OPERATION POSHAN

Machine Learning Project

01. AMARTYA KUMAR

02. ADITYA YADAV

03. AMAN JAIN

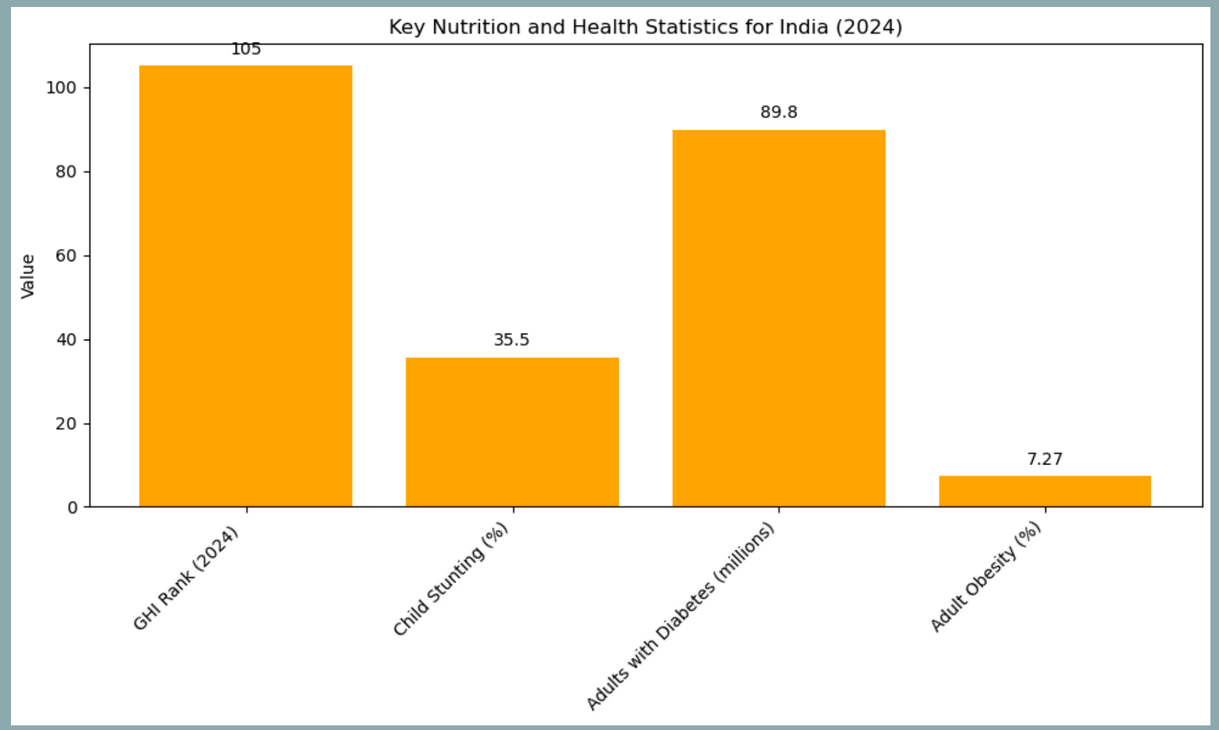
04. DISHITA TIRTHANI

TEAM MEMBERS

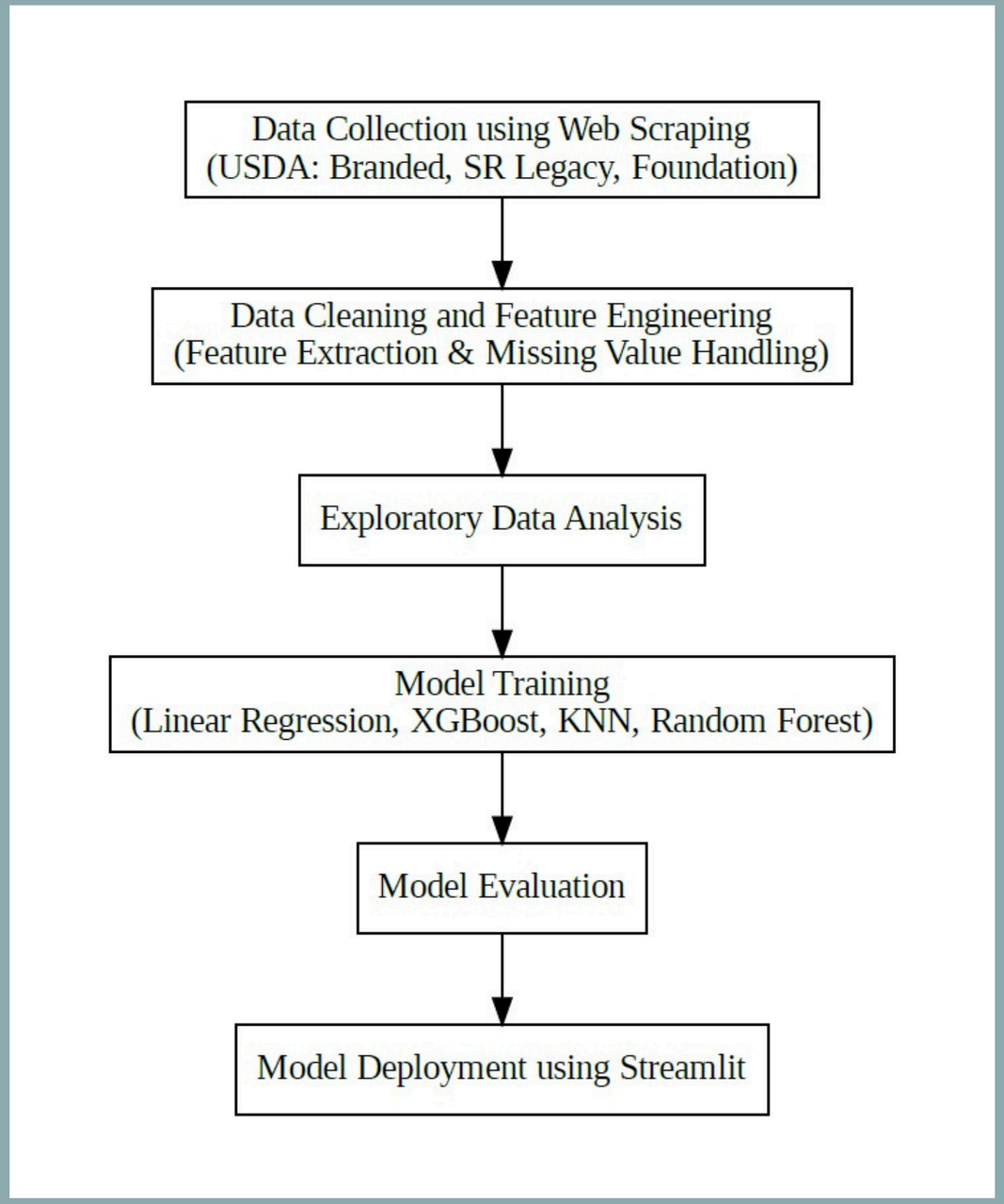
PROBLEM STATEMENT :

- LACK OF EFFECTIVE NUTRITIONAL ANALYSIS TOOLS LIMITS INDIVIDUALS FROM MAKING INFORMED, HEALTH-CONSCIOUS FOOD CHOICES.

- ABSENCE OF PERSONALIZED DIETARY INSIGHTS CONTRIBUTES TO POOR LIFESTYLE HABITS AND LONG-TERM HEALTH ISSUES.



METHODOLOGY!



MADE USING GRAPHIZ

DATA COLLECTION

SOURCE : WWW.USDA.GOV

```
Dataset saved with 39253 entries and 34 columns

Dataset Preview:
                          name       category    brand  serving_size_g  \
0  Butter, Clarified butter (ghee)  indian_dishes  Generic           100.0
1                  Butter, salted  indian_dishes  Generic           100.0
2              Croissants, butter  indian_dishes  Generic           100.0
3                  Chicken spread  indian_dishes  Generic           100.0
4               Chicken, meatless  indian_dishes  Generic           100.0

   calories  protein_g  fat_total_g  carbs_g  fiber_g  sugar_g  ...  \
0     900.0       0.00        100.0     0.00      0.0        0  ...
1    3000.0       0.85         81.1     0.06      0.0        0  ...
2     406.0       8.20         21.0    45.80      2.6        0  ...
3     158.0      18.00         17.6     4.05      0.3        0  ...
4     936.0      23.60         12.7     3.64      3.6        0  ...

   carbs_per_100g  protein_ratio  fat_ratio  carb_ratio  is_high_protein  \
0            0.00           0.00     100.00        0.00           False
1            0.06           0.46      99.50        0.03           False
2           45.80           8.10      46.67       45.23           False
3            4.05          29.20      64.23        6.57            True
4            3.64          42.28      51.20        6.52            True
```

