DS5110 IDMP
Project Proposal
Food Categorization: A Machine Learning Approach to USDA Branded Foods

Ajin Frank Justin
justin.aj@northeastern.edu

Chirag Chivate
chivate.c@northeastern.edu

Yashi Chawla
chawla.y@northeastern.edu

1 Project Participants

Ajin Frank Justin, Chirag Chivate, Yashi Chawla

Database Design & Application Development (Parts 1 & 2)

- **Chirag:** Design a database with 10 relational tables, define data, and implement functions with a command-line interface.
- **Ajin:** Acquire and preprocess data from external datasets, and develop querying functionality with database integration.
- **Yashi:** Ensure integration and consistency across tables, preprocess data from 4 tables, and manage data insertion and I/O handling.

Data Analysis, Visualizations & Machine Learning (Parts 3 & 4)

- **Ajin:** Conduct EDA and summary statistics, train and evaluate the model, and store predictions.
- **Chirag:** Analyze correlations, perform group-by operations, create visualizations, and focus on hyperparameter tuning and metrics storage.
- **Yashi:** Handle missing data, apply transformations, lead model selection, and summarize insights.

Collaborative Efforts

All members will contribute equally to the final report and results discussion.

2 System Description

The system's objective is to provide a comprehensive platform for analyzing and predicting food product trends using the USDA Branded Foods Dataset. This system will support structured querying, data analysis, and machine learning, aimed at deriving actionable insights regarding food consumption, nutritional trends, and product categorization.

2.1 System Requirements & Main Features

The system will include the following core features and components:

- **Data Storage and Management:** A relational database to store branded food products data, nutritional data. The schema will support both querying and data integration ensuring efficient storage and quick access to large amounts of data efficiently with appropriate correlations between tables.

- **Querying and CRUD Operations:** The system will enable users to perform CRUD (Create, Read, Update, Delete) operations on the dataset. It will also support complex queries to retrieve specific information based on nutritional content, product category, or brand.

- **Exploratory Data Analysis (EDA):** Users can explore data trends and insights through statistical analysis with visualizations. The system will also calculate summary statistics, identify correlations, and use plots to visualize key insights like nutrient distributions and product trends.

- **Machine Learning Predictions:** The system will train on classification models to predict the food categories based on product features (e.g., brand name, ingredients, serving size). Each of these predictions will be evaluated and used for further analysis.

- **Visualizations:** Graphical representations, like histograms, bar charts, and heatmaps, will be created to provide visual insights into food consumption patterns, nutrient distributions, etc.

3 Database Design

3.1 Entities and Attributes

- **branded_food:** Contains details about food products, including a unique identifier `fdc_id` (Primary Key), `brand_name`, and `ingredients`.

- **food_attribute_type:** Defines attribute types with a unique identifier `id` (Primary Key), `name`, and `description`.

- **food_attribute:** Stores attributes for food items, linking to `branded_food` via `fdc_id` (Foreign Key) and to attribute types via `food_attribute_type_id` (Foreign Key), with an additional `name`.

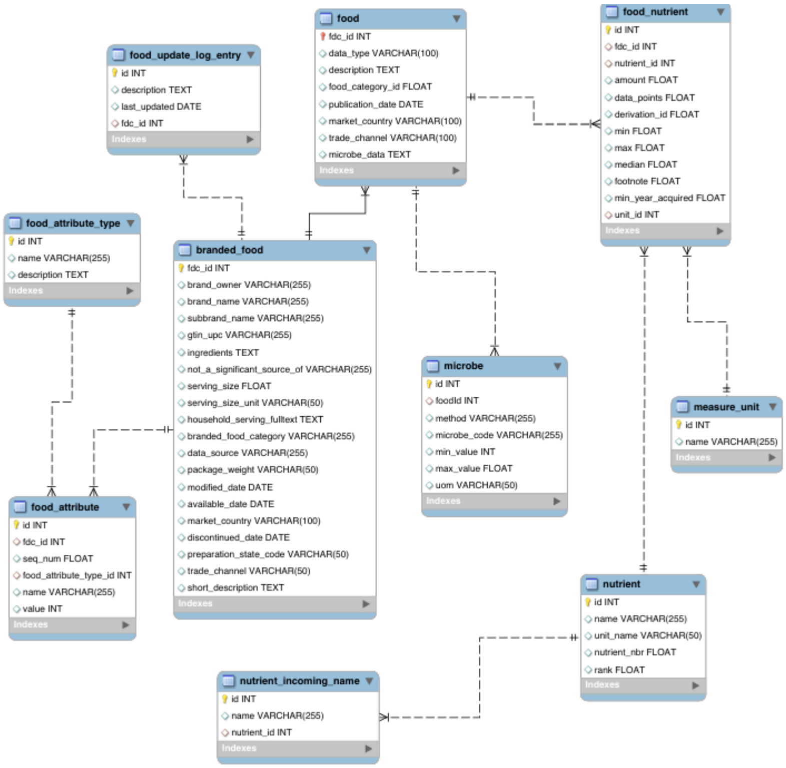

Figure 1: Entity-Relationship Diagram

- **food_update_log_entry:** Tracks updates to food items, recording a `description` of changes, `last_updated` date, and linking to `branded_food` via `fdc_id` (Foreign Key).

- **measure_unit:** Stores measurement units with a unique `id` (Primary Key) and `name` (e.g., grams, milliliters).

- **nutrient:** Contains nutrient information with a unique `id` (Primary Key), `name`, and `unit_name`.

- **food:** General information about food items, with a unique `fdc_id` (Primary Key), `description`, and optional `food_category_id`.

- **food_nutrient:** Links food items to nutrient content using `fdc_id` (Foreign Key referencing `food`), `nutrient_id` (Foreign Key referencing `nutrient`), and includes `amount` and `unit_id` (Foreign Key referencing `measure_unit`).

- **microbe:** Information about microbes in food, linked to the food item via `foodId` (Foreign Key), with fields for detection `method`, and `min_value` and `max_value` of concentration.

- **nutrient_incoming_name:** Maps alternative nutrient names, linking to the standard nutrient via nutrient_id (Foreign Key).

4 Data Sources

The data is obtained from the USDA's FoodData Central data repository. It provides details on a range of food products and we plan to use two key relations:

- branded_food.csv: This file contains product level information, with columns such as brand_name, ingredients, serving_size and the branded_food_category. These fields offer useful insights into the classification and composition of food products.
- food_attribute.csv: This file provides additional details about food products, which may indicate whether a product follows other dietary guidelines. These attributes may influence the food category.

5 Machine Learning and Data Analysis

The purpose of this project is to predict the food category of branded food products using a combination of features through supervised machine learning techniques. We will explore classification models such as logistic regression, random forests to accurately categorize products. To incorporate text-based information like ingredients, we will convert these features into numerical representations using techniques like TF-IDF or bag-of-words. The insights generated from this project could benefit retailers by improving product recommendations and inventory management, leading to more efficient operations and customer satisfaction.

6 Libraries and Tools

We will use a combination of programming libraries and data tools to implement and analyse the USDA's Branded Foods Dataset. Below are the key libraries and tools:

- **Pandas**: Clean, transform, and preprocess the dataset.
- **NumPy**: Handle numerical operations and multidimensional array processing for machine learning tasks.
- **Matplotlib / Seaborn**: Create visualizations such as graphs and heatmaps for data exploration.
- **Scikit-Learn**: Classification, clustering, and predictive modeling tasks like market basket analysis.
- **Jupyter Notebooks**: Prototyping, visualizations, and sharing analysis results.