DSC 672

DATA SCIENCE PROJECT

CAPSTONE FINAL REPORT

Nutri-Recommender – Personalized Food Recommendation System

Team Members (MS Data Science):

Nachiketh Reddy
Aniket Surve
Srinivasan Govindarajan
Arun Subbiah

Abstract:

In today's fast-paced world, maintaining a healthy and balanced diet is paramount for overall well-being. However, with the abundance of food choices available, it can be challenging for individuals to make informed decisions that align with their unique nutritional requirements and dietary preferences. To address this issue, we present Nutri-Recommender, a personalized food recommendation system developed as part of our Data Science Capstone Project. Our objective is to develop a recommendation algorithm that leverages nutritional information and dietary preferences to provide tailored food recommendations.

To achieve this goal, we conducted extensive exploratory data analysis on a comprehensive dataset containing nutritional values for over 8,000 food items. Through data cleaning, transformation, and feature selection, we prepared the dataset for further analysis. Utilizing techniques such as category distribution summary, correlation analysis, and Principal Component Analysis (PCA), we gained valuable insights into the dataset's structure and properties.

Next, we implemented classification models to predict food categories based on nutritional information, achieving significant improvements in accuracy through iterative refinement. Leveraging feature importance analysis and cross-validation techniques, we identified influential features and evaluated the performance of our classification models.

Subsequently, we developed a recommendation system that combines textual descriptions, nutritional features, and GloVe embeddings to generate personalized food recommendations. By integrating content-based filtering techniques and cosine similarity measures, our recommendation system identifies similar food items based on their nutritional composition and textual descriptions.

In our final model, we seamlessly integrate textual descriptions, nutritional features, and GloVe embeddings to provide accurate and personalized food recommendations. Through a series of illustrative examples, we demonstrate the effectiveness of our recommendation system in identifying similar food items tailored to individual preferences.

In conclusion, Nutri-Recommender represents a novel approach to personalized food recommendation, empowering users to make informed dietary choices that align with their nutritional needs and preferences. By leveraging data science techniques and machine learning algorithms, we aim to contribute to improved health outcomes and enhanced user experiences in the realm of nutrition and dietary management.

Introduction:

In an era marked by increasing awareness of health and wellness, the importance of nutrition in maintaining overall well-being cannot be overstated. With the proliferation of food options available in today's market, individuals are often inundated with choices, making it challenging to select foods that meet their specific nutritional requirements and dietary preferences. In response to this challenge, personalized food recommendation systems have emerged as valuable tools for guiding individuals towards healthier dietary choices tailored to their unique needs.

The Nutri-Recommender project represents our endeavor to develop a sophisticated recommendation algorithm that harnesses the power of data science to provide personalized food recommendations. Our project, undertaken as part of the Data Science Capstone course, aims to address the complex interplay between nutrition, dietary preferences, and individual health goals through innovative data-driven approaches.

At the core of our project lies a comprehensive dataset containing detailed nutritional values for thousands of food items. Through exploratory data analysis, we delve into the intricacies of this dataset, uncovering valuable insights that inform our subsequent modeling and recommendation efforts. By leveraging techniques such as data cleaning, transformation, and feature selection, we ensure that our dataset is primed for analysis, setting the stage for the development of our recommendation system.

Building upon this foundation, we explore various machine learning models and classification algorithms to predict food categories based on nutritional information. Through iterative refinement and evaluation, we strive to achieve high levels of accuracy and performance, laying the groundwork for our recommendation system's effectiveness.

Central to our project is the development of a recommendation system that seamlessly integrates textual descriptions, nutritional features, and advanced embedding techniques. By combining content-based filtering with cosine similarity measures, our recommendation system delivers tailored food recommendations that resonate with individual tastes and dietary preferences.

In this report, we present a comprehensive overview of our project, spanning exploratory data analysis, classification modeling, and recommendation system development. Through a combination of data science techniques and machine learning algorithms, we aim to revolutionize the way individuals approach nutrition and dietary management, empowering them to make informed decisions that optimize their health and well-being.

Literature Review:

Introduction

The integration of information technology (IT), information sciences, and nutrition has led to the emergence of nutrition informatics, a field that significantly impacts health care by enhancing the efficiency and effectiveness of dietary management and patient care. Nutrition informatics employs electronic health records, personal health records, and other digital tools to gather, analyze, and utilize data to improve health outcomes (1, 6-17). This summary explores various aspects and developments in food and nutrition recommender systems, emphasizing their role in promoting healthier eating habits and addressing contemporary health challenges.

Nutrition Informatics and Health Recommender Systems

Nutrition informatics has evolved since its inception in 1996, enabling dietitians to leverage electronic tools for patient care, dietary analysis, and nutrient evaluation (1). The concept of personal health records (PHRs) has empowered patients to manage their health data electronically, enhancing their autonomy and access to personalized health information (2). This shift towards digital health records is complemented by health recommender systems (HRS), which use profile-based algorithms to deliver tailored health information and dietary recommendations to users (3).

Challenges and Solutions in Modern Nutrition

The modern lifestyle, characterized by busy schedules and sedentary behavior, contributes to various health issues such as obesity, high blood pressure, and diabetes (4). In industrialized countries, the prevalence of internet access has allowed patients to access global medical knowledge, fostering patient empowerment and informed health decisions (3). However, the abundance of food choices and the complexity of nutrition information often lead to poor dietary habits, necessitating effective recommender systems to guide healthier choices (5).

Recommender Systems in Dietary Management

Food recommender systems have been developed to address the challenges of unhealthy eating habits and to provide personalized dietary advice based on user preferences and nutritional needs. These systems analyze user data to identify patterns and suggest appropriate food choices (6, 7). The integration of nutritional information and user preferences in recommender systems helps users make informed decisions about their diet, ultimately improving health outcomes (8).

Impact on Specific Populations

College students, for instance, are often exposed to environments with high-calorie, low-nutrient foods. Studies have shown that increased nutrition knowledge can positively influence their eating patterns, highlighting the need for effective nutrition labeling and education (9). Similarly, the World Health Organization emphasizes the role of diet in preventing non-communicable diseases, which account for a significant portion of global mortality. Personalized dietary recommendations can play a crucial role in addressing these health issues (10, 12, 15).

Technological Advances and Future Directions

Mobile phone sensor technologies and multimedia data have created new opportunities for advanced health systems to provide real-time dietary guidance (11). The development of more comprehensible nutrition metrics and labeling systems aims to improve consumer choices and health outcomes (12). The role of text mining and information retrieval in e-commerce also extends to nutrition, enabling more effective recommendation systems through the extraction of relevant information from diverse sources (17, 18).

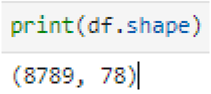
The integration of IT and nutrition sciences through nutrition informatics and recommender systems has the potential to significantly improve dietary habits and health outcomes. By leveraging electronic health records, personal health records, and advanced algorithms, these systems provide personalized dietary advice that addresses individual needs and preferences. As the field continues to evolve, further research and development will enhance the effectiveness of these systems, contributing to better health and well-being on a global scale.

Objective:
Develop a recommendation algorithm that takes into account unique nutritional requirements and dietary preferences and provide tailored results.

1. **EDA:**

DATASET: https://www.kaggle.com/datasets/trolukovich/nutritional-values-for-common-foods-and-products

This dataset contains nutrition values for about 8,789 types of food. The columns consists of the nutritional value of that particular food.



```
print(df.shape)
```

```
(8789, 78)
```

Data Cleaning and Transformation:
Utilized Power Query in Excel to change the data types of all columns to numeric.
Added an additional column named "category" that extracts the first word from the "name" column, which contains food descriptions. This categorization can help organize the data based on food types or categories. This initial preprocessing step helps in preparing the data for further analysis and ensures that the data is in a suitable format for conducting EDA and other tasks.

Category Distribution Summary:

The 'Category' column provides insights into the distribution of different food types or categories in the dataset. After analyzing the data, the top ten most common elements in the 'Category' column are as follows:

1. **BEEF: 967**
2. **CEREALS: 354**
3. **PORK: 336**
4. **LAMB: 295**
5. **BEVERAGES: 282**
6. **BABYFOOD: 243**
7. **FISH: 238**
8. **CHICKEN: 214**
9. **SOUP: 176**
10. **CAMPBELL'S: 156**

These figures provide an overview of the frequency of different food categories present in the dataset. It indicates that beef products are the most commonly occurring category, followed by pork, lamb, and beverages. This distribution can inform further analysis and decision-making processes related to food categorization and classification tasks.

Column Selection and Removal:

In the process of refining the dataset for our Nutri-Recommender system, we've decided to focus on columns that are highly relevant to our domain and essential for personalized food recommendations. This decision is based on domain knowledge and the importance of each feature in providing accurate nutritional insights and recommendations.

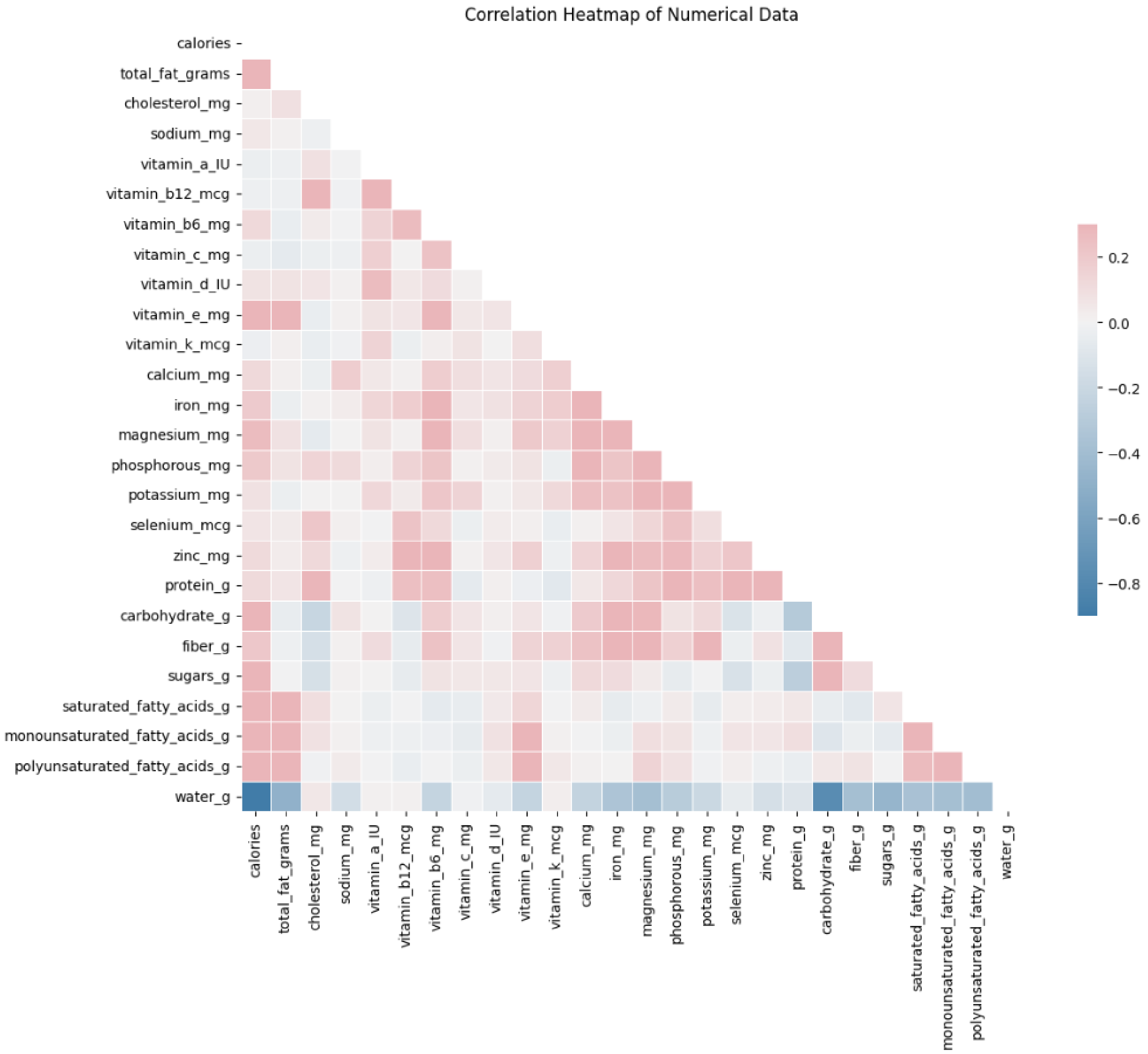
Considering the variation in numerical columns and the number of empty rows, we have selected the following columns for inclusion in our finalized dataset:

