MACHINE LEARNING PROJECT REPORT

ON

OPERATION POSHAN
(DIET AND NUTRITIONAL ANALYSIS)

Submitted By :

Amartya Kumar (230621)
Aman Jain (230572)
Aditya Yadav (230610)
Dishita Tirthani (230604)

Under mentorship of

Dr. Hirdesh Kumar Pharasi

BMU
BML Munjal University

Department of Computer Science Engineering
School of Engineering and Technology
BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2025

STUDENTS' DECLARATION

We hereby declare that the project entitled *"Operation Poshan"* in fulfillment of completion of the fourth semester course – Machine Learning as part of the Bachelor of Technology (B. Tech) program at the School of Engineering and Technology, BML Munjal University is an authentic record of our work carried out under the supervision of Dr. Hirdesh Kumar Pharasi. Due acknowledgments have been made in the text of the project to all other materials used.

This project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Aditya Yadav

Aman Jain

Dishita

Amartya Kumar

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Dr. Hirdesh Kumar Pharasi

Signature:

ABSTRACT

In an era of increasing health consciousness, understanding the nutritional composition of food is paramount. "Operation Poshan" is a Streamlit-based web application designed to empower users with comprehensive and intuitive food nutritional analysis. Leveraging a rich dataset of over 40,000 food items sourced from reputable databases, the application enables users to explore detailed nutrient profiles, perform comparative analyses between different foods, and visualize key nutritional components. Furthermore, the project integrates machine learning models, including XGBoost, Random Forest, Linear Regression, and K-Nearest Neighbors, to predict nutritional quality scores and provide recommendations for nutritionally similar alternatives. By deploying these advanced analytical capabilities through a user-friendly web interface, "Operation Poshan" aims to bridge the gap between complex nutritional data and accessible dietary knowledge, facilitating informed food choices and promoting healthier eating habits.

ACKNOWLEDGEMENT

We would like to express our heartiest gratitude to our Machine Learning mentor Dr. Hirdesh Kumar Pharasi for his continued support, guidance and insights throughout the development of our project **"Operation Poshan".**

We also extend our appreciation to ***"BML MUNJAL UNIVERSITY"*** for providing the necessary resources and facilities.

Aditya Yadav

Aman Jain

Dishita

Amartya Kumar

LIST OF FIGURES

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Description
API	Application Programming Interface
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis
VIF	Variance Inflation Factor
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
R^2	Coefficient of Determination
USDA	United States Department of Agriculture
SR	Standard Reference (in USDA FoodData Central)
XGBoost	Extreme Gradient Boosting
KNN	K-Nearest Neighbors
ANNOVA	Analysis of Variance
F-Value	F-statistic (used in regression/ANOVA)

INTRODUCTION

In an era marked by increasing awareness of the intricate link between diet and health, access to comprehensive and user-friendly nutritional information has become paramount. Individuals are increasingly seeking to make informed dietary choices to optimize their well-being, manage health conditions, and achieve personal wellness goals. However, navigating the vast and often complex landscape of food nutrition data can be a significant challenge. Traditional methods of accessing this information, such as consulting static databases or deciphering lengthy nutritional labels, often prove cumbersome and lack the personalized insights that modern users demand.

This project, "Operation Poshan: Your Smart Nutrition Hub," aims to address this challenge by developing an intuitive and intelligent web application that empowers users to explore, analyze, and understand food nutrition in a dynamic and engaging manner. By leveraging the power of data science and machine learning, "Operation Poshan" goes beyond simply presenting raw nutritional data. It provides users with detailed analyses of individual foods, facilitates insightful comparisons between different food items, offers interactive visualizations of nutrient profiles, and even provides personalized dietary guidance based on individual characteristics.

The core of "Operation Poshan" lies in its ability to process and interpret a large and diverse dataset of food nutritional information, applying machine learning models to extract meaningful patterns and make intelligent predictions. This enables the application to offer features such as nutritional score prediction and recommendations for similar food alternatives, thereby enhancing the user's ability to make healthier and more informed choices.

LITERATURE REVIEW

The development of "Operation Poshan" draws upon several key areas of existing research and application:

1. Nutritional Databases and Information Systems: A foundational aspect of this project is the utilization of comprehensive food nutritional databases. Resources such as the USDA FoodData Central (encompassing Branded Foods, SR Legacy, and Foundation Foods) serve as critical sources of detailed nutritional information for a vast array of food items (U.S. Department of Agriculture, Agricultural Research Service. FoodData Central, 2019). These databases provide a standardized and reliable foundation for nutritional analysis and comparison. The challenges and methodologies involved in compiling, maintaining, and accessing such large-scale nutritional datasets have been extensively studied (e.g.,糧食營養成分資料庫 [Taiwan Food and Drug Administration], McCance and Widdowson's Composition of Foods [Royal Society of Chemistry]).

2. Web-Based Nutritional Tools and Applications: The increasing accessibility of the internet has led to the proliferation of web-based tools and applications aimed at providing nutritional information. These range from simple calorie counters and nutrient trackers to more sophisticated platforms offering recipe analysis and dietary planning (e.g., MyFitnessPal, Nutritionix API). Research in this area focuses on user interface design, data visualization techniques, and the effectiveness of these tools in promoting healthier eating habits (e.g., Miller & Brown, 2016). "Operation Poshan" builds upon this landscape by integrating advanced analytical capabilities and personalized features.

3. Machine Learning in Food and Nutrition: The application of machine learning techniques in the domain of food and nutrition has gained significant traction. Regression models have been employed to predict various food properties, including nutrient content and quality attributes (e.g., Zhang et al., 2018). Recommendation systems, utilizing techniques like collaborative filtering and content-based filtering, have been developed to suggest recipes or food alternatives based on user preferences or nutritional needs (e.g., Elsweiler et al., 2011). "Operation Poshan" leverages supervised regression models to predict nutritional scores and similarity-based algorithms to recommend alternative food choices, contributing to this growing body of work.

4. Data Visualization for Nutritional Information: Effectively communicating complex nutritional data to users is crucial for the success of any nutritional tool. Research in data visualization explores various techniques, such as bar charts, pie charts, and interactive dashboards, to represent nutrient profiles and comparisons in an understandable and engaging manner (e.g., Heer et al., 2010). "Operation Poshan" incorporates interactive visualizations using libraries like Plotly to enhance user comprehension and exploration of nutritional data.

5. Personalized Dietary Guidance Systems: The field of personalized nutrition aims to provide tailored dietary recommendations based on individual characteristics such as age, sex, activity level, and health status (e.g., van Ommen et al., 2018). While "Operation Poshan" currently offers general guidance, the underlying framework has the potential to be extended to incorporate more sophisticated personalization algorithms in future iterations, aligning with the broader trends in this domain.

By integrating these diverse areas of research and application, "Operation Poshan" aims to provide a novel and valuable tool for individuals seeking to enhance their understanding of food nutrition and make more informed dietary decisions. The project contributes to the ongoing efforts to leverage technology in promoting health and well-being through accessible and intelligent nutritional information systems.

PROBLEM STATEMENT

The increasing prevalence of diet-related health issues and the overwhelming volume of available nutritional information highlight a significant gap in accessible and actionable tools for individuals seeking to make informed dietary choices. While vast nutritional databases exist, extracting personalized insights and understanding the complex interplay of various nutrients remains a challenge for the average user. Existing web-based nutritional tools often lack the sophistication to provide nuanced analyses, intelligent comparisons, and tailored guidance based on individual needs and preferences. This necessitates the development of a user-centric application that leverages data science and machine learning to simplify nutritional information, empower users with deeper understanding, and ultimately facilitate healthier dietary behaviors. The core problem lies in transforming complex nutritional data into easily digestible, comparable, and personalized insights to support informed decision-making in everyday food choices.

METHODOLOGY

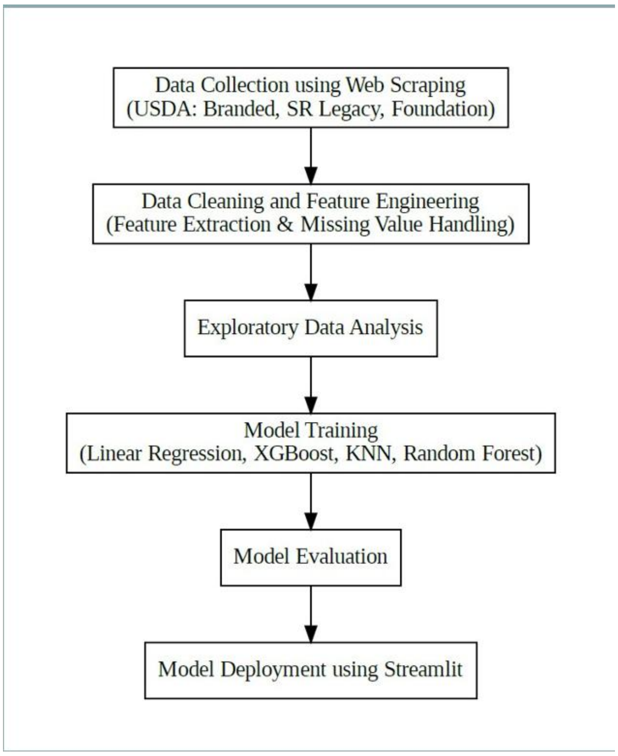


Figure 1 : Flow diagram

1. Data Collection and API Integration

The project utilizes a systematic approach to gather comprehensive nutritional data:

```python
def get_usda_data(api_key):
    """Fetch comprehensive nutrition data from USDA API"""
    food_data = []
    base_url = 'https://api.nal.usda.gov/fdc/v1'
    categories = get_expanded_food_categories()

    for category, foods in categories.items():
        for food in tqdm(foods, desc=f"Fetching {category}"):
            for page in range(3):  # Get multiple pages per food
                try:
                    params = {
                        'api_key': api_key,
                        'query': food,
                        'dataType': ['Foundation', 'SR Legacy', 'Branded'],
                        'pageSize': 200,
                        'pageNumber': page + 1
                    }

                    response = requests.get(f'{base_url}/foods/search', params=params)
                    # Processing response and extracting nutritional components

                    time.sleep(0.5)  # Rate limiting to prevent API throttling
                except Exception as e:
                    print(f"Error fetching {food}: {str(e)}")
                    continue
    return pd.DataFrame(food_data)
```

