Final Report: Predicting Food Categories Based on Nutritional Content

Morgan Lo
Brown University
DATA 1030
December 11, 2024
https://github.com/mlo20030/data1030_midterm

Introduction

The objective of this project is to predict the category of food—either "Meat," "Fruits," "Vegetables," "Grains," or "Dairy"—using detailed nutritional information. Accurate classification of food categories has practical applications in areas such as meal planning, nutrition tracking, and personalized dietary recommendations, making it a valuable tool for dietitians and individuals focusing on health and wellness.

The dataset was sourced from the FoodStruct database on Kaggle, which provides detailed nutritional facts about a variety of food items. Each row in the dataset represents a food item, including macronutrients (e.g., fats, proteins, carbohydrates) and micronutrients (e.g., vitamins, minerals).

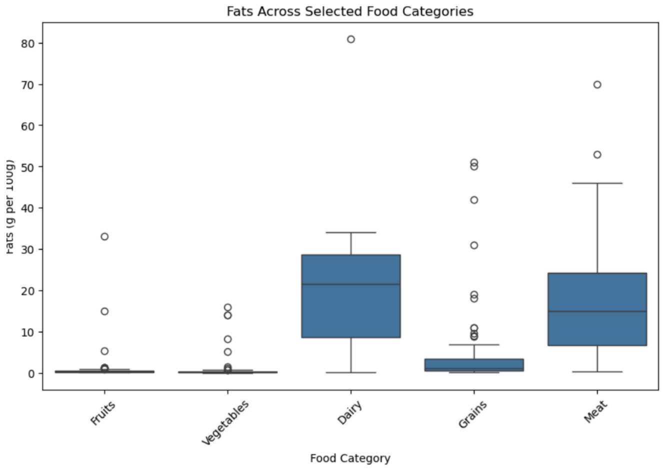
The primary challenges faced in this project included handling missing data, particularly for micronutrient columns such as Fructose and Starch, and identifying which features contribute most to model predictions. The project aimed to address these challenges while evaluating the performance of several ML models and interpreting their results.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis provided crucial insights into the dataset, revealing significant patterns and relationships among features. The analysis focused on visualizing fat content, caloric content, and nutrient distributions across food categories, which were instrumental in understanding the variability in the data.

Fat Content Across Categories

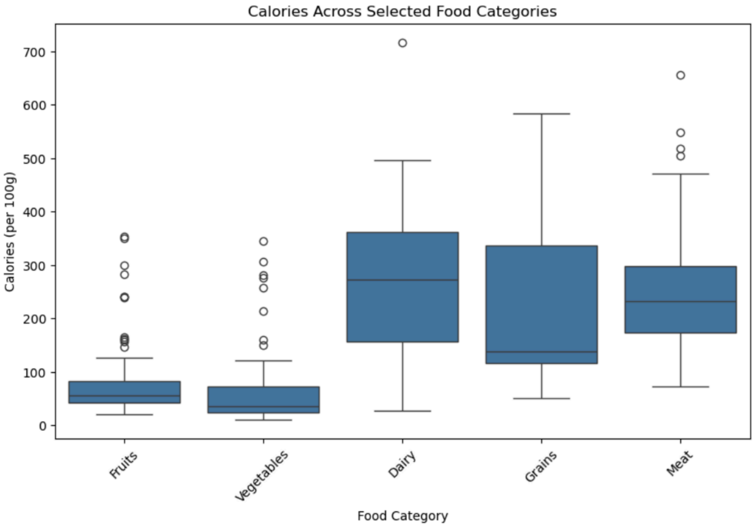
A box plot of fat content across the five selected food categories—Fruits, Vegetables, Dairy, Grains, and Meat—highlighted substantial variability in fat content between categories. "Meat" and "Dairy" exhibited higher median fat values, while "Fruits" and "Vegetables" displayed consistently low fat content. This variability indicates that fat content is a strong differentiating factor among food categories, making it a potentially valuable feature for classification.



Box Plot of Fat Content Across Food Categories Figure 1: Box Plot of Fat Content Across Food Categories
This box plot shows the distribution of fat content across "Fruits," "Vegetables," "Dairy," "Grains," and "Meat." Meat and Dairy exhibit higher fat content, while Fruits and Vegetables have significantly lower fat content.