1. **name**: The description of the food item.
2. **calories**: The energy content of the food item (in calories).
3. **total_fat_grams**: The total amount of fat in the food item (in grams).
4. **cholesterol_mg:** The amount of cholesterol in the food item (in milligrams).
5. **sodium_mg:** The sodium content in the food item (in milligrams).
6. **vitamin_a_IU:** The amount of Vitamin A in the food item (in International Units).
7. **vitamin_b12_mcg:** The amount of Vitamin B12 in the food item (in micrograms).
8. **vitamin_b6_mg:** The amount of Vitamin B6 in the food item (in milligrams).
9. **vitamin_c_mg:** The amount of Vitamin C in the food item (in milligrams).
10. **vitamin_d_IU:** The amount of Vitamin D in the food item (in International Units).
11. **vitamin_e_mg:** The amount of Vitamin E in the food item (in milligrams).
12. **vitamin_k_mcg:** The amount of Vitamin K in the food item (in micrograms).
13. **calcium_mg:** The amount of calcium in the food item (in milligrams).
14. **iron_mg:** The amount of iron in the food item (in milligrams).
15. **magnesium_mg:** The amount of magnesium in the food item (in milligrams).
16. **phosphorous_mg:** The amount of phosphorus in the food item (in milligrams).
17. **potassium_mg:** The amount of potassium in the food item (in milligrams).
18. **selenium_mcg:** The amount of selenium in the food item (in micrograms).
19. **zinc_mg:** The amount of zinc in the food item (in milligrams).
20. **protein_g:** The protein content in the food item (in grams).
21. **carbohydrate_g:** The carbohydrate content in the food item (in grams).
22. **fiber_g:** The fiber content in the food item (in grams).
23. **sugars_g:** The sugar content in the food item (in grams).
24. **saturated_fatty_acids_g:** The amount of saturated fatty acids in the food item (in grams).
25. **monounsaturated_fatty_acids_g:** The amount of monounsaturated fatty acids in the food item (in grams).
26. **polyunsaturated_fatty_acids_g:** The amount of polyunsaturated fatty acids in the food item (in grams).
27. **water_g:** The water content in the food item (in grams).
28. **Category:** The category of the food item.

These columns provide crucial nutritional information necessary for generating personalized food recommendations. By focusing on these features, we aim to simplify the analysis process while ensuring the relevance and accuracy of our recommendations.

Correlation Analysis:

To identify any multicollinearity among the selected features, we conducted a correlation analysis using a correlation plot. The correlation plot shows the correlation coefficients between each pair of features in our dataset. Here's a summary of the findings:



Removed Variable:
Water_g: We observed a strong negative correlation between water content and many other features in the dataset. Since water content is not directly related to the nutritional value of the food items and may introduce multicollinearity issues, we decided to remove this variable from further analysis.

Generating Profile Report using Pandas Profile Reporting for Further EDA:

As part of our exploratory data analysis (EDA), we utilized the Pandas Profile Reporting library to generate a comprehensive profile report for our nutrition dataset. This report provides valuable insights into the structure, distribution, and characteristics of the data, helping us gain a deeper understanding of its properties.

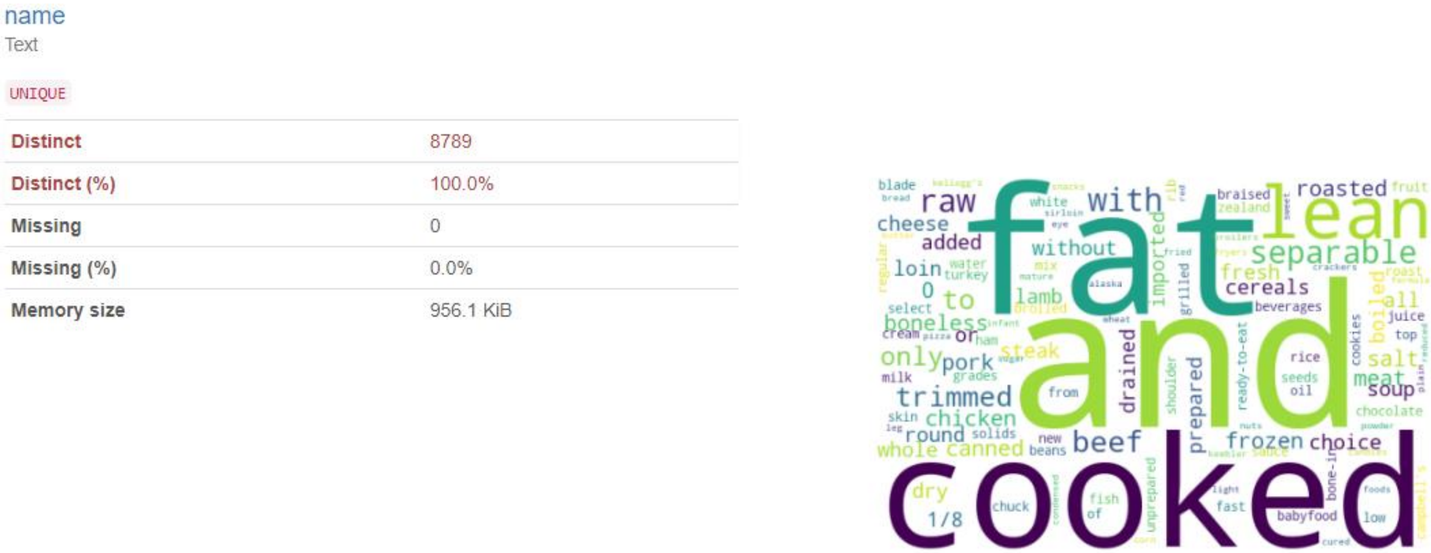
The Profile Report, titled "Nutrition Dataset Report," was created using the ProfileReport function from the ydata_profiling module. By setting explorative=True, we enabled additional exploratory analysis features in the report, allowing us to uncover hidden patterns and anomalies within the data.

The report covers various aspects of the dataset, including:

1. **Overview**: Provides a summary of the dataset, including the number of variables, observations, and missing values.
2. **Variables**: Presents detailed information about each variable, including data type, unique values, and missing value percentages.
3. **Statistics**: Offers descriptive statistics such as mean, median, standard deviation, minimum, maximum, and quantiles for numerical variables.
4. **Correlations**: Displays correlation matrices and heatmaps to visualize the relationships between different variables in the dataset.
5. **Distribution**: Illustrates the distribution of numerical variables through histograms and kernel density estimation plots.
6. **Interactions**: Examines interactions between variables, including scatter plots and pair-wise correlation coefficients.
7. **Missing Values**: Identifies patterns and frequencies of missing values in the dataset.
8. **Warnings and Detections**: Flags potential issues or anomalies in the data, such as high cardinality or constant variables.

By leveraging the insights provided by the Profile Report, we were able to make informed decisions during the data preprocessing and feature selection stages. Additionally, the visualizations and statistical summaries offered valuable guidance for building our Nutri-Recommender system, ensuring its accuracy and effectiveness.

Furthermore, alongside the Profile Report, we supplemented our analysis with word visualizations of the "Name" and "Category" columns, showcasing the distribution of words within these columns. These visualizations provided additional context and helped us better understand the textual data in our dataset.

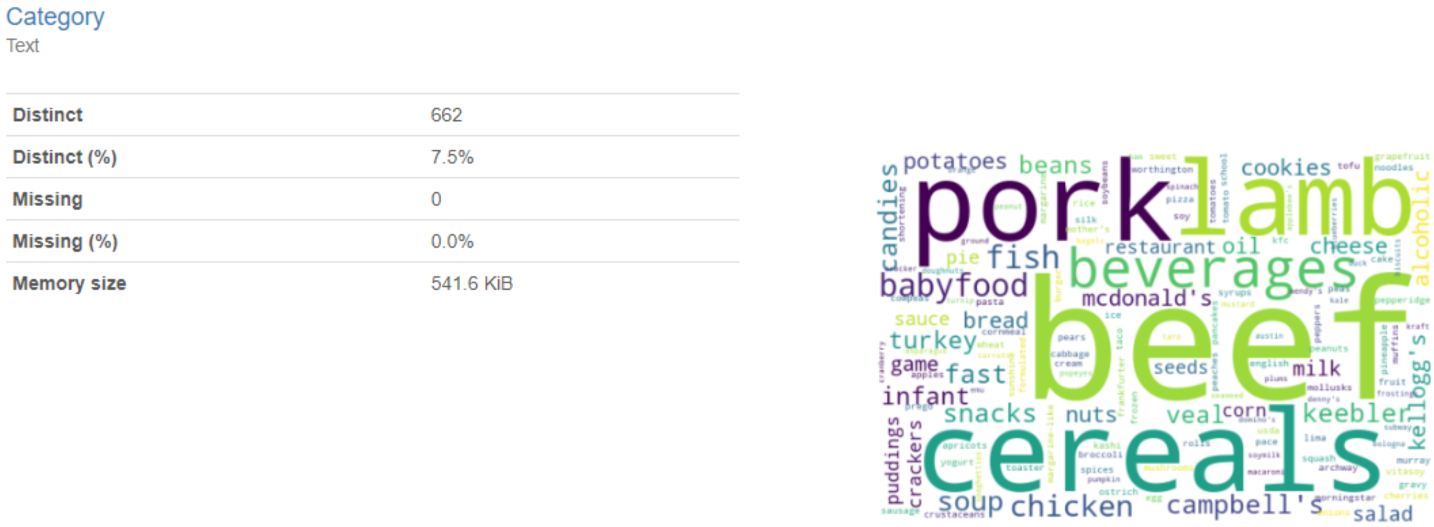


name

Text

UNIQUE

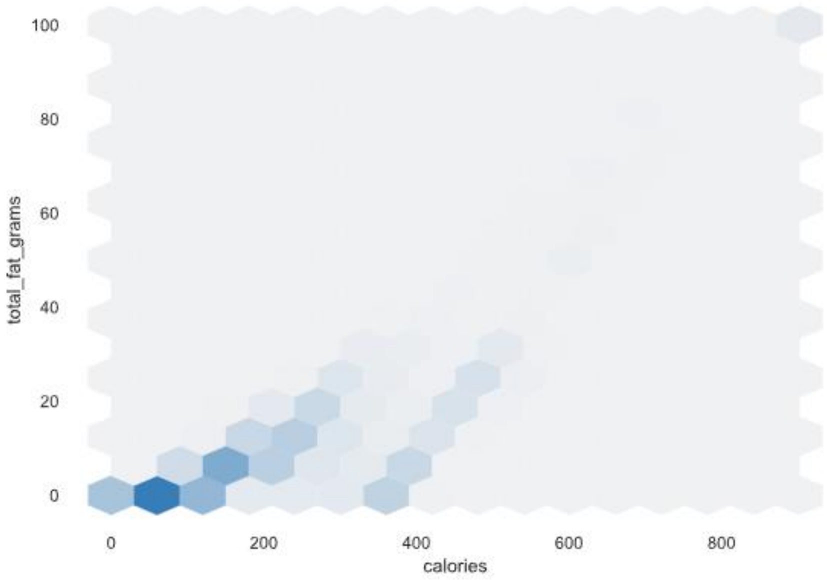
Distinct	8789
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	956.1 KiB



Category

Text

Distinct	662
Distinct (%)	7.5%
Missing	0
Missing (%)	0.0%
Memory size	541.6 KiB



We can also view correlation between different features through correlation plots.