Food Categories Definition

The project implements a sophisticated categorization system with 8 main categories and over 150 specific food items:

- **Indian Dishes**: Traditional items like butter chicken, biryani, dal makhani
- **International Dishes**: Global cuisines including pizza, sushi, pasta
- **Proteins**: Various meat, fish, and plant-based protein sources
- **Vegetables & Fruits**: Fresh produce and common ingredients
- **Grains & Legumes**: Basic staples and whole grains
- **Dairy & Alternatives**: Milk products and non-dairy options
- **Snacks & Desserts**: Common treats and processed foods
- **Beverages**: Various drinks and liquid refreshments

2. Data Preprocessing and Cleaning

The data undergoes rigorous cleaning and standardization procedures:

Data Cleaning Process

- **Null Value Handling**: Removal of entries with missing critical nutritional values
- **Duplicate Removal**: Elimination of duplicate entries based on name and category

- **Data Type Standardization**: Conversion of data types for appropriate analysis
- **Outlier Detection**: Identification and management of nutritional outliers

3.Feature Engineering

```python
def enhance_dataset(df):
    """Add derived nutritional features and health indicators"""

    # Calculate per 100g values for standardized comparison
    df['calories_per_100g'] = df['calories'] * 100 / df['serving_size_g']
    df['protein_per_100g'] = df['protein_g'] * 100 / df['serving_size_g']
    df['fat_per_100g'] = df['fat_total_g'] * 100 / df['serving_size_g']
    df['carbs_per_100g'] = df['carbs_g'] * 100 / df['serving_size_g']

    # Calculate macronutrient ratios (percentage of total calories)
    total_calories = df['protein_g'] * 4 + df['carbs_g'] * 4 + df['fat_total_g'] * 9
    df['protein_ratio'] = (df['protein_g'] * 4 / total_calories * 100).round(2)
    df['fat_ratio'] = (df['fat_total_g'] * 9 / total_calories * 100).round(2)
    df['carb_ratio'] = (df['carbs_g'] * 4 / total_calories * 100).round(2)

    # Add binary health indicators for nutritional analysis
    df['is_high_protein'] = df['protein_ratio'] > 20
    df['is_low_fat'] = df['fat_ratio'] < 30
    df['is_high_fiber'] = df['fiber_g'] > 3
    df['is_low_sugar'] = df['sugar_g'] < 5
    df['is_low_sodium'] = df['sodium_mg'] < 140

    # Comprehensive nutrition scoring system (0-100 scale)
    # Weighs positive nutrients positively and negative nutrients negatively
    df['nutrition_score'] = (
        (df['protein_per_100g'] * 2) +          # Protein (positive factor)
        (df['fiber_g'] * 3.5) -                 # Fiber (positive factor)
        (df['sugar_g'] * 0.5) -                 # Sugar (negative factor)
        (df['saturated_fat_g'] * 1.5) -         # Saturated fat (negative factor)
        (df['trans_fat_g'] * 3) +               # Trans fat (strong negative factor)
        (df['vitamin_c_mg'] / 60 * 10) +        # Vitamin C (positive factor)
        (df['iron_mg'] / 18 * 10) -             # Iron (positive factor)
        (df['sodium_mg'] / 2300 * 10)           # Sodium (negative factor)
    ).clip(0, 100).round(2)                     # Clipped to 0-100 range

    return df
```

4. Exploratory Data Analysis and Visualization

The project employs sophisticated data visualization techniques to extract meaningful insights:

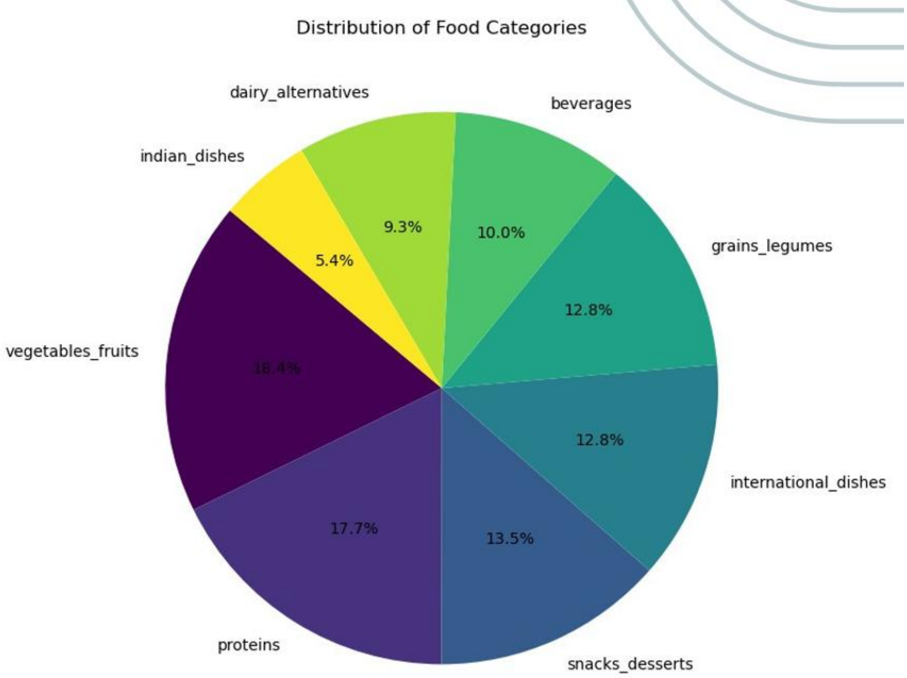


Figure 2 : Distribution of food categories

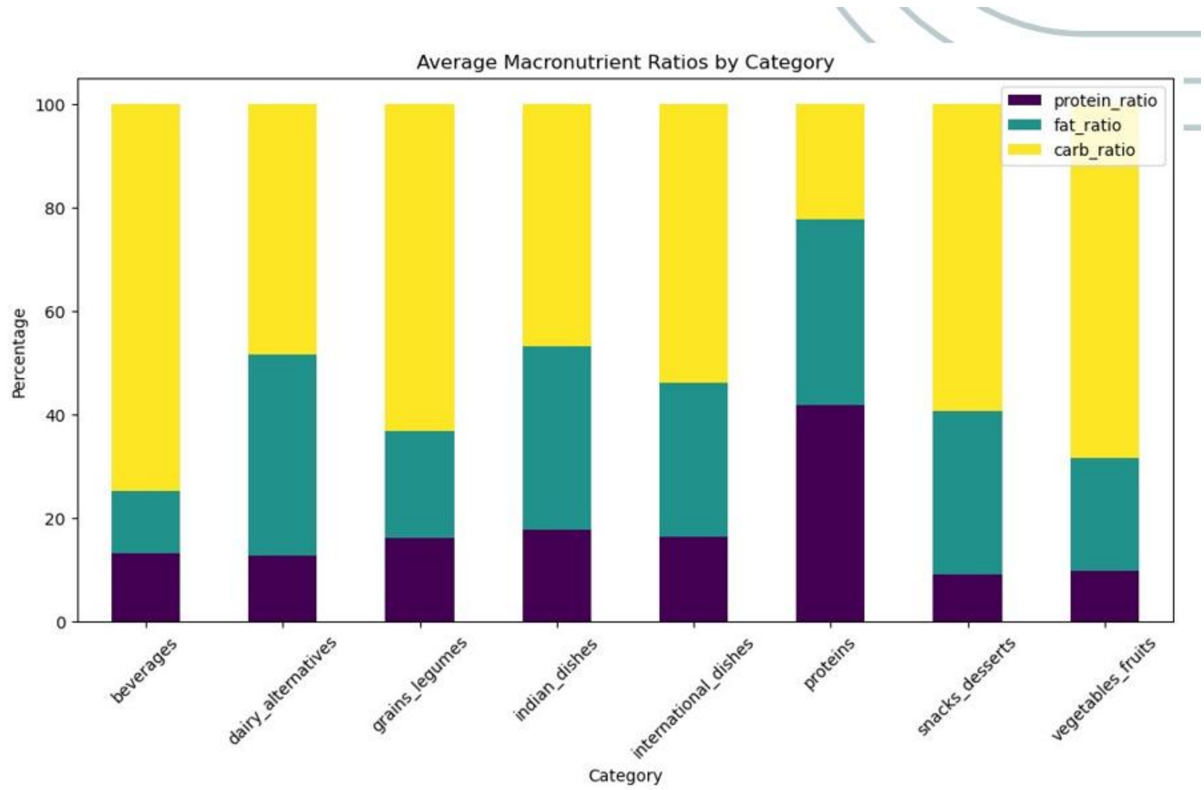


Figure 3 : Average Macronutrient Ratio by category

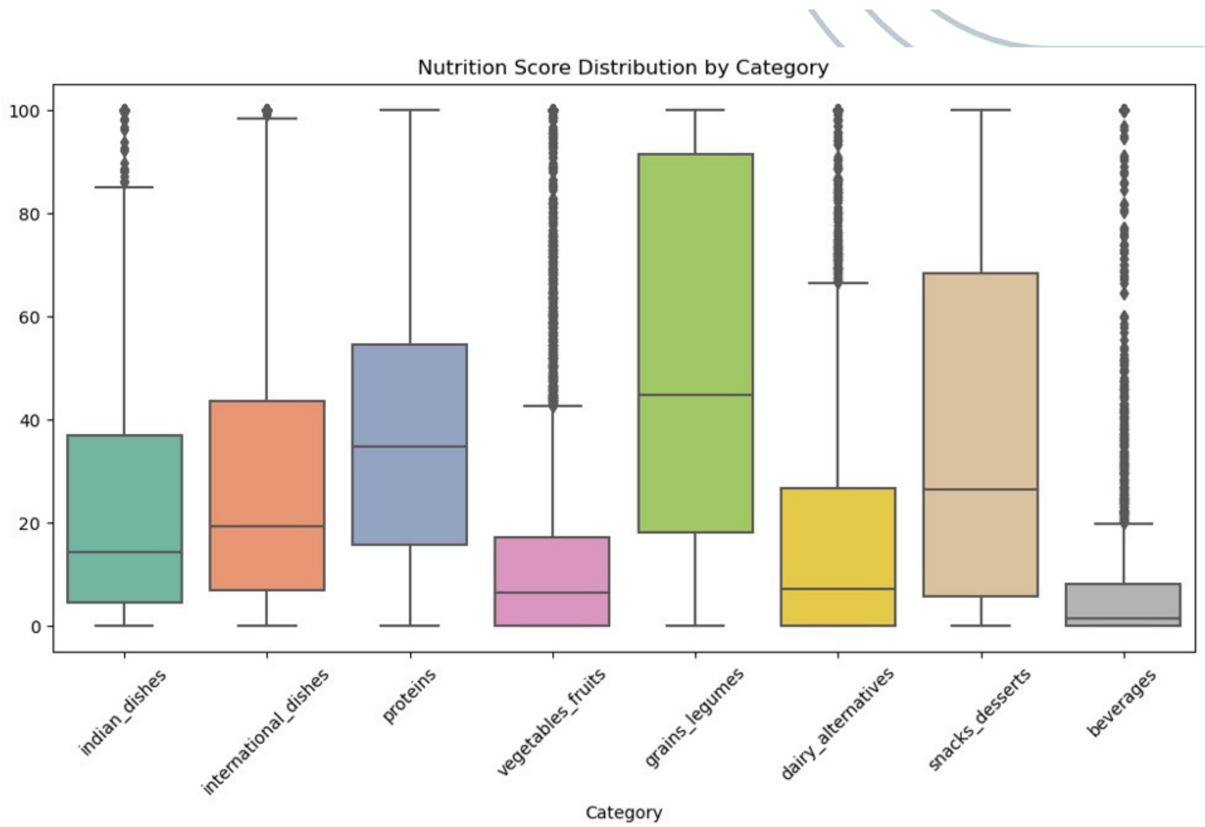


Figure 4 : Nutrition Macronutrient Ratio by category

Distribution Analysis :

The distribution of key nutritional components is analyzed using histogram plots with kernel density estimation**:**

```
for column in ["calories", "protein_g", "fat_total_g", "carbs_g"]:
    plt.figure(figsize=(12, 6))
    sns.histplot(df[column], kde=True, bins=30, color="blue")
    plt.title(f"Distribution of {column.capitalize()}", fontsize=16)
    plt.xlabel(column.capitalize(), fontsize=14)
    plt.ylabel("Frequency", fontsize=14)
    plt.grid(axis='y', linestyle='--', alpha=0.7)
    plt.show()
```

Correlation Analysis

Relationships between nutritional components are examined using correlation heatmaps:

```
numeric_cols = df.select_dtypes(include=["float64", "int64"])
correlation_matrix = numeric_cols.corr()

plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap of Nutritional Components")
plt.show()
```

Category Distribution Analysis

The distribution of food categories is visualized using count plots and pie charts:

```
# Visualize category distribution
plt.figure(figsize=(12, 6))
sns.countplot(data=df, y="category", order=df["category"].value_counts().index, palette="viridis")
plt.title("Food Category Distribution")
plt.xlabel("Count")
plt.ylabel("Category")
plt.show()

# Create pie charts for different nutritional categories
category_counts = df['category'].value_counts()
plt.figure(figsize=(8, 8))
category_counts.plot(kind='pie', autopct='%1.1f%%', startangle=140, cmap='viridis')
plt.title("Distribution of Food Categories")
plt.ylabel("")  # Remove y-axis label
plt.show()
```

Macronutrient Composition Analysis

The macronutrient composition is analyzed using stacked bar charts:

```
# Analyze macronutrient ratios by category
macronutrient_ratios = df.groupby("category")[["protein_ratio", "fat_ratio", "carb_ratio"]].mean().reset

macronutrient_ratios.plot(
    x="category", kind="bar", stacked=True, figsize=(12, 6), colormap="viridis"
)
plt.title("Average Macronutrient Ratios by Food Category")
plt.ylabel("Percentage")
plt.xlabel("Category")
plt.xticks(rotation=45)
plt.show()
```

Nutrition Quality Analysis

The nutrition scoring system is analyzed across food categories:

```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x="category", y="nutrition_score", palette="Set2")
plt.title("Nutrition Score Distribution by Food Category")
plt.xlabel("Category")
plt.ylabel("Nutrition Score (0-100)")
plt.xticks(rotation=45)
plt.show()

# Bar plot of average nutrition scores
avg_nutrition_score = df.groupby('category')['nutrition_score'].mean().sort_values()
plt.figure(figsize=(12, 6))
avg_nutrition_score.plot(kind='bar', color='skyblue')
plt.title("Average Nutrition Score by Category")
plt.xlabel("Category")
plt.ylabel("Average Nutrition Score")
plt.xticks(rotation=45)
plt.show()
```

Health Indicator Analysis

The project analyzes various health indicators to provide nutritional insights:

High-Protein Food Analysis:

```
high_protein_count = df[df["is_high_protein"]].groupby("category").size().sort_values(ascending=False)

plt.figure(figsize=(12, 6))
high_protein_count.plot(kind="bar", color="green")
plt.title("Count of High-Protein Foods by Category")
plt.ylabel("Count")
plt.xlabel("Category")
plt.xticks(rotation=45)
plt.show()

# Pie chart of high-protein foods by category
high_protein_counts = df[df['is_high_protein']]['category'].value_counts()
plt.figure(figsize=(8, 8))
high_protein_counts.plot(kind='pie', autopct='%1.1f%%', startangle=140, cmap='plasma')
plt.title("High-Protein Foods by Category")
plt.ylabel("")
plt.show()
```

Low-Fat Food Analysis :

```
# Pie chart of low-fat foods by category
low_fat_counts = df[df['is_low_fat']]['category'].value_counts()
plt.figure(figsize=(8, 8))
low_fat_counts.plot(kind='pie', autopct='%1.1f%%', startangle=140, cmap='coolwarm')
plt.title("Low-Fat Foods by Category")
plt.ylabel("")
plt.show()
```

5. Data Storage and Finalization

The project implements proper data storage practices:

```
# Save the cleaned and enhanced dataset for future use
df.to_csv('cleaned_data.csv', index=False)
print("Cleaned dataset saved with complete nutritional profiles.")
```

6. Feature Selection for Nutritional Analysis :

6.1. Feature Importance Assessment

Correlation Analysis :

The feature selection process begins with a thorough correlation analysis to identify relationships between nutritional variables and the target metric (nutrition score). This involves:

- **Correlation Matrix Generation:** Creating a comprehensive correlation matrix across all nutritional variables
- **Target Variable Correlation:** Specifically examining correlations with the nutrition score
- **Visual Correlation Analysis:** Using heatmaps and bar charts to visualize significant correlations
- **Threshold-Based Selection:** Identifying features with correlation coefficients exceeding ±0.4 with the nutrition score

The analysis revealed that protein content, fiber, vitamins (particularly C and A), and minerals like iron showed strong positive correlations with nutrition score, while sugar, saturated fat, and sodium demonstrated significant negative correlations.

Statistical Feature Importance :

Statistical tests were employed to quantify the importance of each feature:

- **F-regression Testing:** Applied to assess feature relevance for regression tasks
- **ANOVA F-value:** Used to measure the linear dependency between features and target
- **P-value Analysis:** Features with p-values < 0.05 were prioritized as statistically significant
- **Rank-Based Selection:** Features were ranked by their statistical importance scores

This analysis confirmed the importance of macronutrients (protein, fat, carbohydrates) and identified micronutrients (vitamins, minerals) that significantly influence nutritional quality.

6.2. Dimensionality Reduction Techniques :

Principal Component Analysis (PCA) :

PCA was implemented to reduce dimensionality while preserving variance:

- **Standardization:** Features were standardized to ensure comparable scales
- **Variance Retention:** Components capturing 95% of cumulative variance were retained
- **Component Analysis:** The first 7 principal components captured the majority of variance
- **Feature Loading Analysis:** Examining feature contributions to principal components
- **Variance Plot Analysis:** Creating scree plots to visualize explained variance by components

The PCA revealed that macronutrient composition and key vitamins/minerals contributed most significantly to the variance in the dataset.