NO. OF DATABASES :

USDA-SR LEGACY
USDA-FOUNDATION
USDA-BRANDED

DATA CLEANING AND FEATURE ENGINEERING

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

    # Calculate per 100g values
    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
    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 health indicators
    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

    # Nutrition score (0-100 scale)
    df['nutrition_score'] = (
        (df['protein_per_100g'] * 2) +
        (df['fiber_g'] * 3.5) -
        (df['sugar_g'] * 0.5) -
        (df['saturated_fat_g'] * 1.5) -
        (df['trans_fat_g'] * 3) +
        (df['vitamin_c_mg'] / 60 * 10) +
        (df['iron_mg'] / 18 * 10) -
        (df['sodium_mg'] / 2300 * 10)
    ).clip(0, 100).round(2)
```

FEATURE EXTRACTION:

CREATING 11 DERIVED COLUMNS LIKE : IS_HIGH_PROTEIN , CALORIES_PER_100G , FAT_RATIO , ETC.

HANDLING MISSING VALUES:

IN THE BRAND COLUMN : 299 NULL VALUES
FIXED BY CAPTIONING THEM AS GENERIC (MODE)

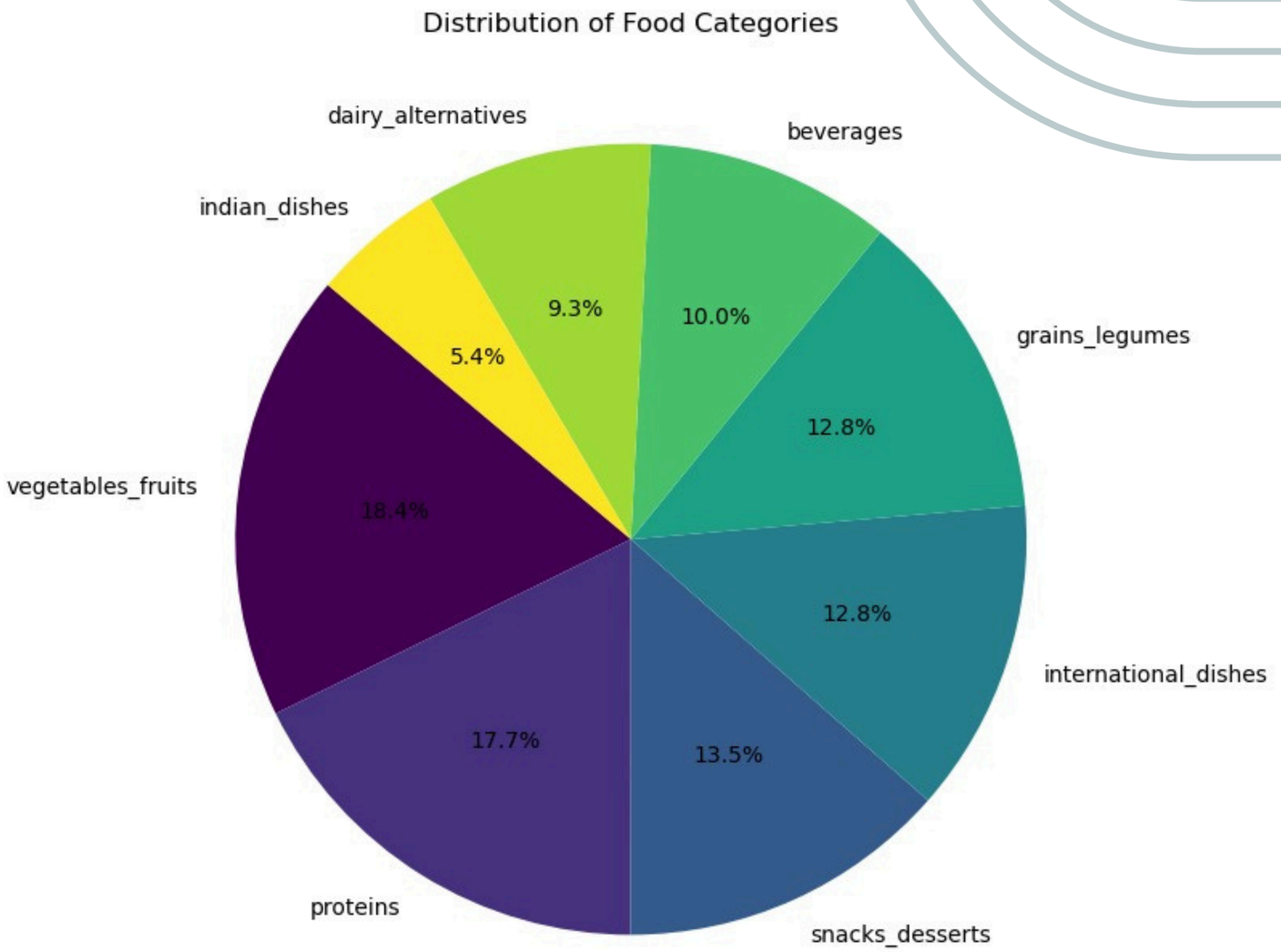
DROPPING DUPLICATE VALUES :

FROM THE COLUMNS “NAME” , “CATEGORY”

EXPLORATORY DATA ANALYSIS

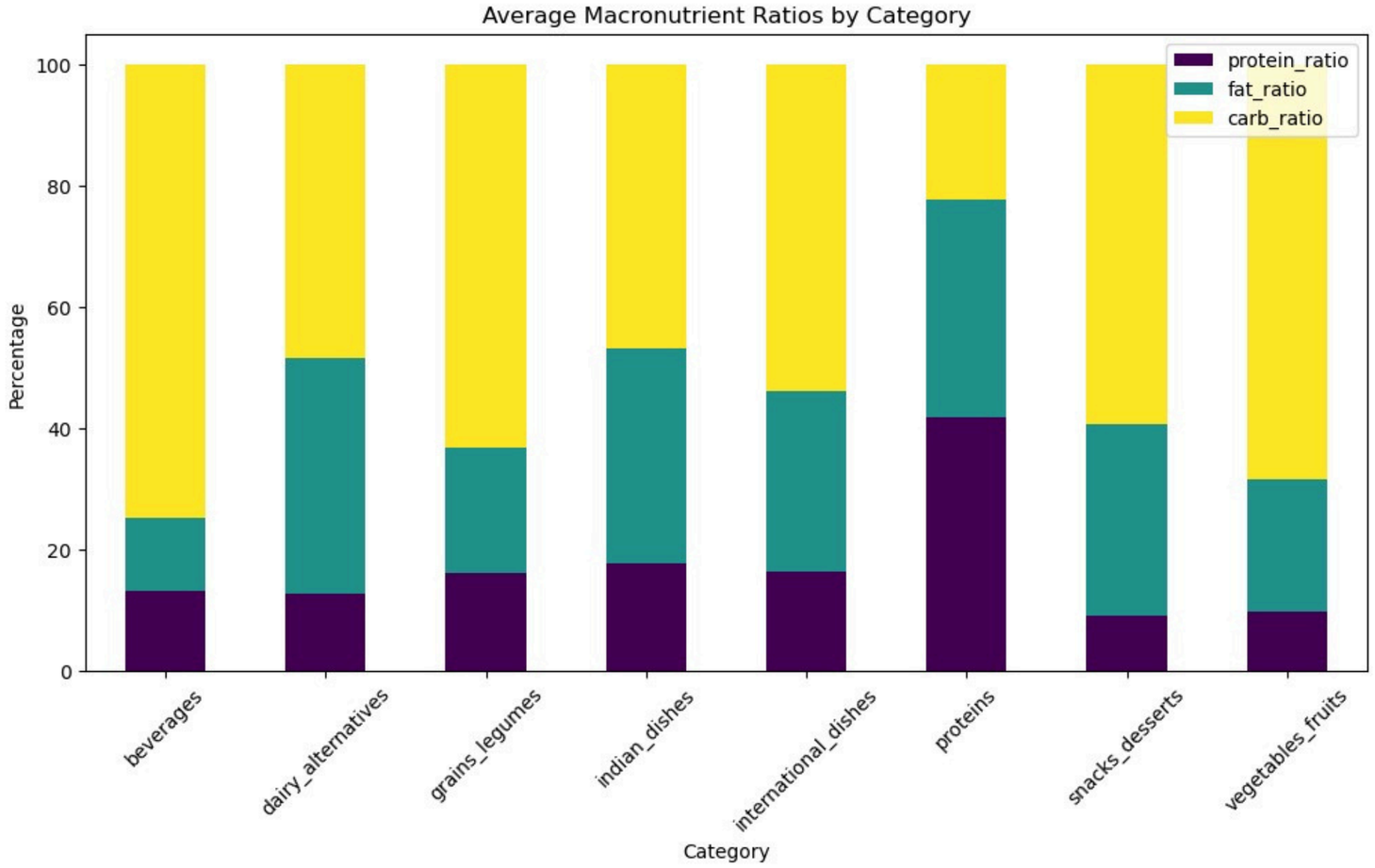
CONTRIBUTION IN THE DATABASE IN DECREASING ORDER :

VEGETABLES_FRUITS
PROTEINS
SNACKS_DESSERTS
INTERNATIONAL_DISHES
GRAINS_LEGUMES
BEVERAGES
DAIRY_ALTERNATIVES
INDIAN_DISHES



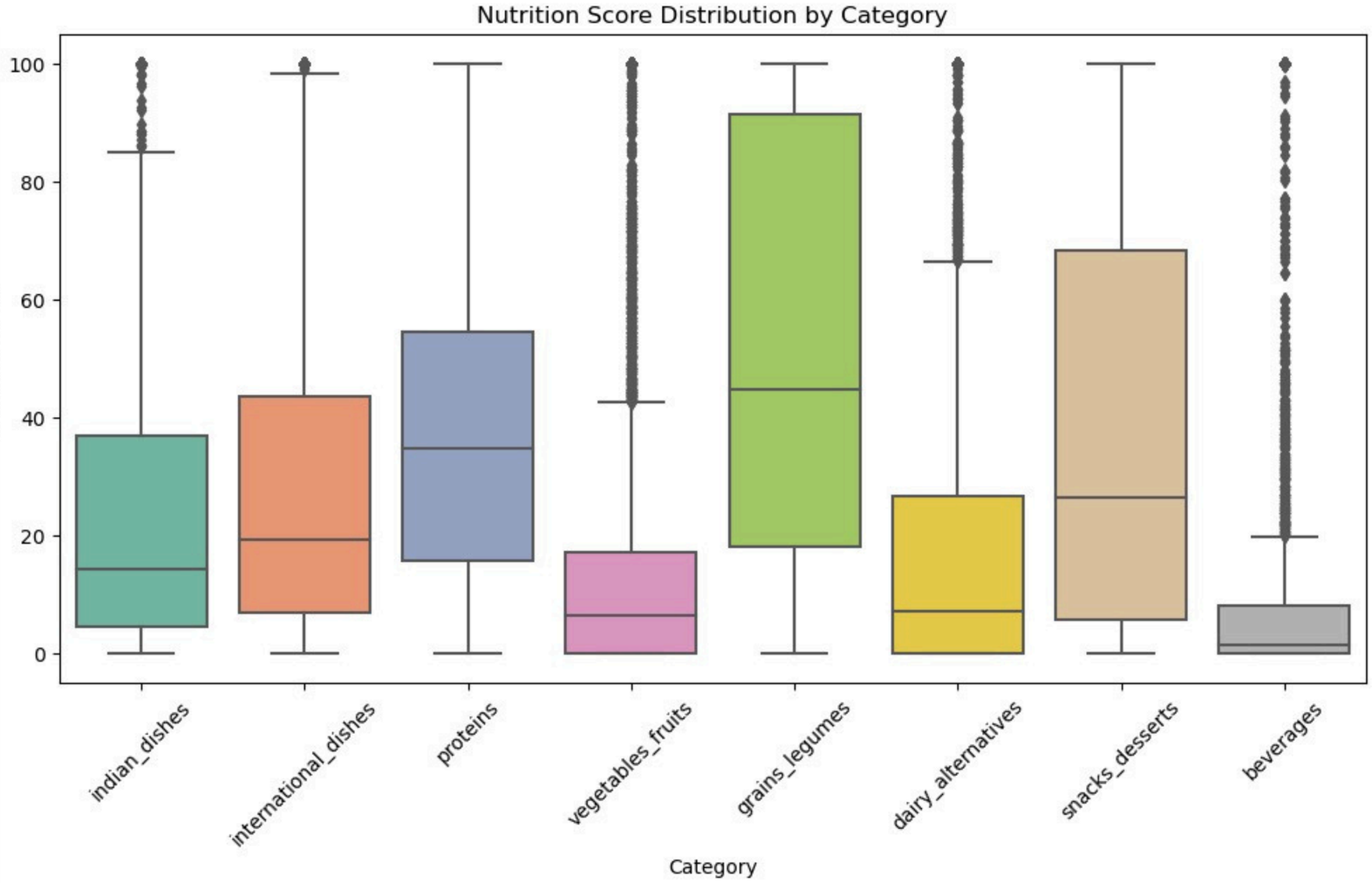
EXPLORATORY DATA ANALYSIS

- HIGH CARB CATEGORIES: BEVERAGES, GRAINS_LEGUMES, SNACKS_DESSERTS, AND VEGETABLES_FRUITS ARE CARB-DOMINANT (>60%).
- HIGH FAT CATEGORIES: DAIRY_ALTERNATIVES AND INDIAN_DISHES HAVE HIGH FAT RATIOS (~35–40%).