2. Performing PCA for Dimensionality Reduction:

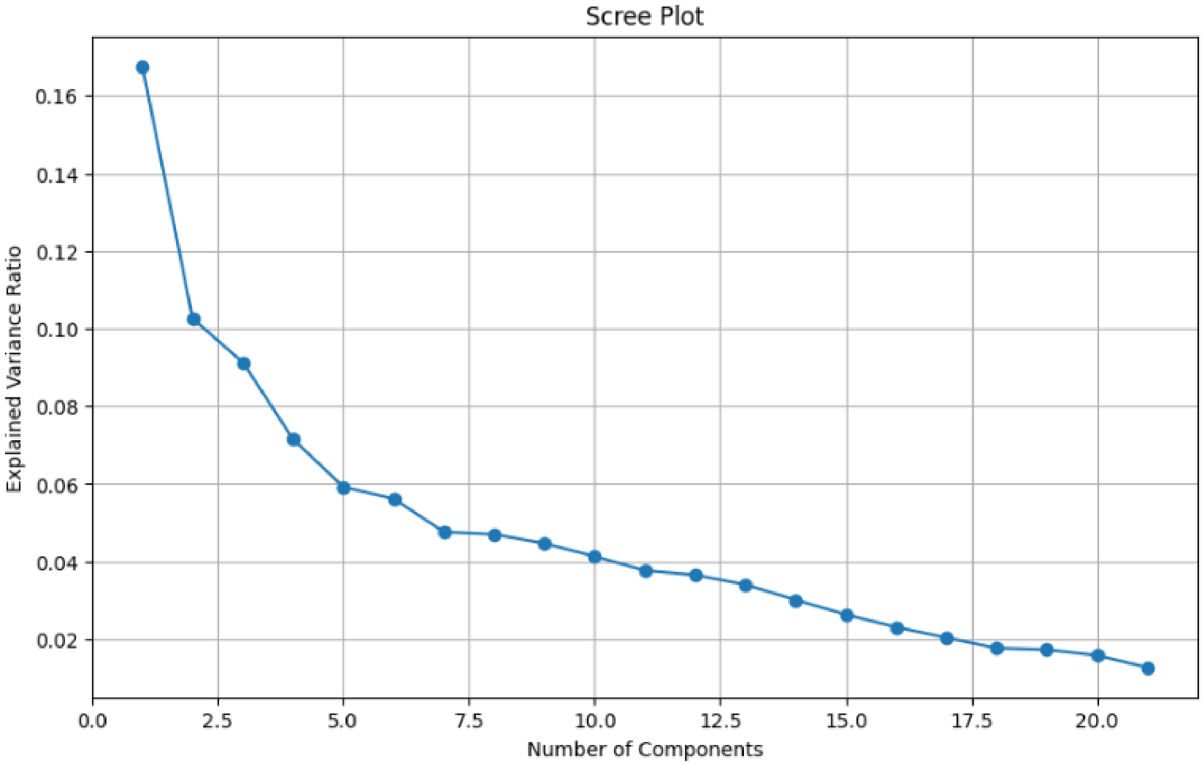
In our quest to simplify the dataset for further analysis while retaining essential information, we employed Principal Component Analysis (PCA). PCA is a powerful technique for reducing the dimensionality of data while preserving its variance. Here's how we executed PCA and its implications for our dataset:

Exclusion of Columns:
We strategically excluded certain columns ('calories', 'total_fat_grams', 'protein_g', 'carbohydrate_g', 'Category', and 'name') from our dataset. These exclusions were made based on domain knowledge and the specific requirements of our Nutri-Recommender system algorithm.

Standardization:
Before applying PCA, we standardized the numerical data using StandardScaler. Standardization is essential for PCA as it ensures that all features contribute equally to the analysis by scaling them to have a mean of 0 and a standard deviation of 1.

PCA Calculation:
We then performed PCA on the standardized numerical data. PCA identifies the principal components—linear combinations of the original features—that capture the maximum variance in the data.

Scree Plot Analysis:
To determine the optimal number of principal components to retain, we plotted a scree plot. This plot visualizes the explained variance ratio of each principal component. We identified the number of components that explain a significant portion of the variance in the data.

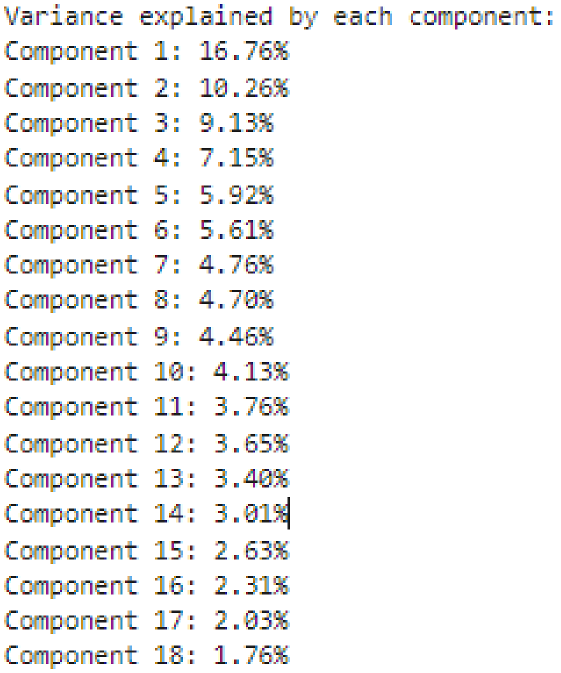


Selecting Components:

We chose to retain the first 10 principal components, which collectively explain 70% of the variance in the dataset. This decision strikes a balance between dimensionality reduction and information retention.

Interpreting Principal Components:

We analyzed the loadings of the original features on each principal component to interpret their meaning. Each principal component represents a unique combination of features that contribute to its variance. For example, PC1 is influenced by nutrients like magnesium, iron, and vitamin B6, while PC2 is driven by fatty acids such as monounsaturated and polyunsaturated fats.



```
Variance explained by each component:
Component 1: 16.76%
Component 2: 10.26%
Component 3: 9.13%
Component 4: 7.15%
Component 5: 5.92%
Component 6: 5.61%
Component 7: 4.76%
Component 8: 4.70%
Component 9: 4.46%
Component 10: 4.13%
Component 11: 3.76%
Component 12: 3.65%
Component 13: 3.40%
Component 14: 3.01%
Component 15: 2.63%
Component 16: 2.31%
Component 17: 2.03%
Component 18: 1.76%
```

Final Dataset:

The final dataset after PCA consists of the original columns (including those excluded earlier) along with the 10 principal components. These principal components serve as compressed representations of the original data, facilitating simpler and more efficient analysis.

By leveraging PCA, we have effectively reduced the dimensionality of our dataset while preserving its essential information. This streamlined dataset will serve as a foundation for our subsequent analysis and the development of our Nutri-Recommender system algorithm.

3. Classification

In this section, we present the results of our classification models for predicting food categories based on nutritional information. We conducted three classifications, each refining our approach to improve accuracy and efficiency. The final classification, referred to as Classification 3, achieved an accuracy of 87% and a cross-validation mean accuracy of 81%. We'll delve into the details of this classification, as well as provide an overview of the previous two classifications for comparison.

Classification 1:
In the initial classification attempt, we employed a Random Forest Classifier without feature selection. The model achieved an accuracy of 66%. While this result provided a baseline understanding of the classification task, we aimed to enhance the model's performance by refining feature selection and parameter tuning.

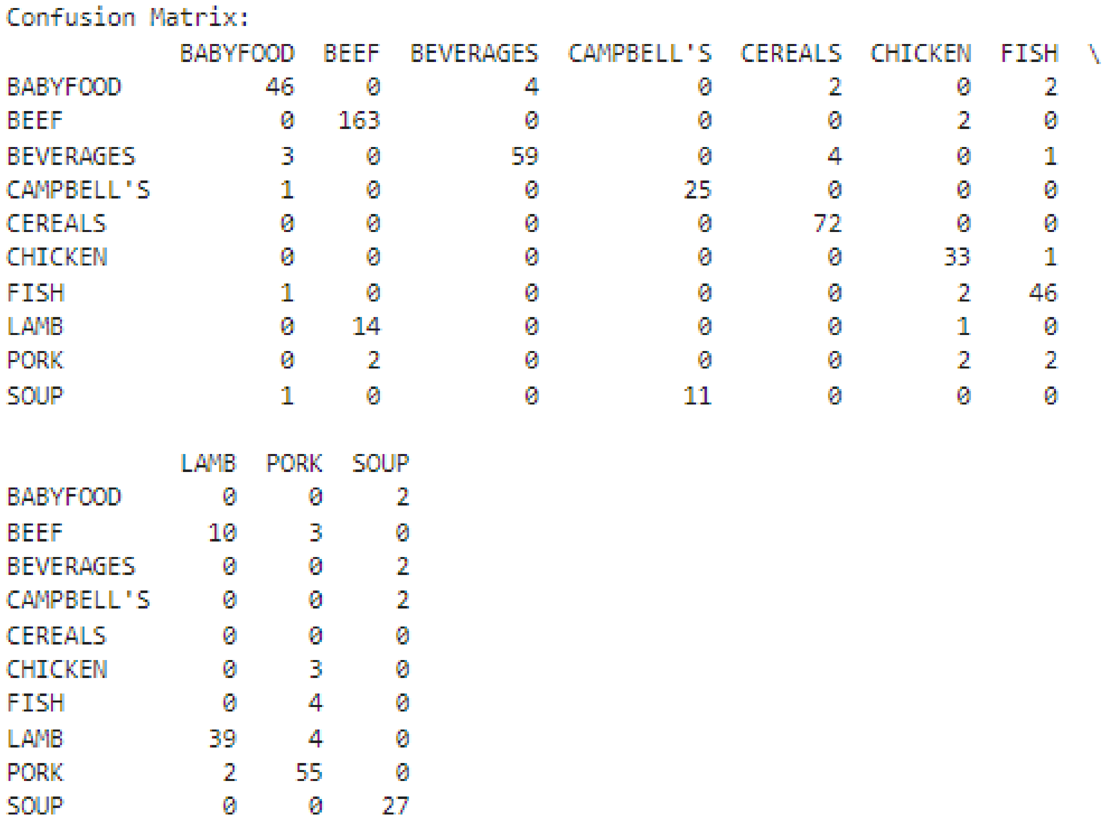
Classification 2:
In the second classification iteration, we focused on the top ten food categories, which represented a significant portion of the dataset. This approach improved the accuracy to 89%. By selecting the most relevant categories, we reduced the complexity of the classification task and achieved a higher accuracy rate.

Classification 3:
The final classification utilized feature importance analysis to select the most influential features for prediction. By considering features such as calories, macronutrients, and principal components (PCs), we refined our model and achieved an accuracy of 87%. The feature importance plot revealed that principal components, particularly PC5, played a significant role in the classification process, followed by protein content and calories.