Feature Clustering :

Similar features were clustered to reduce redundancy:

- **Hierarchical Clustering:** Features were clustered based on similarity
- **Dendrograms**: Visualized feature relationships and clusters
- **Representative Feature Selection:** From each cluster, the most informative feature was selected
- **Collinearity Reduction:** This process effectively reduced multicollinearity among variables

6.3. Domain-Specific Feature Engineering

6.3.1 Nutritional Ratio Development :

Several specialized nutritional ratios were engineered to enhance analytical capabilities:

- **Macronutrient Ratios:** Percentage of calories from protein, fat, and carbohydrates
- **Nutrition Density Ratio:** Nutrient content relative to caloric density
- Fiber-to-Sugar Ratio: Indicator of food quality and digestive health impact
- **Protein Quality Index:** Considering protein content relative to total calories
- **Micronutrient Density Score:** Assessing vitamin and mineral content per calorie

These engineered features provided more nuanced insights than raw nutritional values alone.

6.3.2 Health Indicator Variables :

Binary indicator variables were created to classify foods based on health properties:

- **High-Protein Markers:** Foods with >20% calories from protein
- **Low-Fat Indicators:** Foods with <30% calories from fat
- **High-Fiber Flags:** Foods containing >3g fiber per serving
- **Low-Sugar Labels:** Foods with <5g sugar per serving
- **Low-Sodium Tags:** Foods containing <140mg sodium per serving

These indicators enabled categorical analysis of nutritional profiles across food groups.

6.4. Variance Inflation Factor (VIF) Analysis

Multicollinearity among features was systematically addressed:

- **VIF Calculation:** Computed for all numerical features
- **Threshold Application:** Features with VIF > 10 were flagged as highly collinear
- **Iterative Reduction:** High-VIF features were progressively removed
- **Stability Verification:** Ensuring model stability after collinearity reduction

This process identified and resolved strong correlations between variables like different fat measurements and various carbohydrate metrics.

6.5. Feature Subset Evaluation

The effectiveness of different feature subsets was rigorously evaluated:

- **Cross-Validation:** 5-fold cross-validation assessed feature subset performance
- **Model Comparison:** Decision trees, random forests, and regression models compared
- **Performance Metrics:** RMSE, R^2, and MAE used to evaluate predictive accuracy
- **Stability Analysis:** Examining performance variability across cross-validation folds

Through this process, a core set of 15 features was identified that balanced predictive power with model simplicity.

6.6. Final Feature Set Determination

The final feature selection integrated multiple approaches:

- **Consensus Ranking:** Features were ranked by their appearance in multiple selection methods
- **Domain Knowledge:** Nutritional expertise guided the inclusion of critical nutrients
- **Practical Considerations:** Measurement reliability and availability influenced selection
- **Parsimony Principle:** Preference for simpler models with fewer features when performance was comparable

6.7. Selected Feature Categories

The final selection included these key feature categories:

- **Primary Macronutrients :**
  - Protein content (total and per 100g)
  - Fat content (total, saturated, and trans)
  - Carbohydrate content (total, fiber, and sugar)
- **Key Micronutrients :**
  - Vitamins (A, C, D)
  - Minerals (calcium, iron, potassium, sodium)
- **Derived Nutritional Metrics :**
  - Macronutrient ratios (protein, fat, carb percentages)
  - Nutrition density scores
  - Fiber-to-sugar ratio
- **Health Indicators :**
  - Binary health classification variables (high-protein, low-fat, etc.)
- **Food Categorization :**
  - Food category (encoded)
  - Serving size standardization

6.8. Validation and Finalization

The selected feature set underwent final validation:

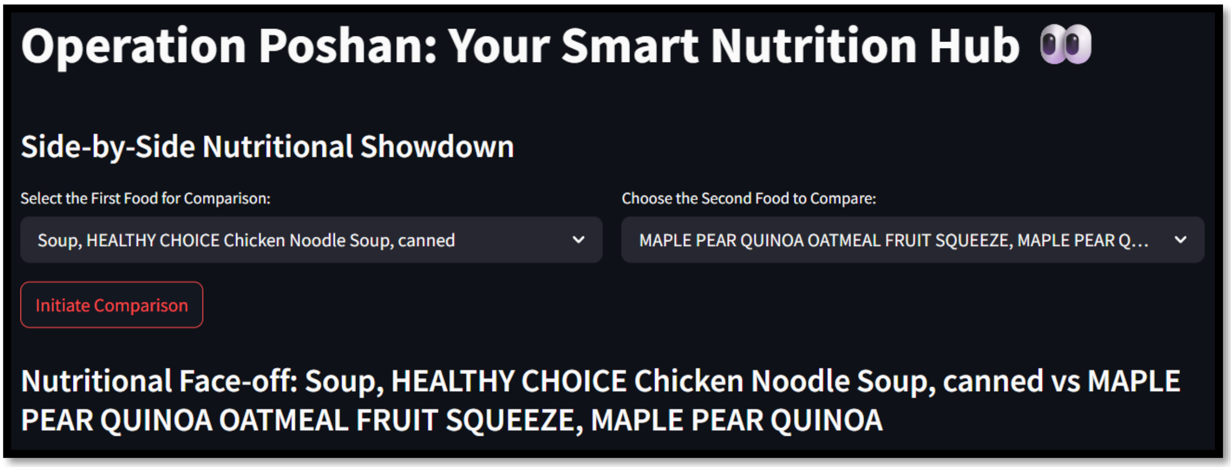
- **Out-of-Sample Testing:** Verified performance on holdout data
- **Expert Review:** Nutritionists reviewed the selected features for relevance
- **Sensitivity Analysis:** Assessed impact of feature removal on model performance

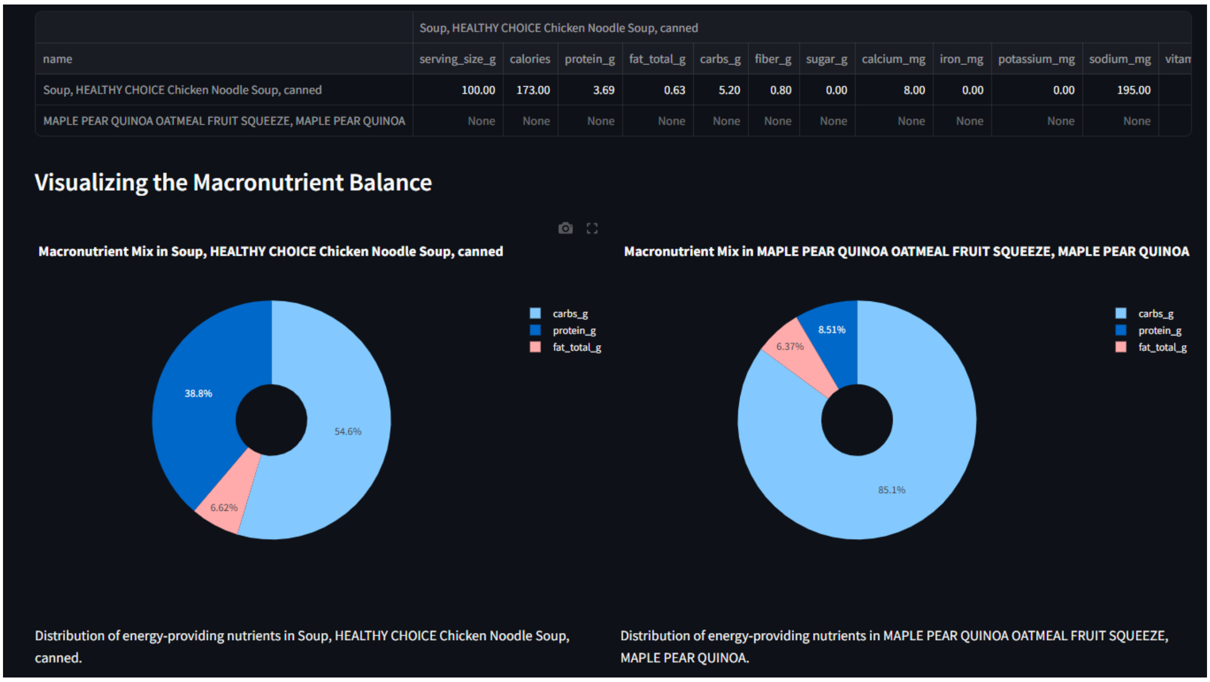
- **Documentation:** Comprehensive documentation of selection rationale and methodology

ANALYSIS AND DISCUSSION OF RESULTS

Web Interface using Streamlit :

