Predictive Model

-FINAL PROJECT-

Classification of Nutritional Content for Dietary Recommendations

[AIM-5004-1]

Team Members

Aaron Meoded

Milan Regmi

Tharun Prabhakar

Syed Ali Hussain

ABSTRACT

This study addresses the complex challenge of dietary recommendation systems by developing a robust predictive model that categorizes food items into 'Nourishing', 'Indulgent', and 'Balanced' based on comprehensive nutritional profiles. Utilizing a dataset comprising 7,793 diverse food items, extensive exploratory data analysis (EDA) was conducted to understand underlying patterns and distributions. Key preprocessing steps included techniques for handling missing data through imputation, normalization of nutritional values to uniform scales, and feature engineering to enhance model input relevance.

Various advanced machine learning models were rigorously evaluated, with a focus on Logistic Regression, Decision Tree, Grid Search Decision Tree Random Forest, and Gradient Boosting techniques. These models were assessed using a range of metrics, including precision, recall, and the F1 score, to determine their efficacy in accurately classifying the nutritional content. Both the Random Forest and Gradient Boosting models outperformed other techniques, demonstrating high precision and robustness, with superior F1 scores that underscore their potential for practical application.

The findings reveal significant potential for these models to contribute to automated nutritional guidance, offering precise and reliable classifications that can support public health initiatives and personalized diet planning. This research not only advances the field of nutritional science but also proposes a scalable and effective approach to food categorization that can be integrated into dietary management tools. Future work will explore the integration of more granular data and the potential for real-time application in consumer health technologies.

Table of Contents

Table of Figures

INTRODUCTION

Nutritional science plays a pivotal role in public health by informing dietary recommendations and influencing individual food choices. In an era marked by an abundance of dietary information and choices, the ability to categorically assess food based on nutritional content is more crucial than ever. However, traditional dietary guidelines often fail to provide the nuanced understanding necessary to cater to diverse dietary needs and preferences. This lack of specificity can lead to generalized advice that may not be applicable or optimal for everyone. The increasing prevalence of diet-related health issues further underscores the need for an advanced, data-driven approach to dietary recommendations.

Challenges in Nutritional Classification

Current methods for classifying foods typically rely on simplistic categorizations that do not consider the complex interplay of nutrients. For example, foods are often labeled as 'healthy' or 'unhealthy' based solely on calorie content or the presence of specific nutrients like fats or sugars. This binary classification overlooks various nutritional factors that contribute to health, such as the balance of macronutrients, the presence of essential vitamins and minerals, and the overall caloric density relative to the volume of food. Such oversimplifications can mislead consumers and may not align with modern nutritional science, which advocates for a more holistic view of diet.

Project Objectives

This project aims to transcend traditional classification schemes by leveraging machine learning techniques to categorize food items into three detailed health-oriented categories: Nourishing, Indulgent, and Balanced. These categories were devised to reflect a more comprehensive understanding of nutritional science:

- Nourishing: Foods that are high in proteins and/or fiber, low in saturated fats and sugars, and not excessively high in calories.
- Indulgent: Foods primarily high in sugars or saturated fats, low in beneficial nutrients, or very high in calories.
- Balanced: Foods that do not distinctly fall into the Nourishing or Indulgent categories, offering a reasonable compromise of nutrients.

Significance of the Study

By developing a model that accurately classifies foods into these categories based on their nutritional content, this study aims to provide a tool that can aid individuals in making more informed dietary choices. This model is particularly relevant in the context of personalized nutrition, where dietary recommendations can be tailored to individual health needs and preferences. Furthermore, the outcomes of this research have the potential to influence public health policy by providing a basis for more nuanced and scientifically sound dietary guidelines.

Structure of the Report

The following report details the methods used for data collection, preprocessing, and analysis, including the specific machine learning models employed and the evaluation of their performance. The results are discussed in the context of their implications for dietary management and future research directions in nutritional science.

METHODS

Data Collection

The dataset utilized in this study comprises nutritional information for 7,793 food items, each characterized by detailed macro and micronutrient profiles based on a 100g serving. Data points included calories, protein, carbohydrates, total fat, cholesterol, fiber, and various vitamins and minerals. This comprehensive dataset allowed for a nuanced analysis of food items across a broad spectrum of dietary categories.