- HIGH PROTEIN CATEGORY: PROTEINS CATEGORY STANDS OUT WITH THE HIGHEST PROTEIN CONTENT (>40%) AND LOWEST CARBS.



EXPLORATORY DATA ANALYSIS

- TOP NUTRITION SCORES: GRAINS_LEGUMES AND SNACKS_DESSERTS.
- MODERATE: PROTEINS AND INTERNATIONAL_DISHES.
- LOWEST: BEVERAGES, DAIRY_ALTERNATIVES, AND VEGETABLES_FRUITS.
- HIGH VARIATION: MOST CATEGORIES HAVE A WIDE RANGE OF NUTRITION SCORES, INDICATING BOTH HEALTHY AND UNHEALTHY ITEMS.



MODEL TRAINING

LINEAR REGRESSION

MEAN ABSOLUTE ERROR (MAE): 10.40
MEAN SQUARED ERROR (MSE): 669.48
R-SQUARED (R2): 0.36

RANDOM FOREST REGRESSION

MEAN ABSOLUTE ERROR (MAE): 3.28
MEAN SQUARED ERROR (MSE): 44.45
R-SQUARED (R2): 0.96

XG-BOOST REGRESSION

MEAN ABSOLUTE ERROR (MAE): 1.84
MEAN SQUARED ERROR (MSE): 11.68