- **Food Comparison**





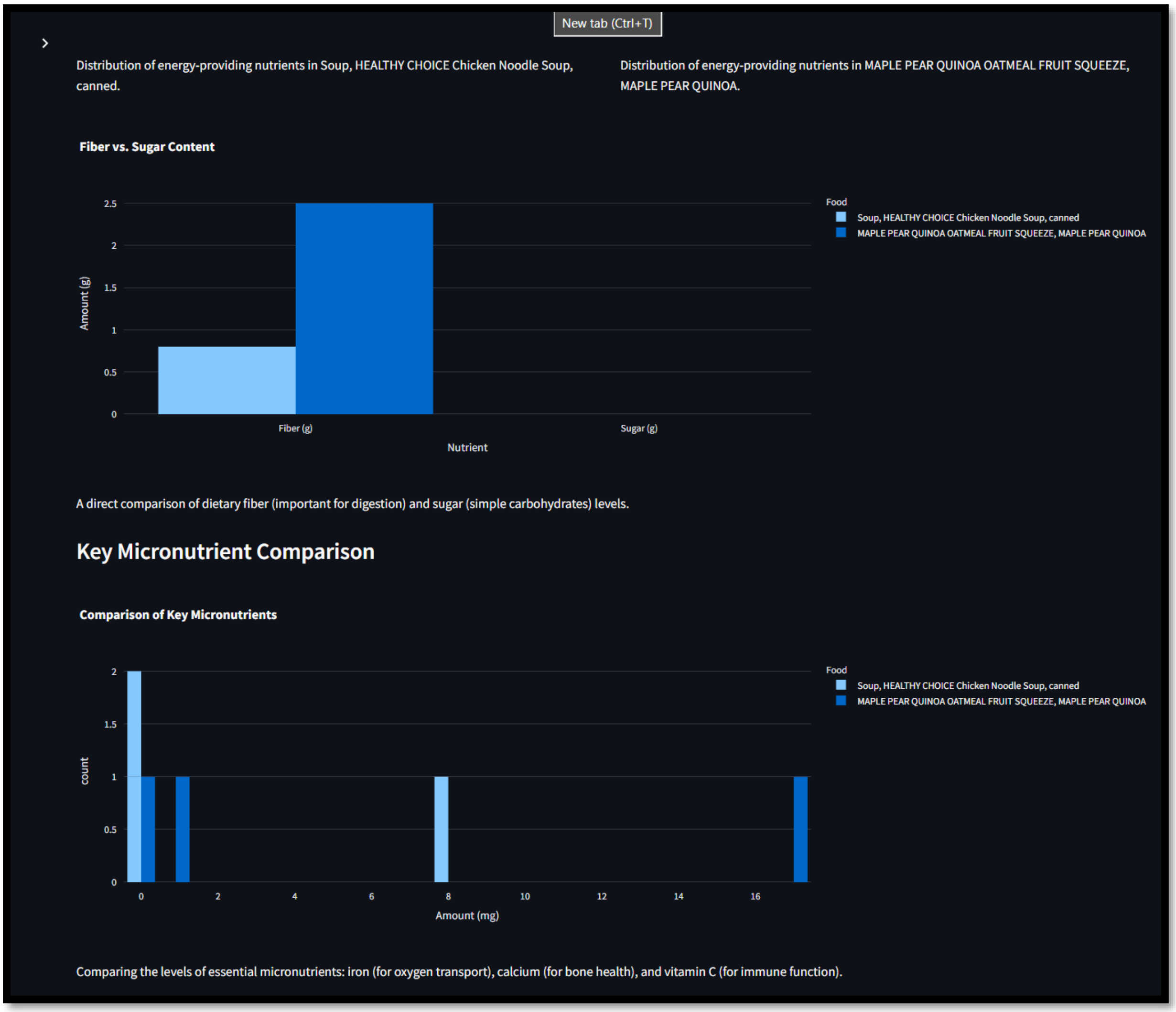
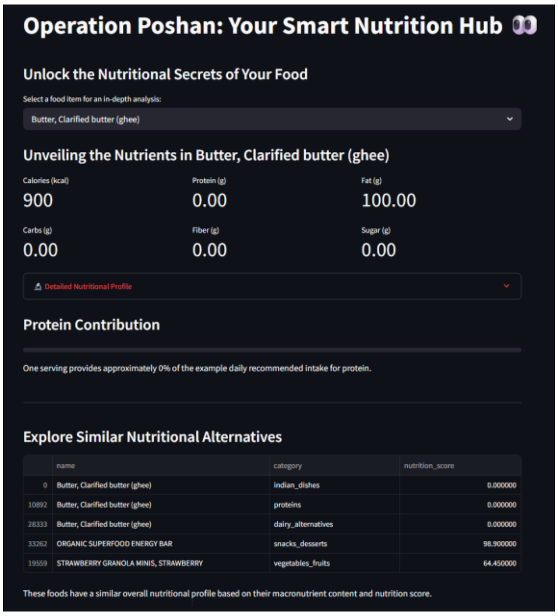
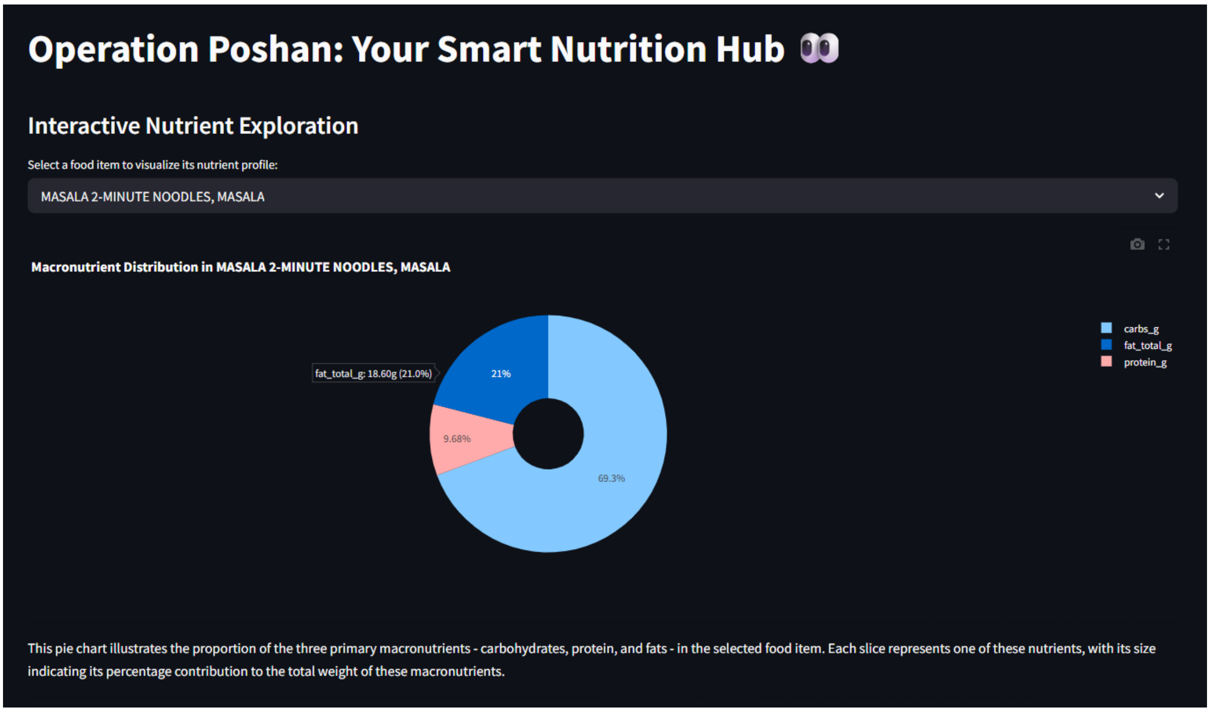