Caloric Content Across Categories

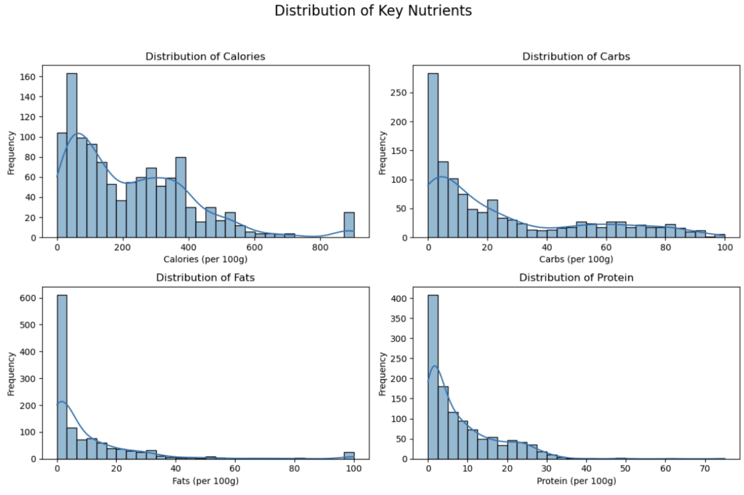
Similarly, a box plot of caloric content showcased distinct patterns. "Dairy" and "Meat" had higher caloric densities on average, whereas "Fruits" and "Vegetables" were comparatively lower. These insights align with dietary expectations, where dairy and meats are energy-dense, while fruits and vegetables are lighter in caloric content. The clear separation of caloric ranges across categories provides another informative feature for classification models.



Box Plot of Caloric Content Across Food Categories Figure 2: Box Plot of Caloric Content Across Food Categories
This box plot illustrates the caloric distribution across the five food categories. Dairy and Meat have the highest median caloric values, while Fruits and Vegetables have lower caloric density, consistent with their dietary profiles.

Nutrient Distribution

Bar plots with smooth density curves revealed the distribution of key nutrients, including Calories, Carbs, Protein, and Fat. Each nutrient exhibited a right-skewed distribution, indicating that most foods in the dataset have relatively low values for these nutrients, with a few outliers having significantly higher values. Among these, Calories displayed the most spread-out distribution, reflecting greater variability in caloric content across food items, but it remained right-skewed overall. Carbs, Protein, and Fat showed more concentrated distributions, with most food items having relatively low levels of these nutrients. The smooth density curves highlighted the central tendency and spread of each nutrient, providing a comprehensive view of the nutrient profiles and helping to identify potential outliers or deviations from expected patterns.



Nutrient Distribution for Calories, Carbs, Protein, and Fat Figure 3: Distribution of Calories, Carbs, Protein, and Fat
This figure shows bar plots with smooth density curves for four key nutrients: Calories, Carbohydrates, Protein, and Fat. All distributions are right-skewed, with Calories having the most dispersed distribution, reflecting the natural variability in caloric content among foods.

Methods

The dataset was split into training (60%), validation (20%), and test (20%) sets, ensuring stratified sampling to preserve class proportions. For preprocessing, numeric columns with any missing values were dropped for Logistic Regression, Random Forest, and KNN models, followed by reduced features method for the remaining missing values.

Feature scaling was applied to Logistic Regression, Random Forest, and KNN models using StandardScaler, while XGBoost was left unscaled to leverage its native handling of numerical data.

Four machine learning models were evaluated: Logistic Regression, Random Forest, XGBoost, and KNN. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. The tuned parameters for each model were as follows:

- **Logistic Regression**: Regularization parameter (C = [0.01, 0.1, 1, 10]).
- **Random Forest**: Number of estimators (n_estimators = [50, 100, 200]), maximum depth (max_depth = [None, 10, 20]), and minimum samples split (min_samples_split = [2, 5, 10]).
- **XGBoost**: Number of estimators (n_estimators = [100, 200]), maximum depth (max_depth = [3, 5, 7]), and column subsampling by tree (colsample_bytree = [0.6, 0.8, 1.0]).
- **KNN**: Number of neighbors (n_neighbors = [3, 5, 7, 9]) and weights (weights = ['uniform', 'distance']).

The primary evaluation metric was accuracy, with additional insights provided by confusion matrices and global/local feature importance. Baseline accuracy, calculated as the proportion of the most frequent class, was 20%.