R-SQUARED (R2): 0.99

KNN REGRESSION

MEAN ABSOLUTE ERROR (MAE): 4.55
MEAN SQUARED ERROR (MSE): 80.36
R-SQUARED (R2): 0.92

MODEL DEPLOYMENT

INDIVIDUAL FOOD ANALYSIS

- SEARCH AND VIEW DETAILED NUTRITIONAL BREAKDOWN (CALORIES, MACROS, MICROS).
- VISUALIZATIONS: PIE CHARTS FOR MACRONUTRIENTS, BAR CHARTS FOR KEY NUTRIENTS.

FOOD COMPARISON

- COMPARE MULTIPLE FOODS SIDE-BY-SIDE.
- HIGHLIGHT DIFFERENCES VIA TABLES AND GROUPED BAR CHARTS.
- FOCUS ON SPECIFIC NUTRIENTS OF INTEREST.

PERSONALIZED DIETARY GUIDANCE

- INPUTS: AGE, GENDER, ACTIVITY LEVEL, HEALTH GOALS.
- TAILORED FOOD/PORTION RECOMMENDATIONS AND HEALTHY ALTERNATIVES.

WHOLE DAY NUTRITION TRACKER

- LOG DAILY FOOD INTAKE.
- TOTAL INTAKE VS. PERSONALIZED TARGETS OR RDAS.
- FEEDBACK AND SUGGESTIONS FOR A BALANCED DIET.

THANK YOU