive data. Users should note that dietary labels (e.g., vegetarian or halal) are provided based on general definitions and may not reflect all cultural or individual interpretations.

8. Data Availability

The dataset is available on Harvard Dataverse under an open-access license (e.g., CC BY 4.0). Users can access and reuse the dataset in accordance with the terms of the chosen license.

9. Conclusion

I present a structured dataset of Sri Lankan foods with clear attributes designed to support AI-driven research and recommendation systems. By addressing a gap in culturally relevant food data, this dataset provides a foundational resource for academic and applied work. We encourage reuse, extension, and adaptation of this dataset to foster diverse food recommendation research.