Results

The results of the model evaluation are summarized in the table and supported by visualizations of model accuracy, confusion matrix, and global and local feature importances.

Model	**Best Parameters**	**Validation Accuracy**	**Test Accuracy**
Logistic Regression	C=10	0.7407	0.8148

Random Forest	max_depth=None,
	n_estimators=50
		0.7901	0.8395

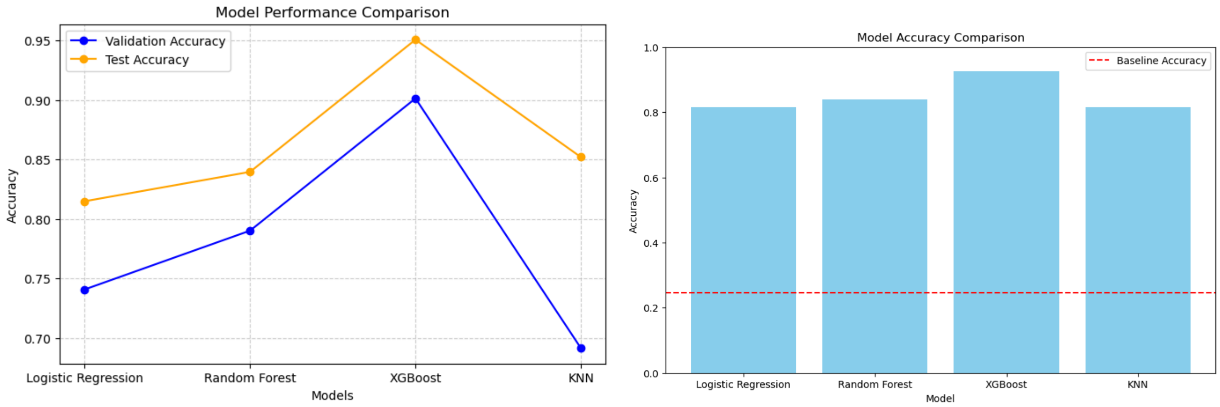
XGBoost	max_depth=5,
	n_estimators=100,
	colsample_bytree=0.6
		0.9012	0.9259

KNN	n_neighbors=5	0.6420	0.8148

Model Performance Table Figure 4: Model Performance Summary
This table summarizes the best hyperparameters, validation accuracies, and test accuracies for the four models: Logistic Regression, Random Forest, XGBoost, and KNN. XGBoost outperformed the other models with a test accuracy of 92.59%.

Model Performance

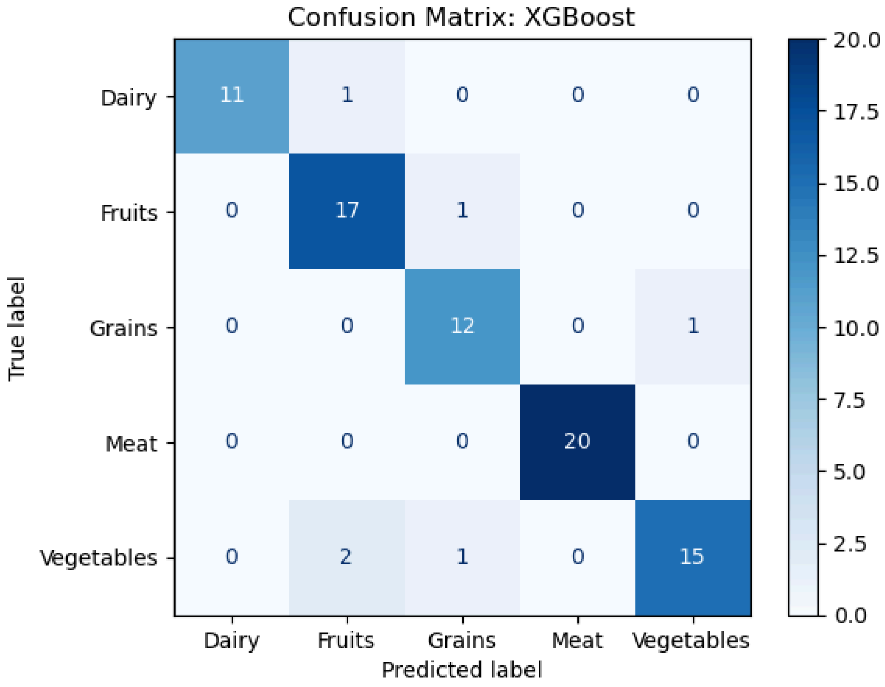
Each model was evaluated using validation and test accuracies. As shown in the accompanying bar plot, XGBoost outperformed other models, achieving a test accuracy of **92.59%**, compared to the baseline accuracy of **20%**. Random Forest and Logistic Regression also performed well, with test accuracies of **83.95%** and **81.48%**, respectively. KNN showed lower performance, matching Logistic Regression’s test accuracy of **81.48%**, but it lagged in validation accuracy, suggesting it was less robust.