Among the features considered, principal components (PCs) emerged as particularly influential, with PC5 standing out as the most significant contributor, followed closely by PC9. These principal components likely encapsulate complex relationships within the dataset, capturing nuanced patterns that aid in distinguishing between different food categories. Additionally, macronutrients such as protein, carbohydrates, and total fat grams exhibited substantial importance, underscoring the relevance of nutritional composition in categorizing food items. Calories also played a notable role, albeit to a slightly lesser extent, suggesting that overall energy content contributes significantly to category differentiation. This comprehensive understanding of feature importance informs the development of more effective classification models and highlights the multidimensional nature of food categorization, integrating both macronutrient composition and underlying structural patterns within the dataset.

Evaluation Metrics for Classification 3:

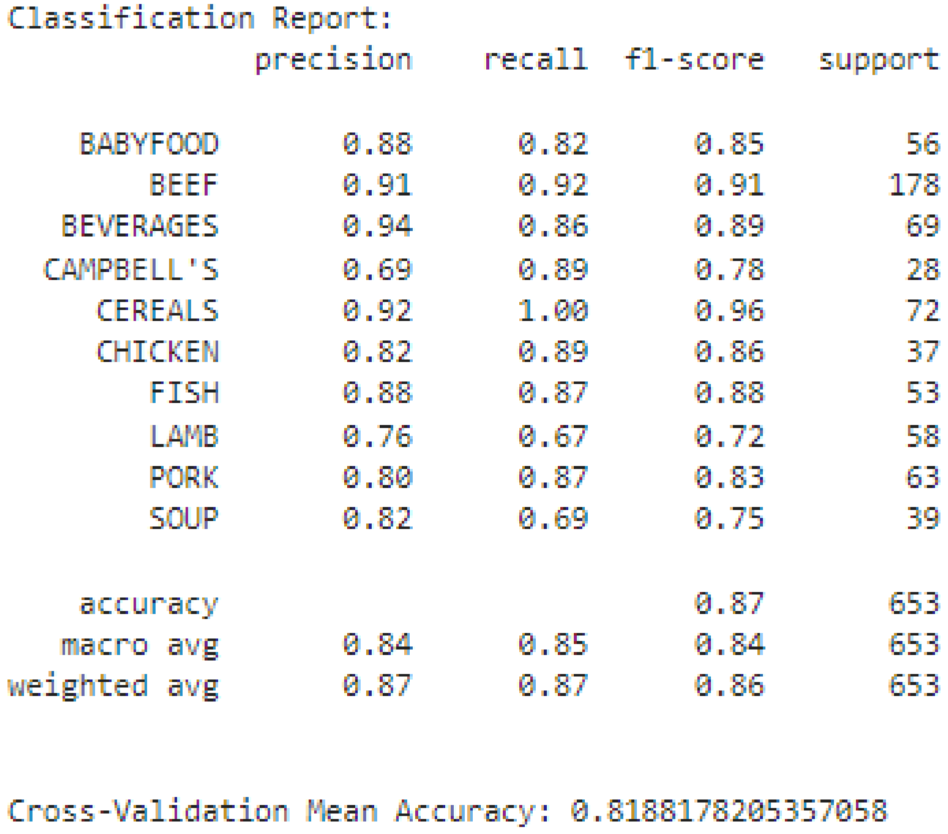
Confusion Matrix:
The confusion matrix provides a detailed breakdown of the classifier's performance across different categories. It enables us to assess both correct and incorrect predictions for each category.



```
Confusion Matrix:
            BABYFOOD  BEEF  BEVERAGES  CAMPBELL'S  CEREALS  CHICKEN  FISH  \
BABYFOOD          46     0          4           0        2        0     2
BEEF               0   163          0           0        0        2     0
BEVERAGES          3     0         59           0        4        0     1
CAMPBELL'S         1     0          0          25        0        0     0
CEREALS            0     0          0           0       72        0     0
CHICKEN            0     0          0           0        0       33     1
FISH               1     0          0           0        0        2    46
LAMB               0    14          0           0        0        1     0
PORK               0     2          0           0        0        2     2
SOUP               1     0          0          11        0        0     0

            LAMB  PORK  SOUP
BABYFOOD       0     0     2
BEEF          10     3     0
BEVERAGES      0     0     2
CAMPBELL'S     0     0     2
CEREALS        0     0     0
CHICKEN        0     3     0
FISH           0     4     0
LAMB          39     4     0
PORK           2    55     0
SOUP           0     0    27
```

Precision, Recall, and F1-Score:

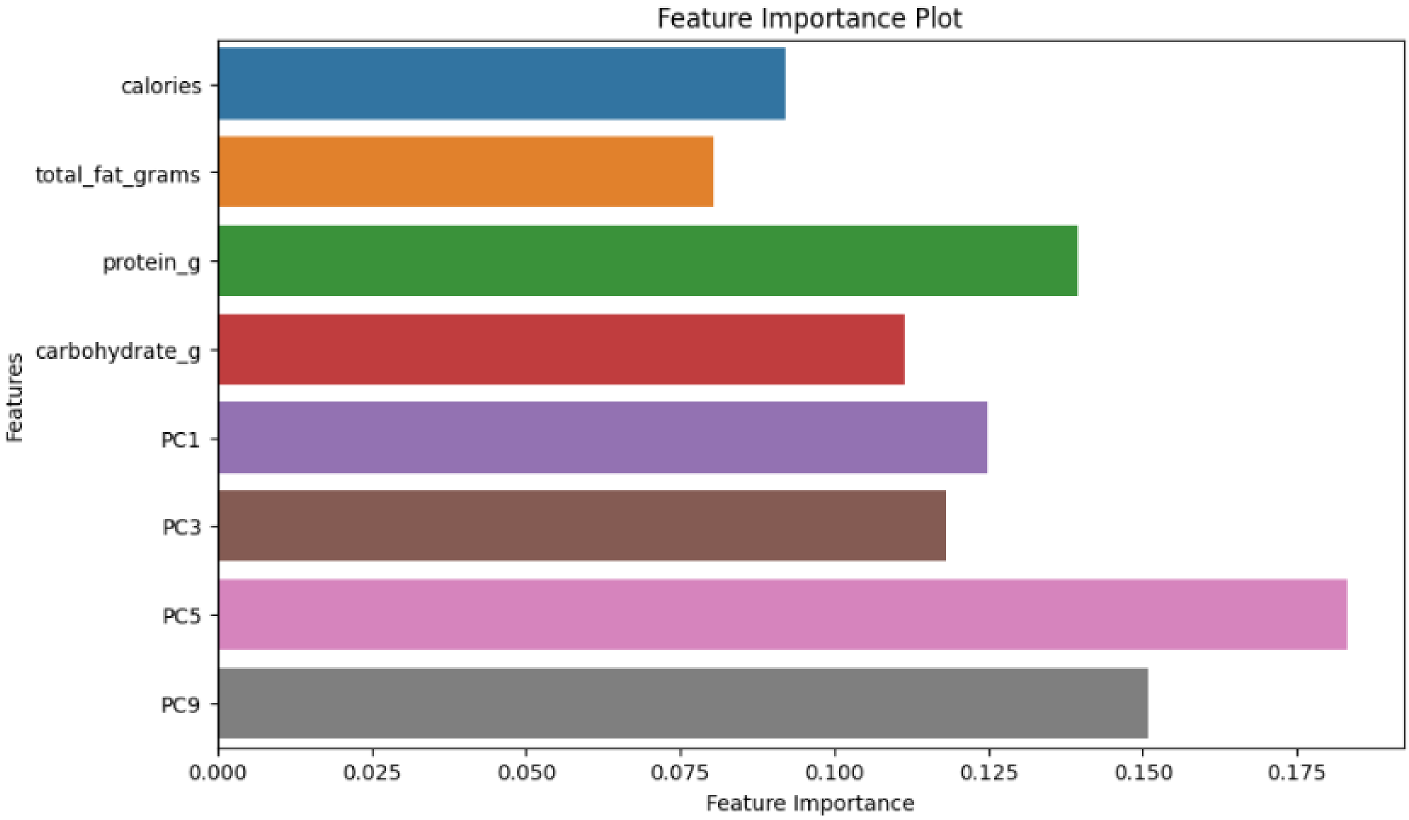
Precision measures the ratio of correctly predicted positive observations to the total predicted positives, while recall measures the ratio of correctly predicted positive observations to the total actual positives. F1-score provides a balance between precision and recall.

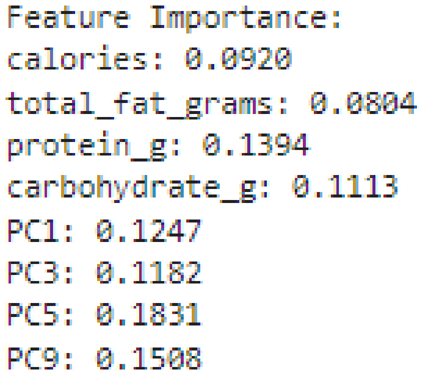


```
Classification Report:
              precision    recall  f1-score   support

    BABYFOOD       0.88      0.82      0.85        56
        BEEF       0.91      0.92      0.91       178
   BEVERAGES       0.94      0.86      0.89        69
  CAMPBELL'S       0.69      0.89      0.78        28
     CEREALS       0.92      1.00      0.96        72
     CHICKEN       0.82      0.89      0.86        37
        FISH       0.88      0.87      0.88        53
        LAMB       0.76      0.67      0.72        58
        PORK       0.80      0.87      0.83        63
        SOUP       0.82      0.69      0.75        39

    accuracy                           0.87       653
   macro avg       0.84      0.85      0.84       653
weighted avg       0.87      0.87      0.86       653


Cross-Validation Mean Accuracy: 0.8188178205357058
```

Cross-Validation Mean Accuracy:

Cross-validation is a robust technique for estimating the performance of a machine learning model on unseen data. The cross-validation mean accuracy of 81% indicates that the model generalizes well to new data and is less likely to overfit.



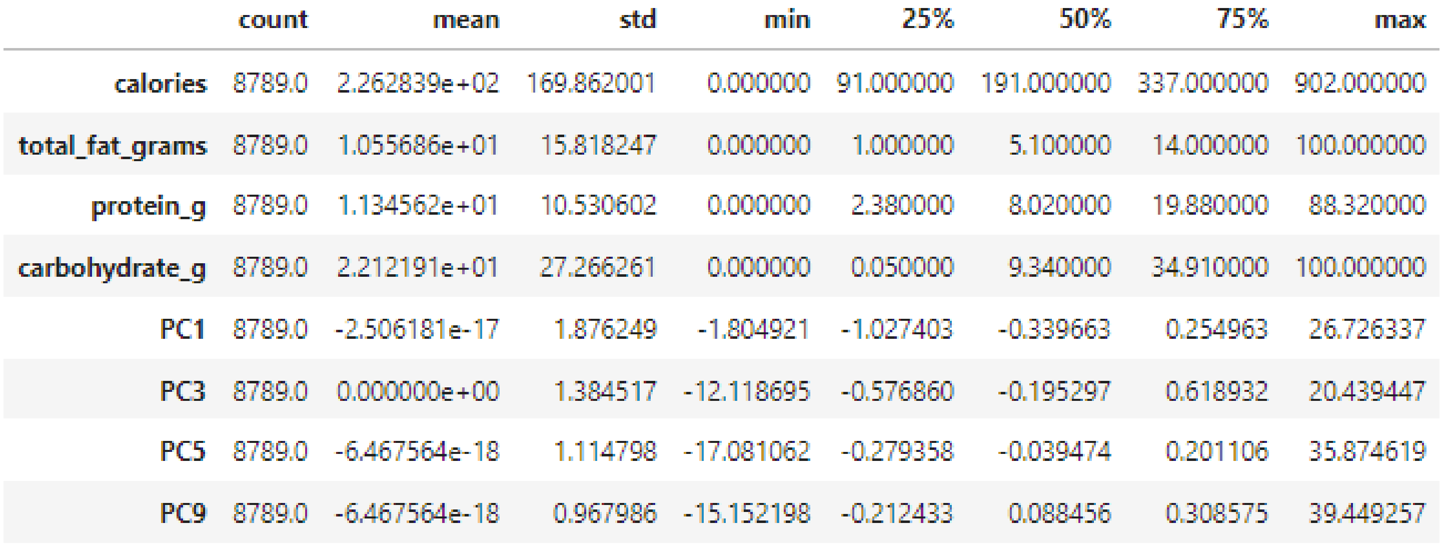


```
Feature Importance:
calories: 0.0920
total_fat_grams: 0.0804
protein_g: 0.1394
carbohydrate_g: 0.1113
PC1: 0.1247
PC3: 0.1182
PC5: 0.1831
PC9: 0.1508
```

Classification 3 represents the culmination of our efforts to improve the accuracy and efficiency of our food category prediction model. By leveraging feature importance analysis and selecting relevant features, we achieved a significant improvement in accuracy compared to the initial classification attempts. The insights gained from this classification process can inform dietary recommendations, food labeling, and nutritional research, contributing to improved health outcomes and food product development.