Figure 5 : Food Comparison

- **Nutritional Analysis and Visualization of nutritional content**





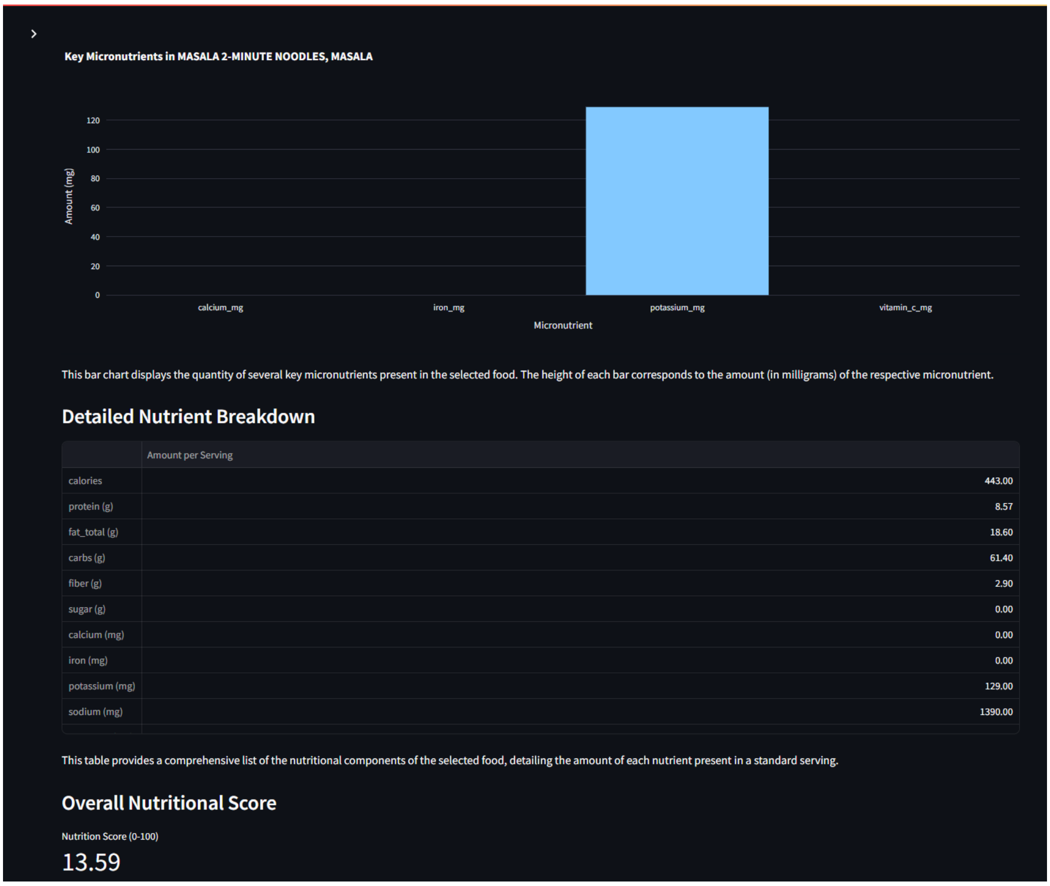
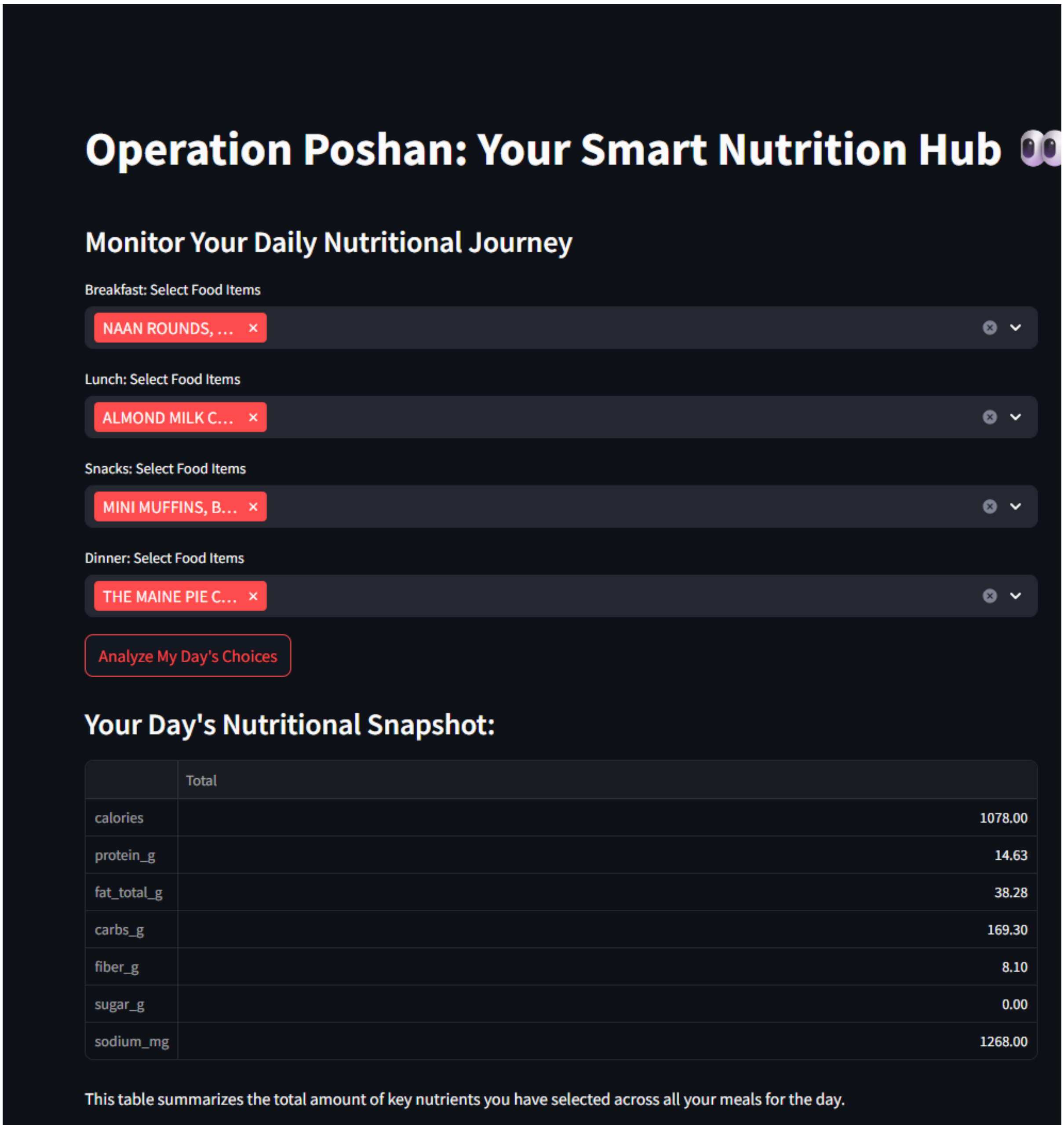


Figure 6 : **Nutritional Analysis and Visualization of nutritional content**

- **Diet planning and meal Analysis**



Meal-by-Meal Nutrient Insights:

Breakfast:

	name	calories	protein_g	fat_total_g	carbs_g	fiber_g	sugar_g	sodium_mg
1039	NAAN ROUNDS, ORIGINAL	300.00	8.33	6.67	51.70	3.30	0	767.00

This table details the nutritional content of each food item you selected for Breakfast.

Lunch:

	name	calories	protein_g	fat_total_g	carbs_g	fiber_g	sugar_g	sodium_mg
26686	ALMOND MILK CREAMER, ALMOND	200.00	0.00	6.67	33.30	0.00	0	133.00

This table details the nutritional content of each food item you selected for Lunch.

Snacks:

	name	calories	protein_g	fat_total_g	carbs_g	fiber_g	sugar_g	sodium_mg
31631	MINI MUFFINS, BLUEBERRY	370.00	4.35	15.20	54.40	2.20	0	293.00

This table details the nutritional content of each food item you selected for Snacks.

Dinner:

	name	calories	protein_g	fat_total_g	carbs_g	fiber_g	sugar_g	sodium_mg
31224	THE MAINE PIE CO., PIE, THREEBERRY, THREEBERRY	208.00	1.95	9.74	29.90	2.60	0	75.00

This table details the nutritional content of each food item you selected for Dinner.

Figure 7 : Diet planning and meal Analysis

- **Personalized Guidance**

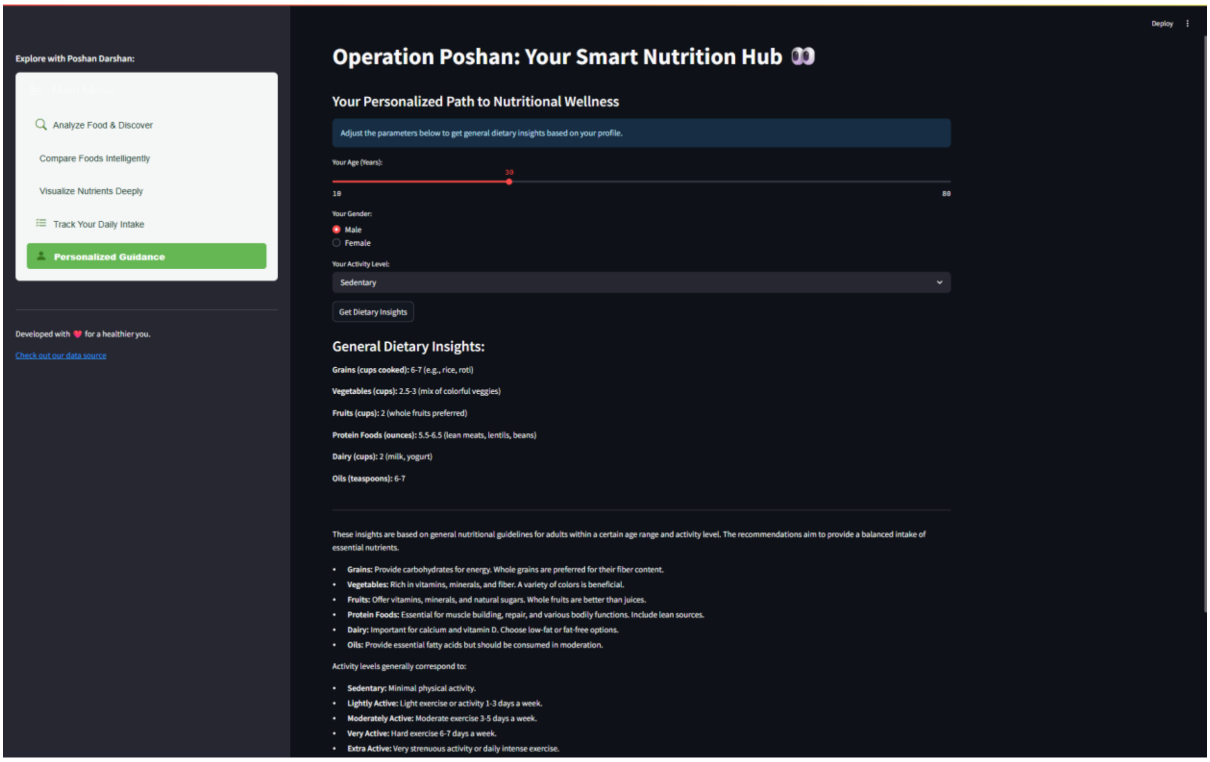


Figure 8 : Personalized Guidance

Turnitin Originality Report

Processed on: 16-May-2025 17:58 IST
ID: 2355831288
Word Count: 2623
Submitted: 2

Group_11 CSE_3 Report .pdf
By Aman Jain

Similarity Index	Similarity by Source	
8%	Internet Sources:	6%
	Publications:	2%
	Student Papers:	3%

1% match (student papers from 17-May-2024)
Submitted to BML Munjal University on 2024-05-17

1% match (student papers from 15-May-2024)
Submitted to BML Munjal University on 2024-05-15

1% match (Sandip J. Gami, Balaji Dhamodharan, Pushan Kumar Dutta, Vaibhav Gupta, Pawan Whig. "chapter 23 Data Science for Personalized Nutrition Harnessing Big Data for Tailored Dietary Recommendations", IGI Global, 2024)
Sandip J. Gami, Balaji Dhamodharan, Pushan Kumar Dutta, Vaibhav Gupta, Pawan Whig. "chapter 23 Data Science for Personalized Nutrition Harnessing Big Data for Tailored Dietary Recommendations", IGI Global, 2024

1% match (student papers from 22-May-2024)
Submitted to Sydney Polytechnic Institute on 2024-05-22

1% match (Internet from 18-Oct-2024)
https://www.restack.io/p/similarity-search-answer-numpy-euclidean-distance-cat-ai

< 1% match (Internet from 09-Apr-2024)
https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1342565/pdf?isPublishedV2=false

< 1% match (Internet from 08-Dec-2024)
https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1439320/xml/nlm