Model Accuracy Comparison Figure 5: Accuracy Comparison Across ML Models
This bar plot compares the accuracy of the four machine learning models on the test set. XGBoost achieved the highest test accuracy (92.59%), significantly outperforming the baseline accuracy of 20% (shown as a red dashed line).

Confusion Matrix for XGBoost

The confusion matrix for the XGBoost model shows excellent classification performance across all categories. Most predictions are concentrated along the diagonal, indicating high precision and recall. Misclassifications were minimal, with only a few instances of "Vegetables" being misclassified as "Fruits." This suggests the model can effectively distinguish between nutritionally similar categories.

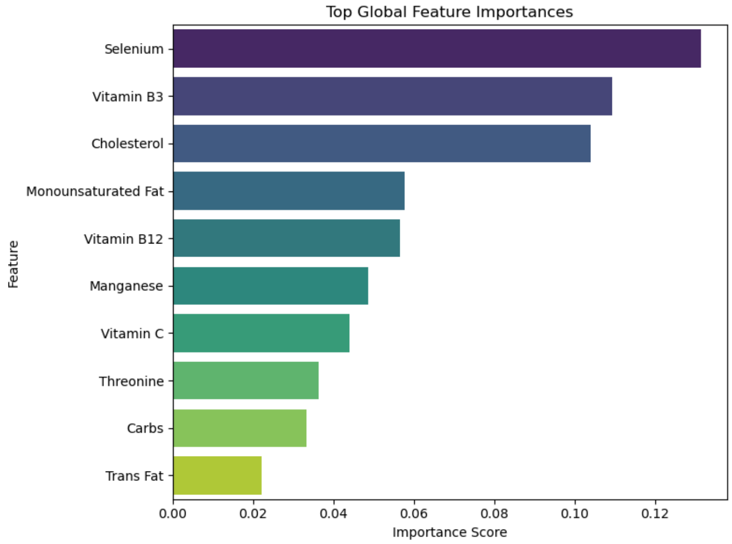


	Dairy	Fruits	Grains	Meat	Vegetables
Dairy	11	1	0	0	0
Fruits	0	17	1	0	0
Grains	0	0	12	0	1
Meat	0	0	0	20	0
Vegetables	0	2	1	0	15

XGBoost Confusion Matrix Figure 6: Confusion Matrix for XGBoost Model
This confusion matrix illustrates the classification results of the XGBoost model on the test set. The majority of predictions lie on the diagonal, indicating high accuracy across all food categories. Misclassifications were minimal, with most errors involving misclassifying "Vegetables" as "Fruits."

Global Feature Importances

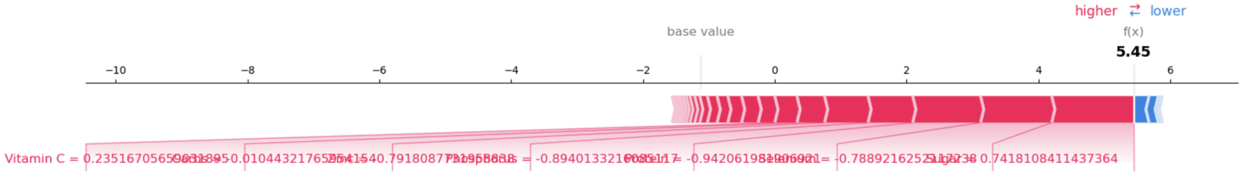
The global feature importance analysis from XGBoost highlights the most predictive features. Selenium, Vitamin B3, and Cholesterol were the top contributors, indicating their strong relationship with the target categories. This aligns with domain knowledge, as Selenium and Cholesterol are commonly associated with "Meat," while Vitamin B3 is crucial for energy production in foods like grains and meat. These insights provide a biological basis for the model's predictive accuracy, demonstrating how specific nutrients influence categorization.