4. Recommender Analysis:

we aim to construct a recommender system that suggests similar food items based on their nutritional composition and categorical similarities. The dataset used for this project contains information about various food items, including their macronutrient content and principal component features derived from nutritional attributes.



	count	mean	std	min	25%	50%	75%	max
calories	8789.0	2.262839e+02	169.862001	0.000000	91.000000	191.000000	337.000000	902.000000
total_fat_grams	8789.0	1.055686e+01	15.818247	0.000000	1.000000	5.100000	14.000000	100.000000
protein_g	8789.0	1.134562e+01	10.530602	0.000000	2.380000	8.020000	19.880000	88.320000
carbohydrate_g	8789.0	2.212191e+01	27.266261	0.000000	0.050000	9.340000	34.910000	100.000000
PC1	8789.0	-2.506181e-17	1.876249	-1.804921	-1.027403	-0.339663	0.254963	26.726337
PC3	8789.0	0.000000e+00	1.384517	-12.118695	-0.576860	-0.195297	0.618932	20.439447
PC5	8789.0	-6.467564e-18	1.114798	-17.081062	-0.279358	-0.039474	0.201106	35.874619
PC9	8789.0	-6.467564e-18	0.967986	-15.152198	-0.212433	0.088456	0.308575	39.449257

4.1 Get the Descriptions of the Top Similar Items Using Item Description

In this subsection, we detail the process of retrieving the descriptions of the top similar items based on a given item description. The objective is to recommend similar food items by leveraging their textual descriptions and nutritional features. The steps involved include:

Data Preprocessing and Feature Selection:
We start by selecting relevant columns from the dataset, including the food category, item name, and key nutritional features such as calories, total fat grams, protein grams, carbohydrate grams, and principal component features (PC1, PC3, PC5, PC9). These features provide a comprehensive representation of each food item's nutritional profile and structural characteristics.

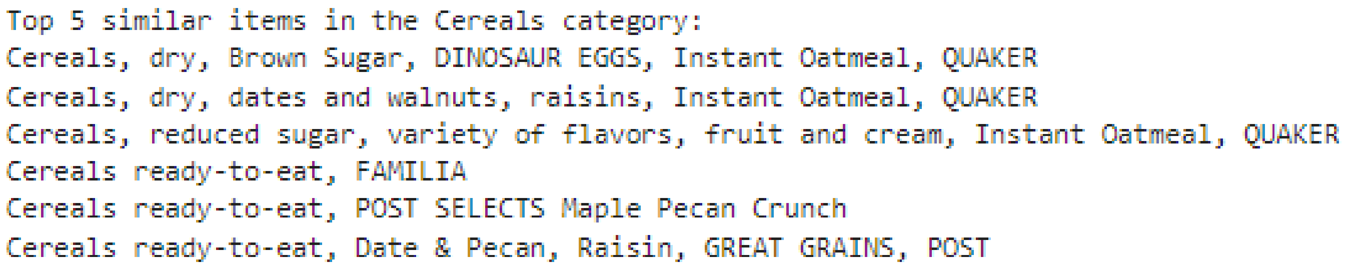
Similarity Calculation Based on Item Features:
We define a similarity function to compute the cosine similarity between the features of a given item and those of other items within the same category. This function takes into account the nutritional composition and structural similarities between food items.

Retrieving Top Similar Items:
Using the defined similarity function, we identify the top similar items within the same food category as the given item. The similarity scores are computed based on the cosine similarity measure, and the top similar items are ranked accordingly. We retrieve the top 5 similar items and present their descriptions, including category, name, calories, total fat grams, protein grams, carbohydrate grams, and cosine similarity measure.

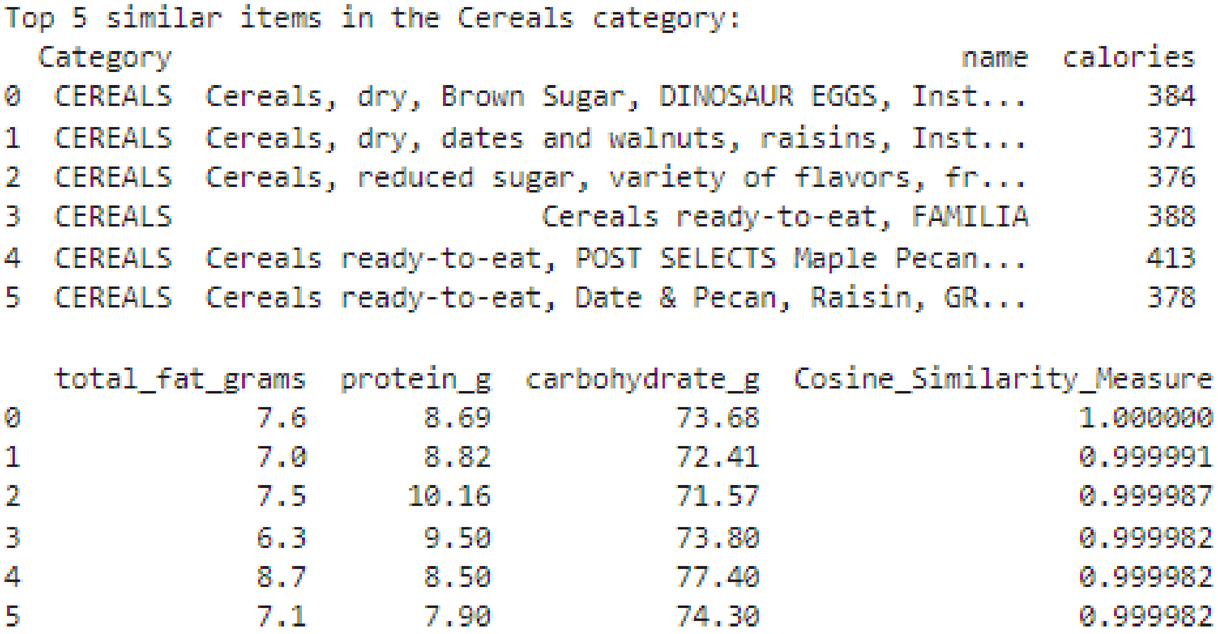
Results and Insights:

For demonstration purposes, we consider the "Cereals" category and provide examples of the top 5 similar items based on their nutritional composition and structural characteristics. The recommender system effectively identifies items with similar macronutrient profiles and principal component features, facilitating personalized recommendations for users interested in exploring similar food options.

```
cat_items = df_selected[df_selected['Category'] == 'CEREALS']
# Given item description
given_item_description = "Cereals, dry, Brown Sugar, DINOSAUR EGGS, Instant Oatmeal, QUAKER"
```



```
Top 5 similar items in the Cereals category:
Cereals, dry, Brown Sugar, DINOSAUR EGGS, Instant Oatmeal, QUAKER
Cereals, dry, dates and walnuts, raisins, Instant Oatmeal, QUAKER
Cereals, reduced sugar, variety of flavors, fruit and cream, Instant Oatmeal, QUAKER
Cereals ready-to-eat, FAMILIA
Cereals ready-to-eat, POST SELECTS Maple Pecan Crunch
Cereals ready-to-eat, Date & Pecan, Raisin, GREAT GRAINS, POST
```



```
Top 5 similar items in the Cereals category:
  Category                                               name  calories  \
0  CEREALS  Cereals, dry, Brown Sugar, DINOSAUR EGGS, Inst...       384
1  CEREALS  Cereals, dry, dates and walnuts, raisins, Inst...       371
2  CEREALS  Cereals, reduced sugar, variety of flavors, fr...       376
3  CEREALS                      Cereals ready-to-eat, FAMILIA       388
4  CEREALS  Cereals ready-to-eat, POST SELECTS Maple Pecan...       413
5  CEREALS  Cereals ready-to-eat, Date & Pecan, Raisin, GR...       378

   total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure
0              7.6       8.69           73.68                   1.000000
1              7.0       8.82           72.41                   0.999991
2              7.5      10.16           71.57                   0.999987
3              6.3       9.50           73.80                   0.999982
4              8.7       8.50           77.40                   0.999982
5              7.1       7.90           74.30                   0.999982
```

4.2 Recommend Items Based on Their Features (Calories, Fats, Proteins, Carbohydrates)

In this section, we outline the process of recommending food items based on their nutritional features, including calories, fats, proteins, and carbohydrates. The approach involves leveraging content-based filtering techniques to identify items with similar nutritional profiles to a given set of features. The steps are as follows:

Sample Feature Space and Data Preprocessing:

We start by defining a sample feature vector representing the desired nutritional composition, including calories, protein grams, total fat grams, and carbohydrate grams. Additionally, we preprocess the textual data in the "name" column using TF-IDF vectorization to convert it into numerical form. This enables us to incorporate textual information into the recommendation process.

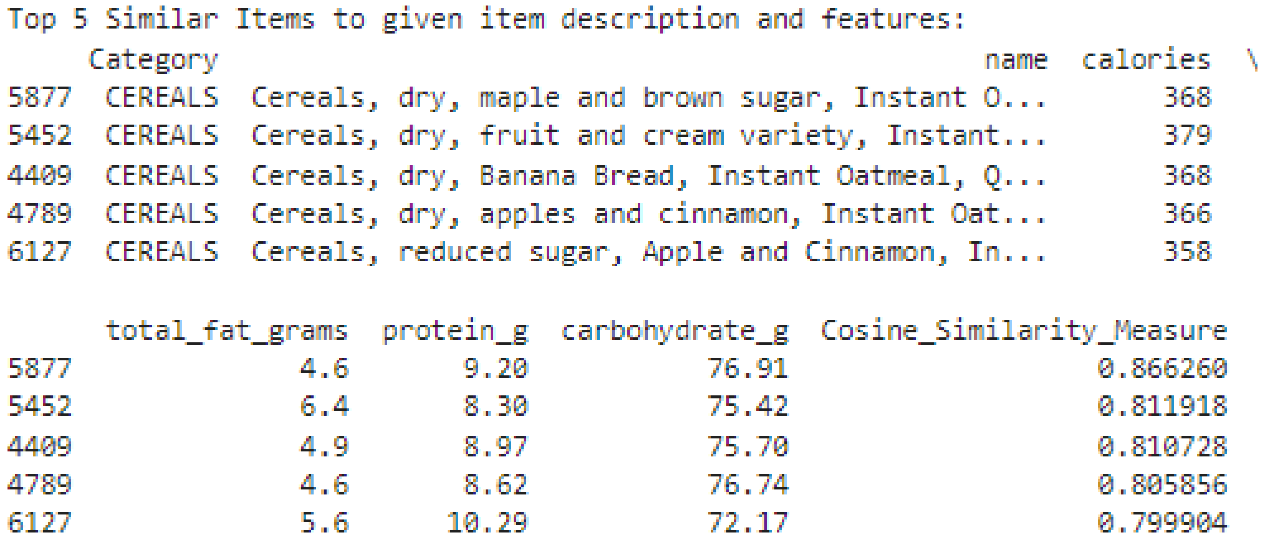
Standardization and Feature Concatenation:
Next, we standardize the numerical features to ensure consistency in scale across different dimensions. We then concatenate the standardized numerical features with the TF-IDF vectors representing the textual descriptions. This combined feature representation captures both the nutritional content and textual information of each food item.