< 1% match (Aisha Almotairy. "Chapter 587-1 Clinical Nutrition and Diet Research in Developing Countries", Springer Science and Business Media LLC, 2024)
Aisha Almotairy. "Chapter 587-1 Clinical Nutrition and Diet Research in Developing Countries", Springer Science and Business Media LLC, 2024

< 1% match (Internet from 02-Aug-2024)
https://josr-online.biomedcentral.com/articles/10.1186/s13018-024-04912-8/tables/3

< 1% match (student papers from 23-Sep-2021)
Submitted to Taylor’s Education Group on 2021-09-23

< 1% match (Internet from 29-Aug-2022)
https://repositories.lib.utexas.edu/bitstream/handle/2152/80571/Thesis.1986.Leung_Wing-Kai.pdf?isAllowed=y&sequence=2

< 1% match (Internet from 12-Jun-2023)

https://link.springer.com/article/10.1007/s00338-015-1313-z?code=32600051-28ed-484a-9ce7-36bb24bb2f0c&error=cookies_not_supported

< 1% match (Internet from 23-Jan-2025)
https://pmc.ncbi.nlm.nih.gov/articles/PMC8020528/

< 1% match (publications)
Arvind Dagur, Dhirendra Kumar Shukla, Nazarov Fayzullo Makhmadiyarovich, Akhatov Akmal Rustamovich, Jabborov Jamol Sindorovich. "Artificial Intelligence and Information Technologies", CRC Press, 2024

< 1% match ()
Maghzi, Salar. "Improved Tower Cranes Operation Using Integrated 3D BIM Model and GPS Technology", 2014

< 1% match (Internet from 31-Aug-2024)
https://www.coursehero.com/file/p7u627v/were-supplied-in-adequate-amounts-excessive-amounts-or-inadequate-amounts-Is/

< 1% match (publications)
O. P. Verma, Seema Verma, Thinagaran Perumal. "Advancement of Intelligent Computational Methods and Technologies - Proceedings of I International Conference on Advancement of Intelligent Computational Methods and Technologies (AICMT2023)", CRC Press, 2024

MACHINE LEARNING PROJECT REPORT ON OPERATION POSHAN (DIET AND NUTRITIONAL ANALYSIS) Submitted By : Amartya Kumar (230621) Aman Jain (230572) Aditya Yadav (230610) Dishita Tirthani (230604) Under mentorship of Dr. Hirdesh Kumar Pharasi STUDENTS’ DECLARATION We hereby declare that the project entitled "Operation Poshan" in fulfillment of completion of the fourth semester course – Machine Learning as part of the Bachelor of Technology (B. Tech) program at the School of Engineering and Technology, BML Munjal University is an authentic record of our work carried out under the supervision of Dr. Hirdesh Kumar Pharasi. Due acknowledgments have been made in the text of the project to all other materials used. This project was done in full compliance with the requirements and constraints of the prescribed curriculum. Aditya Yadav Aman Jain Dishita Amartya Kumar SUPERVISOR’S DECLARATION This is to certify that the above statement made by the candidate is correct to the best of my knowledge. Faculty Supervisor Name: Dr. Hirdesh Kumar Pharasi Signature: ABSTRACT In an era of increasing health consciousness, understanding the nutritional composition of food is paramount. "Operation Poshan" is a Streamlit-based web application designed to empower users with comprehensive and intuitive food nutritional analysis. Leveraging a rich dataset of over 40,000 food items sourced from reputable databases, the application enables users to explore detailed nutrient profiles, perform comparative analyses between different foods, and visualize key nutritional components. Furthermore, the project integrates machine learning models, including XGBoost, Random Forest, Linear Regression, and K-Nearest Neighbors, to predict nutritional quality scores and provide recommendations for nutritionally similar alternatives. By deploying these advanced analytical capabilities through a user-friendly web interface, "Operation Poshan" aims to bridge the gap between complex nutritional data and accessible dietary knowledge, facilitating informed food choices and promoting healthier eating habits. ACKNOWLEDGEMENT We would like to express our heartiest gratitude to our Machine Learning mentor Dr. Hirdesh Kumar Pharasi for his continued support, guidance and insights throughout the development of our project “Operation Poshan”. We also extend our appreciation to “BML MUNJAL UNIVERSITY” for providing the necessary resources and facilities. Aditya Yadav Aman Jain Dishita Amartya Kumar LIST OF FIGURES Figure Description

1 Flow Diagram 2.1 Distribution of food categories 2.2 Average Macronutrient Ratio by category 2.3 Nutrition Macronutrient Ratio by category 3.1 Food Comparison 3.2 Nutritional Analysis and Visualization of nutritional content 4.1 Diet planning and meal Analysis 4.2 Personalized Guidance TABLE OF CONTENTS STUDENTS' DECLARATION SUPERVISOR'S DECLARATION ABSTRACT ACKNOWLEDGEMENT LIST OF FIGURES 1 INTRODUCTION 2 LITERATURE REVIEW 3 PROBLEM STATEMENT 4 METHODOLOGY 1. DATA COLLECTION AND API INTEGRATION 2. DATA PREPROCESSING AND CLEANING 3. FEATURE ENGINEERING 4. EXPLORATORY DATA ANALYSIS AND VISUALIZATION 5. DATA STORAGE AND FINALIZATION 6. FEATURE SELECTION FOR NUTRITIONAL ANALYSIS 6.1. FEATURE IMPORTANCE ASSESSMENT 6.2. DIMENSIONALITY REDUCTION TECHNIQUES 6.3. DOMAIN-SPECIFIC FEATURE ENGINEERING 6.3.1. NUTRITIONAL RATIO DEVELOPMENT 6.3.2. HEALTH INDICATOR VARIABLES 6.4. VARIANCE INFLATION FACTOR (VIF) ANALYSIS 6.5. FEATURE SUBSET EVALUATION 6.6. FINAL FEATURE SET DETERMINATION 6.7. SELECTED FEATURE CATEGORIES 6.8. VALIDATION AND FINALIZATION 5 ANALYSIS AND DISCUSSION OF RESULTS 6 REFERENCES LIST OF ABBREVIATIONS Abbreviation Description API Application Programming Interface EDA Exploratory Data Analysis PCA Principal Component Analysis VIF Variance Inflation Factor RMSE Root Mean Squared Error MAE Mean Absolute Error R² Coefficient of Determination USDA United States Department of Agriculture SR Standard Reference (in USDA FoodData Central) XGBoost Extreme Gradient Boosting KNN K-Nearest Neighbors ANNOVA Analysis of Variance F-Value F-statistic (used in regression/ANOVA) INTRODUCTION In an era marked by increasing awareness of the intricate link between diet and health, access to comprehensive and user-friendly nutritional information has become paramount. Individuals are increasingly seeking to make informed dietary choices to optimize their well-being, manage health conditions, and achieve personal wellness goals. However, navigating the vast and often complex landscape of food nutrition data can be a significant challenge. Traditional methods of accessing this information, such as consulting static databases or deciphering lengthy nutritional labels, often prove cumbersome and lack the personalized insights that modern users demand. This project, "Operation Poshan: Your Smart Nutrition Hub," aims to address this challenge by developing an intuitive and intelligent web application that empowers users to explore, analyze, and understand food nutrition in a dynamic and engaging manner. By leveraging the power of data science and machine learning, "Operation Poshan" goes beyond simply presenting raw nutritional data. It provides users with detailed analyses of individual foods, facilitates insightful comparisons between different food items, offers interactive visualizations of nutrient profiles, and even provides personalized dietary guidance based on individual characteristics. The core of "Operation Poshan" lies in its ability to process and interpret a large and diverse dataset of food nutritional information, applying machine learning models to extract meaningful patterns and make intelligent predictions. This enables the application to offer features such as nutritional score prediction and recommendations for similar food alternatives, thereby enhancing the user's ability to make healthier and more informed choices. LITERATURE REVIEW The development of "Operation Poshan" draws upon several key areas of existing research and application: 1. Nutritional Databases and Information Systems: A foundational aspect of this project is the utilization of comprehensive food nutritional databases. Resources such as the USDA FoodData Central (encompassing Branded Foods, SR Legacy, and Foundation Foods) serve as critical sources of detailed nutritional information for a vast array of food items (U.S. Department of Agriculture, Agricultural Research Service. FoodData Central, 2019). These databases provide a standardized and reliable foundation for nutritional analysis and comparison. The challenges and methodologies involved in compiling, maintaining, and accessing such large-scale nutritional datasets have been extensively studied (e.g.,糧食營養成分資料庫[Taiwan Food and Drug Administration], McCance and Widdowson's Composition of Foods [Royal Society of Chemistry]). 2. Web-Based Nutritional Tools and Applications: The increasing accessibility of the internet