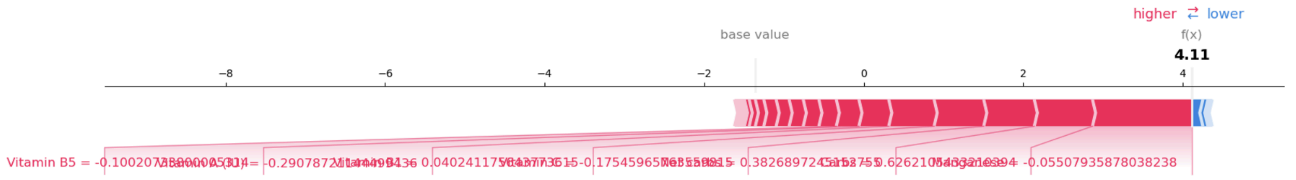
Global Feature Importance Figure 7: Top Global Feature Importances from XGBoost

This table highlights the most influential features for food categorization as identified by the XGBoost model. Selenium, Vitamin B3, and Cholesterol are the most important features, indicating their predictive relevance in distinguishing food categories.

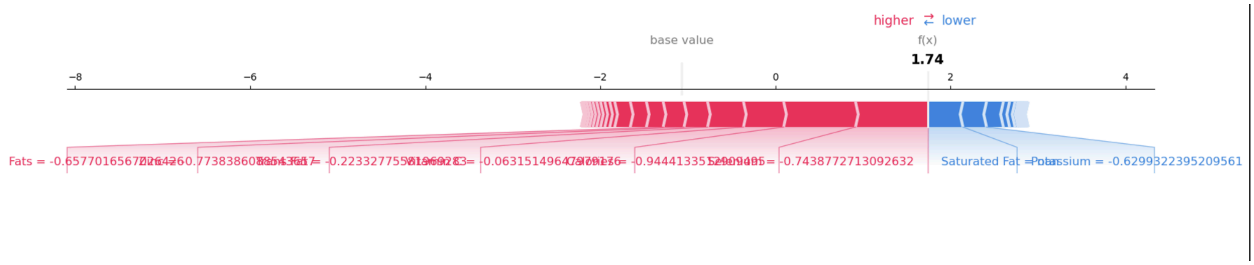
The bar plot of feature importances underscores the dominance of these key variables. This not only supports the model's robustness but also offers practical guidance for future applications in food classification systems.



Force Plot for Index 0 Figure 8: *This SHAP force plot illustrates the model's prediction for Index 0. The key positive contributors to the prediction are Selenium (SHAP: 0.82), Calories (SHAP: 0.82), and Vitamin C (SHAP: 0.46), which collectively push the prediction higher. Conversely, the negative contributors are Saturated Fat (SHAP: -0.40), Potassium (SHAP: -0.27), and Folate (SHAP: -0.19), which reduce the prediction's score.*



Force Plot for Index 1 Figure 9: *The SHAP force plot for Index 1 highlights the primary drivers of the model's prediction. Positive contributions come from Sugar (SHAP: 1.24), Selenium (SHAP: 1.08), and Protein (SHAP: 1.03), while negative contributions from Monounsaturated Fat (SHAP: -0.23), Fiber (SHAP: -0.14), and Fats (SHAP: -0.02) decrease the prediction score. These factors influence the final prediction outcome for this data point.*



Force Plot for Index 2 Figure 10: *For Index 2, the SHAP force plot reveals that Manganese (SHAP: 1.24), Carbs (SHAP: 0.73), and Net Carbs (SHAP: 0.63) are the most significant positive contributors to the model's prediction. On the negative side, Potassium (SHAP: -0.14), Saturated Fat (SHAP: -0.06), and Copper (SHAP: -0.02) reduce the prediction score. These contributions collectively explain the model's final prediction for this instance.*

Outlook

Future work could involve incorporating additional features, such as cuisine type or preparation method, to enhance the context of predictions. Addressing class imbalance through class-weighted models would improve performance for underrepresented categories. Advanced techniques, such as neural networks, could be explored to capture complex nonlinear relationships. Finally, expanding SHAP analysis would provide deeper insights into feature interactions and model behavior.

References

1. Kaggle Dataset: FoodStruct Nutritional Facts
2. Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.
3. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*.