Cosine Similarity Calculation:
We compute the cosine similarity between the feature vectors of the given sample and all items in the dataset. This similarity measure quantifies the degree of similarity between the features of different food items. Higher cosine similarity values indicate greater similarity in nutritional composition and textual description.

Recommending Similar Items:
Based on the calculated cosine similarity scores, we identify the top N similar items to the given sample features. These recommended items exhibit similar nutritional profiles and textual descriptions to the input features. The recommendation process enables users to explore alternative food options that closely match their dietary preferences and nutritional requirements.

Results and Insights:
As an illustrative example, we showcase the top 5 recommended items based on the sample feature vector. These recommendations include items from various categories, such as cereals, oats, wheat, and cocoa products, with similar nutritional compositions to the input features. The cosine similarity measure provides a quantitative assessment of the similarity between the recommended items and the given sample, facilitating personalized food recommendations for users.

```
# Sample feature space including only 'calories', 'protein_g', 'total_fat_grams', 'carbohydrate_g'
sample_features = np.array([[350, 10, 10, 70]])  # Example feature vector
# Given item description
given_item_description = "Cereals, dry, Brown Sugar, DINOSAUR EGGS, Instant Oatmeal, QUAKER"
```



```
Top 5 Similar Items to given item description and features:
     Category                                               name  calories  \
5877  CEREALS  Cereals, dry, maple and brown sugar, Instant O...       368
5452  CEREALS  Cereals, dry, fruit and cream variety, Instant...       379
4409  CEREALS  Cereals, dry, Banana Bread, Instant Oatmeal, Q...       368
4789  CEREALS  Cereals, dry, apples and cinnamon, Instant Oat...       366
6127  CEREALS  Cereals, reduced sugar, Apple and Cinnamon, In...       358

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure
5877              4.6       9.20           76.91                   0.866260
5452              6.4       8.30           75.42                   0.811918
4409              4.9       8.97           75.70                   0.810728
4789              4.6       8.62           76.74                   0.805856
6127              5.6      10.29           72.17                   0.799904
```

4.3 Recommend Items Based on Both Description and Features Using GloVe Embeddings

In this section, we extend the recommendation process by incorporating both textual descriptions and nutritional features using GloVe word embeddings. The approach combines textual embeddings derived from item descriptions with numerical features to enhance the recommendation accuracy. The steps involved are as follows:

Loading GloVe Embeddings:
We begin by loading pre-trained GloVe embeddings of a specified dimension (e.g., 50 dimensions). GloVe embeddings provide dense vector representations of words, capturing semantic similarities between them.

Tokenization and Text Embedding:
Next, we tokenize the textual data in the "Category" and "name" columns, preprocess it, and create embeddings using GloVe for each tokenized word. These embeddings capture the semantic meaning of words in the item descriptions.

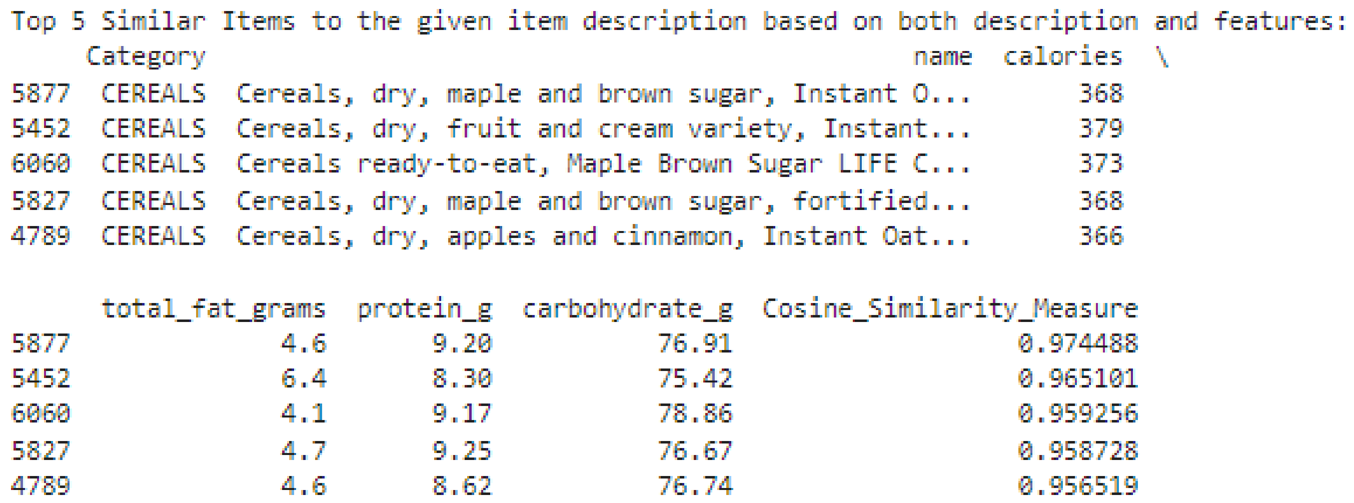
Feature Concatenation:
We standardize the numerical features representing calories, total fat grams, protein grams, and carbohydrate grams. These standardized features are then concatenated with the GloVe embeddings of the textual descriptions, forming a combined feature representation.

Cosine Similarity Calculation:
We compute the cosine similarity between the combined feature vectors of the given item and all items in the dataset. This similarity measure quantifies the similarity in both textual descriptions and nutritional features.

Recommending Similar Items:
Based on the calculated cosine similarity scores, we identify the top N similar items to the given item description and features. These recommended items exhibit similarities in both textual descriptions and nutritional compositions to the input features, providing users with more relevant recommendations.

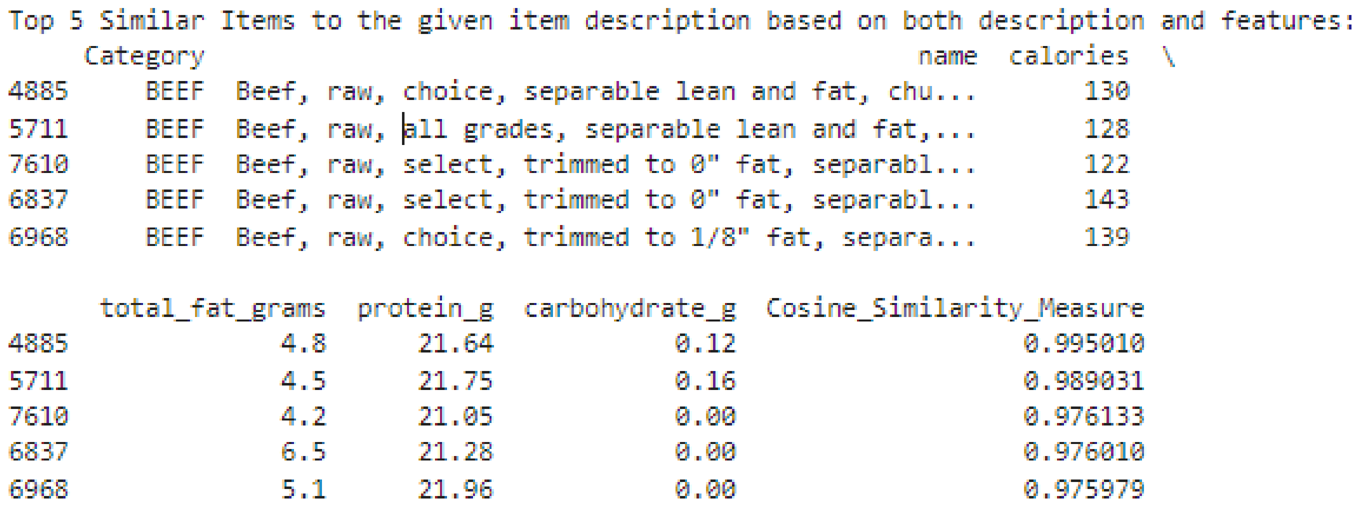
Results and Insights:
As an example, we showcase the top 5 recommended items based on the given item description and sample features. These recommendations include items from various categories, with both textual descriptions and nutritional features closely matching those of the input item. The cosine similarity measure provides a quantitative assessment of the similarity between the recommended items and the given item description and features, facilitating personalized and context-aware recommendations for users.

```
# Sample feature space including only 'calories', 'protein_g', 'total_fat_grams', 'carbohydrate_g'
sample_features = np.array([[250, 10, 10, 80]])
# Given item description
given_item_description = "Cereals, dry, Brown Sugar, DINOSAUR EGGS, Instant Oatmeal, QUAKER"
```



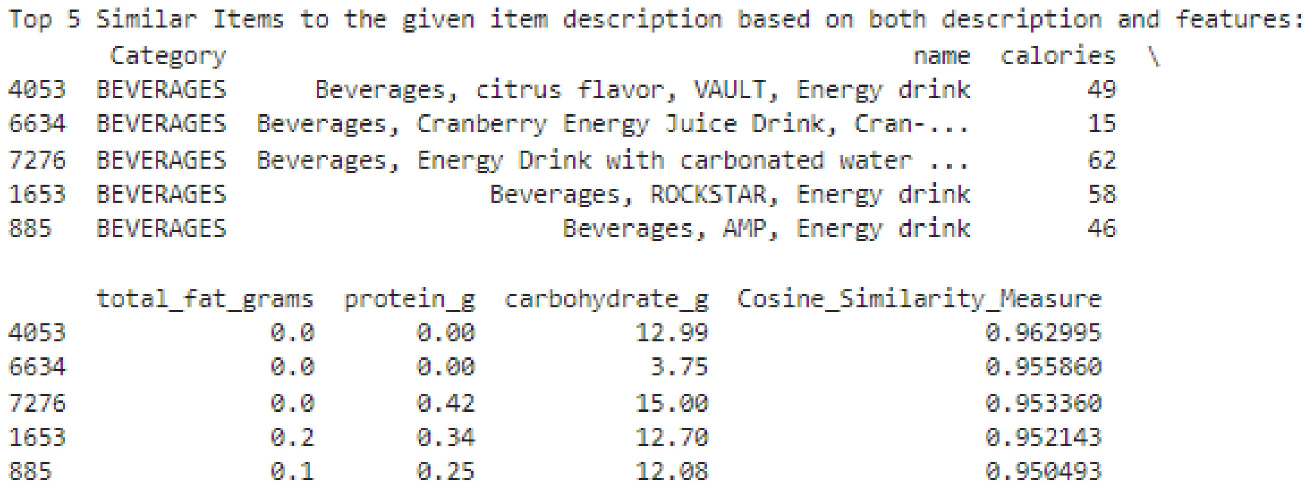
```
Top 5 Similar Items to the given item description based on both description and features:
     Category                                              name  calories  \
5877  CEREALS  Cereals, dry, maple and brown sugar, Instant O...       368
5452  CEREALS  Cereals, dry, fruit and cream variety, Instant...       379
6060  CEREALS  Cereals ready-to-eat, Maple Brown Sugar LIFE C...       373
5827  CEREALS  Cereals, dry, maple and brown sugar, fortified...       368
4789  CEREALS  Cereals, dry, apples and cinnamon, Instant Oat...       366

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure
5877              4.6       9.20           76.91                   0.974488
5452              6.4       8.30           75.42                   0.965101
6060              4.1       9.17           78.86                   0.959256
5827              4.7       9.25           76.67                   0.958728
4789              4.6       8.62           76.74                   0.956519
```