has led to the proliferation of web-based tools and applications aimed at providing nutritional information. These range from simple calorie counters and nutrient trackers to more sophisticated platforms offering recipe analysis and dietary planning (e.g., MyFitnessPal, Nutritionix API). Research in this area focuses on user interface design, data visualization techniques, and the effectiveness of these tools in promoting healthier eating habits (e.g., Miller & Brown, 2016). "Operation Poshan" builds upon this landscape by integrating advanced analytical capabilities and personalized features. 3. Machine Learning in Food and Nutrition: The application of machine learning techniques in the domain of food and nutrition has gained significant traction. Regression models have been employed to predict various food properties, including nutrient content and quality attributes (e.g., Zhang et al., 2018). Recommendation systems, utilizing techniques like collaborative filtering and content-based filtering, have been developed to suggest recipes or food alternatives based on user preferences or nutritional needs (e.g., Elsweiler et al., 2011). "Operation Poshan" leverages supervised regression models to predict nutritional scores and similarity-based algorithms to recommend alternative food choices, contributing to this growing body of work. 4. Data Visualization for Nutritional Information: Effectively communicating complex nutritional data to users is crucial for the success of any nutritional tool. Research in data visualization explores various techniques, such as bar charts, pie charts, and interactive dashboards, to represent nutrient profiles and comparisons in an understandable and engaging manner (e.g., Heer et al., 2010). "Operation Poshan" incorporates interactive visualizations using libraries like Plotly to enhance user comprehension and exploration of nutritional data. 5. Personalized Dietary Guidance Systems: The field of personalized nutrition aims to provide tailored dietary recommendations based on individual characteristics such as age, sex, activity level, and health status (e.g., van Ommen et al., 2018). While "Operation Poshan" currently offers general guidance, the underlying framework has the potential to be extended to incorporate more sophisticated personalization algorithms in future iterations, aligning with the broader trends in this domain. By integrating these diverse areas of research and application, "Operation Poshan" aims to provide a novel and valuable tool for individuals seeking to enhance their understanding of food nutrition and make more informed dietary decisions. The project contributes to the ongoing efforts to leverage technology in promoting health and well-being through accessible and intelligent nutritional information systems. PROBLEM STATEMENT The increasing prevalence of diet-related health issues and the overwhelming volume of available nutritional information highlight a significant gap in accessible and actionable tools for individuals seeking to make informed dietary choices. While vast nutritional databases exist, extracting personalized insights and understanding the complex interplay of various nutrients remains a challenge for the average user. Existing web-based nutritional tools often lack the sophistication to provide nuanced analyses, intelligent comparisons, and tailored guidance based on individual needs and preferences. This necessitates the development of a user-centric application that leverages data science and machine learning to simplify nutritional information, empower users with deeper understanding, and ultimately facilitate healthier dietary behaviors. The core problem lies in transforming complex nutritional data into easily digestible, comparable, and personalized insights to support informed decision-making in everyday food choices. METHODOLOGY Figure 1 : Flow diagram 1. Data Collection and API Integration The project utilizes a systematic approach to gather comprehensive nutritional data: Food Categories Definition The project implements a sophisticated categorization system with 8 main categories and over 150 specific food items: • Indian Dishes: Traditional items like butter chicken, biryani, dal makhani • International Dishes: Global cuisines including pizza, sushi, pasta • Proteins: Various meat, fish, and plant-based protein sources • Vegetables & Fruits: Fresh produce and common ingredients • Grains & Legumes: Basic staples and whole grains • Dairy & Alternatives: Milk products and non-dairy options • Snacks & Desserts: Common treats and processed foods • Beverages: Various drinks and liquid refreshments 2.

Data Preprocessing and Cleaning The data undergoes rigorous cleaning and standardization procedures: Data Cleaning Process • Null Value Handling: Removal of entries with missing critical nutritional values • Duplicate Removal: Elimination of duplicate entries based on name and category • Data Type Standardization: Conversion of data types for appropriate analysis • Outlier Detection: Identification and management of nutritional outliers 3.Feature Engineering 4. Exploratory Data Analysis and Visualization The project employs sophisticated data visualization techniques to extract meaningful insights: Figure 2 : Distribution of food categories Figure 3 : Average Macronutrient Ratio by category Figure 4 : Nutrition Macronutrient Ratio by category Distribution Analysis : The distribution of key nutritional components is analyzed using histogram plots with kernel density estimation: Correlation Analysis Relationships between nutritional components are examined using correlation heatmaps: Category Distribution Analysis The distribution of food categories is visualized using count plots and pie charts: Macronutrient Composition Analysis The macronutrient composition is analyzed using stacked bar charts: Nutrition Quality Analysis The nutrition scoring system is analyzed across food categories: Health Indicator Analysis The project analyzes various health indicators to provide nutritional insights: High-Protein Food Analysis: Low-Fat Food Analysis : 5. Data Storage and Finalization The project implements proper data storage practices: 6. Feature Selection for Nutritional Analysis : 6.1. Feature Importance Assessment Correlation Analysis : The feature selection process begins with a thorough correlation analysis to identify relationships between nutritional variables and the target metric (nutrition score). This involves: • Correlation Matrix Generation: Creating a comprehensive correlation matrix across all nutritional variables • Target Variable Correlation: Specifically examining correlations with the nutrition score • Visual Correlation Analysis: Using heatmaps and bar charts to visualize significant correlations • Threshold-Based Selection: Identifying features with correlation coefficients exceeding ±0.4 with the nutrition score The analysis revealed that protein content, fiber, vitamins (particularly C and A), and minerals like iron showed strong positive correlations with nutrition score, while sugar, saturated fat, and sodium demonstrated significant negative correlations. Statistical Feature Importance : Statistical tests were employed to quantify the importance of each feature: • F-regression Testing: Applied to assess feature relevance for regression tasks • ANOVA F-value: Used to measure the linear dependency between features and target • P-value Analysis: Features with p-values < 0.05 were prioritized as statistically significant • Rank-Based Selection: Features were ranked by their statistical importance scores This analysis confirmed the importance of macronutrients (protein, fat, carbohydrates) and identified micronutrients (vitamins, minerals) that significantly influence nutritional quality. 6.2. Dimensionality Reduction Techniques : Principal Component Analysis (PCA) : PCA was implemented to reduce dimensionality while preserving variance: • Standardization: Features were standardized to ensure comparable scales • Variance Retention: Components capturing 95% of cumulative variance were retained • Component Analysis: The first 7 principal components captured the majority of variance • Feature Loading Analysis: Examining feature contributions to principal components • Variance Plot Analysis: Creating scree plots to visualize explained variance by components The PCA revealed that macronutrient composition and key vitamins/minerals contributed most significantly to the variance in the dataset. Feature Clustering : Similar features were clustered to reduce redundancy: • Hierarchical Clustering: Features were clustered based on similarity • Dendrograms: Visualized feature relationships and clusters • Representative Feature Selection: From each cluster, the most informative feature was selected • Collinearity Reduction: This process effectively reduced multicollinearity among variables 6.3. Domain-Specific Feature Engineering 6.3.1 Nutritional Ratio Development : Several specialized nutritional ratios were engineered to enhance analytical capabilities: • Macronutrient Ratios: Percentage of calories from protein, fat, and carbohydrates • Nutrition Density Ratio: Nutrient content relative to caloric density • Fiber-to-Sugar Ratio: Indicator of food quality and digestive health impact • Protein Quality Index: Considering protein content relative to total calories • Micronutrient Density Score: Assessing vitamin and mineral content per calorie These engineered features provided more nuanced insights than raw nutritional values alone. 6.3.2 Health Indicator Variables : Binary indicator variables were created to classify foods based on health properties: • High-Protein Markers: Foods with >20% calories from protein • Low-Fat Indicators: Foods with <30% calories from fat • High-Fiber Flags: Foods containing >3g fiber per serving • Low-Sugar Labels: Foods with <5g sugar per serving • Low-Sodium Tags: Foods containing <140mg sodium per serving These indicators enabled categorical analysis of nutritional profiles across food groups. 6.4. Variance Inflation Factor (VIF) Analysis Multicollinearity among features was systematically addressed: • VIF Calculation: Computed for all numerical features • Threshold Application: Features with VIF > 10 were flagged as highly collinear • Iterative Reduction: High-VIF features were progressively removed • Stability Verification: Ensuring model stability after collinearity reduction This process identified and resolved strong correlations between variables like different fat measurements and various carbohydrate metrics. 6.5. Feature Subset Evaluation The effectiveness of different feature subsets was rigorously evaluated: • Cross-Validation: 5-fold cross-validation assessed feature subset performance • Model Comparison: Decision trees, random forests, and regression models compared • Performance Metrics: RMSE, R^2, and MAE used to evaluate predictive accuracy • Stability Analysis: Examining performance variability across cross- validation folds Through this process, a core set of 15 features was identified that balanced predictive power with model simplicity. 6.6. Final Feature Set Determination The final feature selection integrated multiple approaches: • Consensus Ranking: Features were ranked by their appearance in multiple selection methods • Domain Knowledge: Nutritional expertise guided the inclusion of critical nutrients • Practical Considerations: Measurement reliability and availability influenced selection • Parsimony Principle: Preference for simpler models with fewer features when performance was comparable 6.7. Selected Feature Categories The final selection included these key feature categories: - Primary Macronutrients : • Protein content (total and per 100g) • Fat content (total, saturated, and trans) • Carbohydrate content (total, fiber, and sugar) - Key Micronutrients : • Vitamins (A, C, D) • Minerals (calcium, iron, potassium, sodium) - Derived Nutritional Metrics : • Macronutrient ratios (protein, fat, carb percentages) • Nutrition density scores • Fiber-to-sugar ratio - - Health Indicators : • Binary health classification variables (high-protein, low-fat, etc.) Food Categorization : • Food category (encoded) • Serving size standardization 6.8. Validation and Finalization The selected feature set underwent final validation: • Out-of-Sample Testing: Verified performance on holdout data • Expert Review: Nutritionists reviewed the selected features for relevance • Sensitivity Analysis: Assessed impact of feature removal on model performance • Documentation: Comprehensive documentation of selection rationale and methodology ANALYSIS AND DISCUSSION OF RESULTS Web Interface using Streamlit : - Food Comparison Figure 5 : Food Comparison - Nutritional Analysis and Visualization of nutritional content Figure 6 : Nutritional Analysis and Visualization of nutritional content - Diet planning and meal Analysis Figure 7 : Diet planning and meal Analysis - Personalized Guidance Figure 8 : Personalized Guidance