sample_features = np.array([[150, 5, 20, 5]]) # Example feature vector
given_item_description = "Beef, raw, select, separable lean and fat, chuck for stew"



```
Top 5 Similar Items to the given item description based on both description and features:
     Category                                              name  calories  \
4885     BEEF  Beef, raw, choice, separable lean and fat, chu...       130
5711     BEEF  Beef, raw, all grades, separable lean and fat,...       128
7610     BEEF  Beef, raw, select, trimmed to 0" fat, separabl...       122
6837     BEEF  Beef, raw, select, trimmed to 0" fat, separabl...       143
6968     BEEF  Beef, raw, choice, trimmed to 1/8" fat, separa...       139

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure
4885              4.8      21.64            0.12                   0.995010
5711              4.5      21.75            0.16                   0.989031
7610              4.2      21.05            0.00                   0.976133
6837              6.5      21.28            0.00                   0.976010
6968              5.1      21.96            0.00                   0.975979
```

sample_features = np.array([[50, 5, 5, 15]]) # Example feature vector
given_item_description = "Beverages, Citrus, Energy drink"



```
Top 5 Similar Items to the given item description based on both description and features:
       Category                                               name  calories  \
4053  BEVERAGES      Beverages, citrus flavor, VAULT, Energy drink        49
6634  BEVERAGES  Beverages, Cranberry Energy Juice Drink, Cran-...        15
7276  BEVERAGES  Beverages, Energy Drink with carbonated water ...        62
1653  BEVERAGES                Beverages, ROCKSTAR, Energy drink        58
885   BEVERAGES                     Beverages, AMP, Energy drink        46

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure
4053              0.0       0.00           12.99                   0.962995
6634              0.0       0.00            3.75                   0.955860
7276              0.0       0.42           15.00                   0.953360
1653              0.2       0.34           12.70                   0.952143
885               0.1       0.25           12.08                   0.950493
```

4.4 Final Model: Recommendation System Integration

In this section, we present the final model for our recommendation system, which seamlessly integrates textual descriptions, nutritional features, and GloVe embeddings to provide personalized item recommendations. The model architecture and functionality are outlined as follows:

Loading GloVe Embeddings:
We start by loading pre-trained GloVe embeddings of a specified dimension (e.g., 50 dimensions). These embeddings capture semantic similarities between words, essential for understanding textual descriptions.

Tokenization and Text Embedding:
Textual descriptions are tokenized, preprocessed, and converted into GloVe embeddings. Each word in the description contributes to the creation of an embedding vector representing the overall semantic meaning of the text.

Feature Concatenation:
The numerical features representing calories, total fat grams, protein grams, and carbohydrate grams are standardized. These standardized features are then concatenated with the GloVe embeddings of the textual descriptions and category labels, forming a comprehensive feature representation.

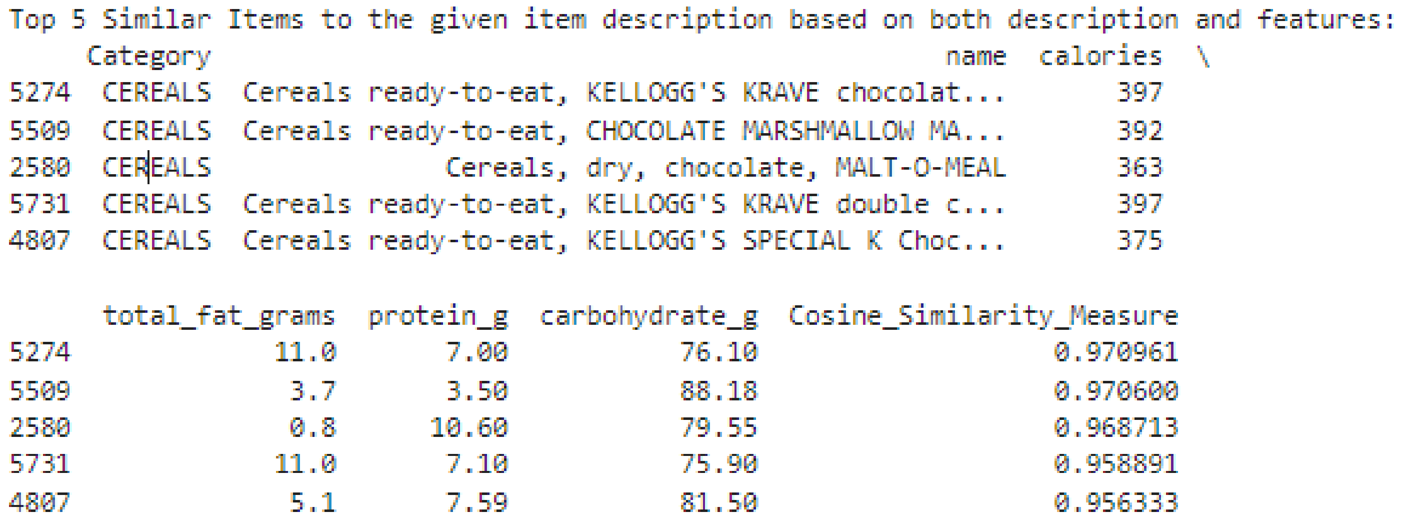
Cosine Similarity Calculation:
Cosine similarity is computed between the combined feature vectors of the given item description, category, and features, and all items in the dataset. This similarity measure quantifies the similarity in both textual descriptions, category labels, and nutritional features.

Recommending Similar Items:
Based on the calculated cosine similarity scores, the top N similar items to the given item description, category, and features are identified. These recommended items exhibit similarities in both textual descriptions, category labels, and nutritional compositions to the input features, providing users with highly relevant recommendations.

Results and Insights:
As an example, we demonstrate the functionality of the final recommendation model by inputting a sample item description ("cereals chocolate") and category ("CEREALS") along with sample features. The model retrieves the top 5 similar items based on both description and features, considering textual semantics, category information, and nutritional attributes. The cosine similarity measure provides a quantitative assessment of the similarity between the recommended items and the input description, category, and features, facilitating accurate and personalized recommendations for users.

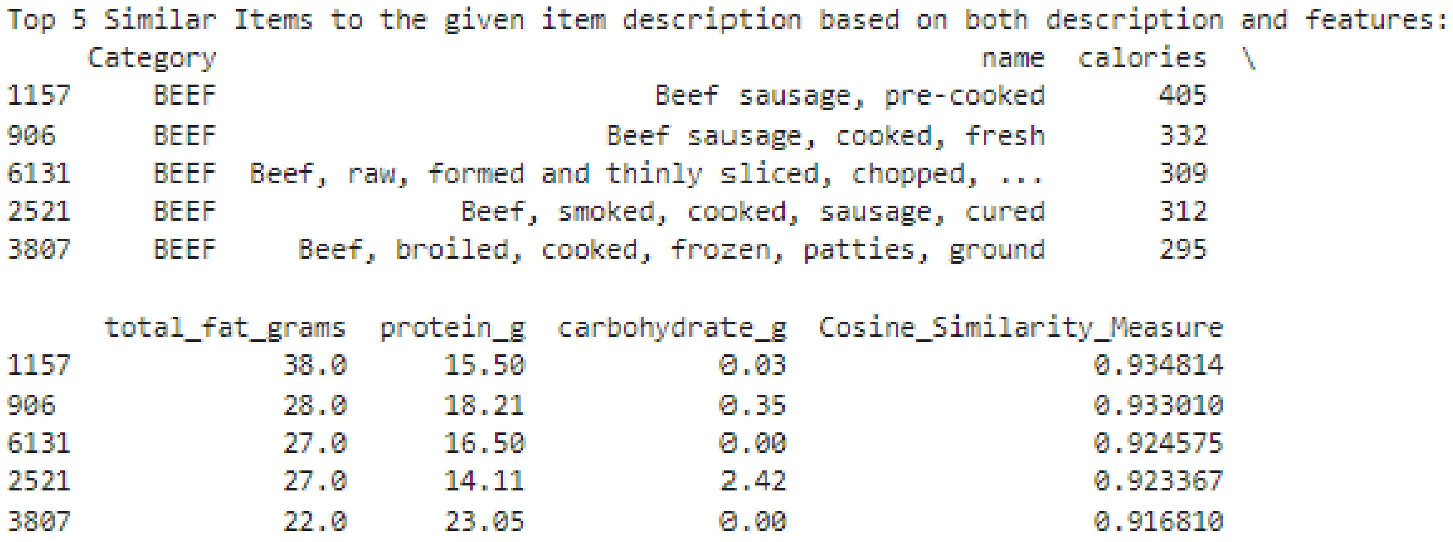
```
# Sample feature space including only 'calories', 'protein_g', 'total_fat_grams', 'carbohydrate_g'
sample_features = np.array([[250, 10, 10, 80]])  # Example feature vector
# Example: Input item description and category
given_item_description = "cereals chocolate"
```

given_item_category = "CEREALS"



```
 Top 5 Similar Items to the given item description based on both description and features:
     Category                                                name  calories  \
5274  CEREALS  Cereals ready-to-eat, KELLOGG'S KRAVE chocolat...       397   
5509  CEREALS  Cereals ready-to-eat, CHOCOLATE MARSHMALLOW MA...       392   
2580  CEREALS               Cereals, dry, chocolate, MALT-O-MEAL       363   
5731  CEREALS  Cereals ready-to-eat, KELLOGG'S KRAVE double c...       397   
4807  CEREALS  Cereals ready-to-eat, KELLOGG'S SPECIAL K Choc...       375   

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure  
5274             11.0       7.00           76.10                   0.970961  
5509              3.7       3.50           88.18                   0.970600  
2580              0.8      10.60           79.55                   0.968713  
5731             11.0       7.10           75.90                   0.958891  
4807              5.1       7.59           81.50                   0.956333  
```

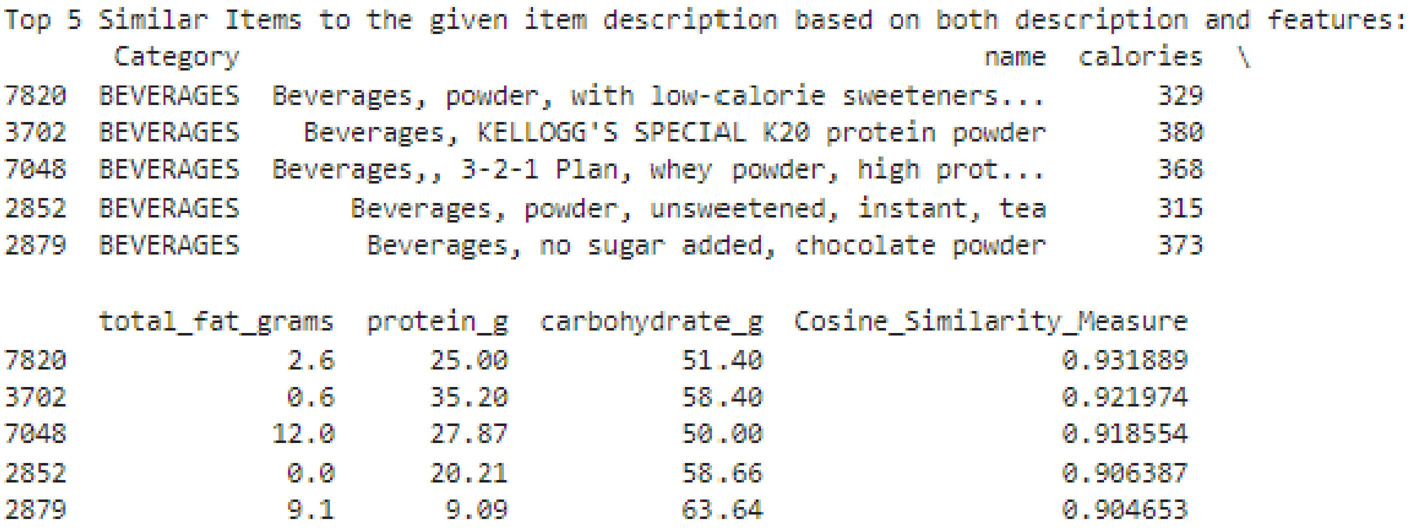
# Sample feature space including only 'calories', 'protein_g', 'total_fat_grams', 'carbohydrate_g'
sample_features = np.array([[350, 30, 20, 10]]) # Example feature vector
# Example: Input item description and category
given_item_description = "sausage grill"
given_item_category = "BEEF"



```
Top 5 Similar Items to the given item description based on both description and features:
     Category                                             name  calories  \
1157     BEEF                           Beef sausage, pre-cooked       405   
906      BEEF                        Beef sausage, cooked, fresh       332   
6131     BEEF  Beef, raw, formed and thinly sliced, chopped, ...       309   
2521     BEEF                 Beef, smoked, cooked, sausage, cured       312   
3807     BEEF     Beef, broiled, cooked, frozen, patties, ground       295   

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure  
1157             38.0      15.50            0.03                   0.934814  
906              28.0      18.21            0.35                   0.933010  
6131             27.0      16.50            0.00                   0.924575  
2521             27.0      14.11            2.42                   0.923367  
3807             22.0      23.05            0.00                   0.916810  
```

DUE LIMITED DATA AVAILABILITY WE CANNOT TEST THIS BETA MODEL TO ITS FULL POTENTIAL FOR EXAMPLE BELOW:

sample_features = np.array([[300, 5, 25, 50]]) # Example feature vector
# Example: Input item description and category
given_item_description = "strawberry bannana chocolate protein smoothie"
given_item_category = "BEVERAGES"



```
Top 5 Similar Items to the given item description based on both description and features:
       Category                                              name  calories  \
7820  BEVERAGES  Beverages, powder, with low-calorie sweeteners...       329   
3702  BEVERAGES    Beverages, KELLOGG'S SPECIAL K20 protein powder       380   
7048  BEVERAGES  Beverages,, 3-2-1 Plan, whey powder, high prot...       368   
2852  BEVERAGES       Beverages, powder, unsweetened, instant, tea       315   
2879  BEVERAGES        Beverages, no sugar added, chocolate powder       373   

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure  
7820              2.6      25.00           51.40                   0.931889  
3702              0.6      35.20           58.40                   0.921974  
7048             12.0      27.87           50.00                   0.918554  
2852              0.0      20.21           58.66                   0.906387  
2879              9.1       9.09           63.64                   0.904653  
```

Top 5 Similar Items to the given item description based on both description and features:

Item 7821
Category: BEVERAGES
Name: Beverages, powder, with low-calorie sweeteners, reduced calorie, chocolate, Dairy drink mix
Calories: 329

Item 3703
Category: BEVERAGES
Name: Beverages, KELLOGG'S SPECIAL K20 protein powder
Calories: 380

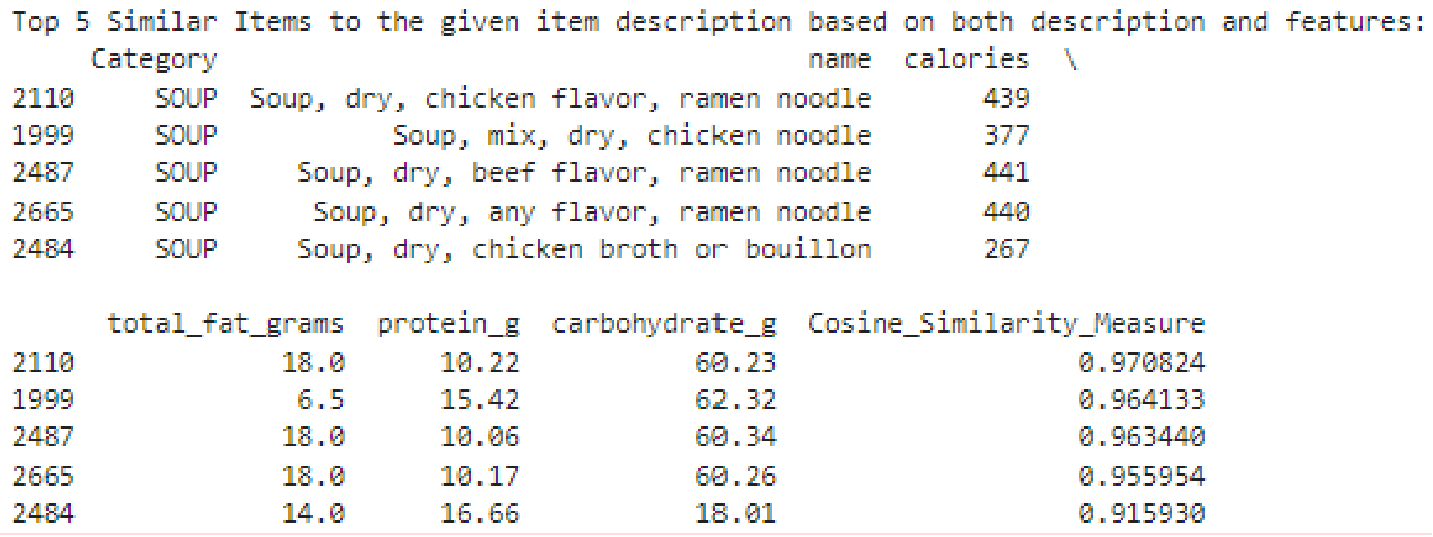
Item 7049
Category: BEVERAGES
Name: Beverages,, 3-2-1 Plan, whey powder, high protein, SLIMFAST Shake Mix, UNILEVER
Calories: 368

Item 2853
Category: BEVERAGES
Name: Beverages, powder, unsweetened, instant, tea
Calories: 315

Item 2880
Category: BEVERAGES
Name: Beverages, no sugar added, chocolate powder
Calories: 373

# Sample feature space including only 'calories', 'protein_g', 'total_fat_grams', 'carbohydrate_g'
sample_features = np.array([[450, 20, 20, 60]]) # Example feature vector
# Example: Input item description and category
given_item_description = "chicken noodle soup"

given_item_category = "SOUP"



```
Top 5 Similar Items to the given item description based on both description and features:
     Category                                    name  calories  \
2110     SOUP  Soup, dry, chicken flavor, ramen noodle       439
1999     SOUP           Soup, mix, dry, chicken noodle       377
2487     SOUP     Soup, dry, beef flavor, ramen noodle       441
2665     SOUP      Soup, dry, any flavor, ramen noodle       440
2484     SOUP     Soup, dry, chicken broth or bouillon       267

      total_fat_grams  protein_g  carbohydrate_g  Cosine_Similarity_Measure
2110             18.0      10.22           60.23                   0.970824
1999              6.5      15.42           62.32                   0.964133
2487             18.0      10.06           60.34                   0.963440
2665             18.0      10.17           60.26                   0.955954
2484             14.0      16.66           18.01                   0.915930
```

Top 5 Similar Items to the given item description based on both description and features:

Item 2111
Category: SOUP
Name: Soup, dry, chicken flavor, ramen noodle
Calories: 439

Item 2000
Category: SOUP
Name: Soup, mix, dry, chicken noodle
Calories: 377

Item 2488
Category: SOUP
Name: Soup, dry, beef flavor, ramen noodle
Calories: 441

Item 2666
Category: SOUP
Name: Soup, dry, any flavor, ramen noodle
Calories: 440

Item 2485
Category: SOUP
Name: Soup, dry, chicken broth or bouillon
Calories: 267

5. Conclusion and Future Prospects:

The Nutri-Recommender project has culminated in the creation of a sophisticated food recommendation system, underpinned by extensive data analysis, machine learning algorithms, and innovative feature engineering techniques. Our endeavor began with the meticulous exploration of a vast dataset containing nutritional information for thousands of food items. Through rigorous exploratory data analysis and feature selection, we unearthed invaluable insights into the dataset's composition, facilitating subsequent modeling endeavors.

Our journey through classification modeling yielded remarkable results, as we successfully predicted food categories with a high degree of accuracy. By iteratively refining our models and evaluating their performance, we achieved classification accuracies that underscore the efficacy of our approach. However, it is worth noting that the journey did not end here. Rather, it served as a springboard for the development of our flagship Nutri-Recommender system.

At the heart of our project lies the Nutri-Recommender system—a cutting-edge recommendation engine designed to provide users with personalized food recommendations tailored to their unique preferences and nutritional requirements. Leveraging advanced embedding techniques and content-based filtering algorithms, our recommendation system delivers curated recommendations that resonate with individual tastes and dietary goals.

Looking ahead, the future prospects for the Nutri-Recommender project are rife with possibilities. One avenue for exploration involves the development of an intuitive application interface to democratize access to personalized food recommendations. By harnessing user data such as dietary preferences, health objectives, and demographic information, the application can enhance recommendation accuracy and relevance, thus empowering users to make informed dietary choices.

Furthermore, user engagement and feedback mechanisms will play a pivotal role in the continuous improvement and refinement of our recommendation system. Through ongoing user testing and validation studies, we can gather invaluable insights into user satisfaction, usability, and the overall effectiveness of the application. Armed with this feedback, we can iteratively enhance the recommendation algorithms to better serve the needs of our users.

In summary, the Nutri-Recommender project represents a convergence of data science, machine learning, and nutrition science, with the aim of revolutionizing the way individuals approach dietary management. By harnessing the power of technology, we have developed a tool that empowers users to make informed food choices, thereby promoting healthier lifestyles and improved well-being. As we look to the future, we remain committed to pushing the boundaries of innovation and leveraging technology to create meaningful impact in the realm of personalized nutrition.

References:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6943843/
2. https://ieeexplore.ieee.org/document/7877463
3. https://dl.acm.org/doi/10.1145/1882992.1883053
4. https://dl.acm.org/doi/10.1145/2792838.2796554
5. https://www.jandonline.org/article/S0002-8223(07)00744-4/abstract
6. https://www.researchgate.net/publication/334529528_A_Food_Recommender_System_Considering_Nutritional_Information_and_User_Preferences
7. https://www.researchgate.net/publication/374418599_Food_Recommendation_Systems_Based_On_Content-based_and_Collaborative_Filtering_Techniques
8. https://www.researchgate.net/publication/313723698_Diet-Right_A_Smart_Food_Recommendation_System
9. https://arxiv.org/pdf/2306.16528
10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6581448/
11. https://pubmed.ncbi.nlm.nih.gov/19465193/
12. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8765311
13. https://sciencescholar.us/journal/index.php/ijhs/article/view/9031
14. https://www.neliti.com/publications/53646/relevant-words-extraction-method-for-recommendation-system
15. https://www.sciencedirect.com/science/article/pii/S0957417423026684
16. https://pubmed.ncbi.nlm.nih.gov/32039089/
17. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9775081