Data Preprocessing

Data preprocessing is a critical step in preparing raw data for machine learning models. It basically refining data so that its easier for machine learning models to understand and predict. The preprocessing steps implemented in this study include:

- Handling Missing Values: Missing data were imputed with median and not mean as the features had outliers which would influence the mean.

- Feature Scaling: Nutrient data were scaled using the MinMaxScaler, which scales the data features to a given range, typically 0 to 1. This is essential as each feature has its own range and influence the model behavior, it ensures the range of data features remain the same, ensuring that no variable dominates another in terms of scale.

- Outlier Treatment: All the features had outliers. Hence outlier treatment was done by clipping the values to 25^{th} Quantile – 1.5 times the Interquartile Range and 75^{th} Quantile + 1.5 times the Interquartile Range so as to not loose importance of the features. Alcohol columns was not treated as treating them would loose the importance of the feature as most values of feature zero and only certain recipes had alcohol in them.

- Feature Engineering: Target Column “Health Category” and the “Category” features were transformed using Ordinal Encoder as it provides a level of importance to each category in the feature meaning “Nourishing” has the high priority and encoded as 2 whereas “Indulgent” is encoded as 0 with least priority, same goes for the “Category” feature.

Model Development

Several machine learning models were evaluated to identify the one that best classifies the nutritional content into Nourishing, Indulgent, and Balanced categories:

- Logistic Regression: A baseline model due to its simplicity and interpretability. However, it may not perform well with non-linear relationships unless explicitly modeled.

- Decision Tree Classifier: Offers easy interpretability, which is valuable for decision-making processes. Nonetheless, it is prone to overfitting, especially with complex datasets.

- Random Forest Classifier: An ensemble model that mitigates overfitting problems of decision trees and is robust to noise. Its main disadvantage is that it can be computationally intensive and less interpretable compared to a single decision tree.

- Gradient Boosting Classifier: Another ensemble model that often provides high accuracy through boosting techniques that focus on correcting the mistakes of prior trees. The trade-off includes being more sensitive to overfitting and requiring careful tuning of parameters.

Model Evaluation

Model performance was primarily evaluated using the F1 score, which balances precision (the accuracy of positive predictions) and recall (the ability to find all relevant instances). This metric is particularly useful when dealing with imbalanced datasets, as is common with food classification tasks. Additional metrics included accuracy, precision, and recall for a comprehensive assessment. Validation techniques such as cross-validation were employed to ensure the model's robustness and generalizability.

- Confusion Matrix: Used to visualize the performance of the classification model. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. This is crucial for seeing the model's performance across different classes.
- ROC Curves and AUC Scores: Provided insights into the trade-offs between true positive rate and false positive rate at various threshold settings, which is critical for models where the balance between sensitivity and specificity is crucial.

RESULTS

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the underlying structure and distribution of the data, which is crucial for effective model application and interpretation.

- Distribution Analysis: The analysis began with the examination of nutrient distributions across different food categories. Key nutrients such as proteins, carbohydrates, fats, and fibers showed varied distributions, indicating diverse food profiles that necessitate nuanced classification approaches.
- Correlation Analysis: Correlation matrices were created to explore the relationships between different nutrients. High correlations were found between fats and calories, which is expected, and between fibers and proteins, highlighting potential groups of nourishing foods.
- Outlier Detection: Outliers were identified using IQR (Interquartile Range) for each nutrient. Foods with extreme values were scrutinized to understand whether they represent special cases like fortified foods or errors in data entry.

Handling Missing Values

Missing values pose significant challenges in predictive modeling as they can introduce bias and affect model accuracy. The approach to handling missing data included:

- Imputation Strategy: Where possible, missing nutrient values were imputed using the median imputation method, chosen due to its robustness against outliers. For categorical data like food categories, the mode was used.
- Deletion: In cases where imputation was not appropriate, particularly when a significant proportion of key data was missing, rows were removed. This was considered a last resort to maintain data integrity.

Scaling of Data

Scaling is critical to ensure that no single feature disproportionately influences the model outcome.

- MinMax Scaling: The MinMaxScaler was employed to transform nutrient amounts, scaling them to a [0, 1] range. This scaling was particularly important for models that are sensitive to the magnitude of input features, such as Logistic Regression and Neural Networks.

Heatmaps of Correlation Matrices

From the heatmap we can see that the multicollinearity exists in the data which would affect the performance of the model. But for intial analysis multicollinearity was not treated as the dataset is small and this could lead to the loosing of important trends in the dataset.

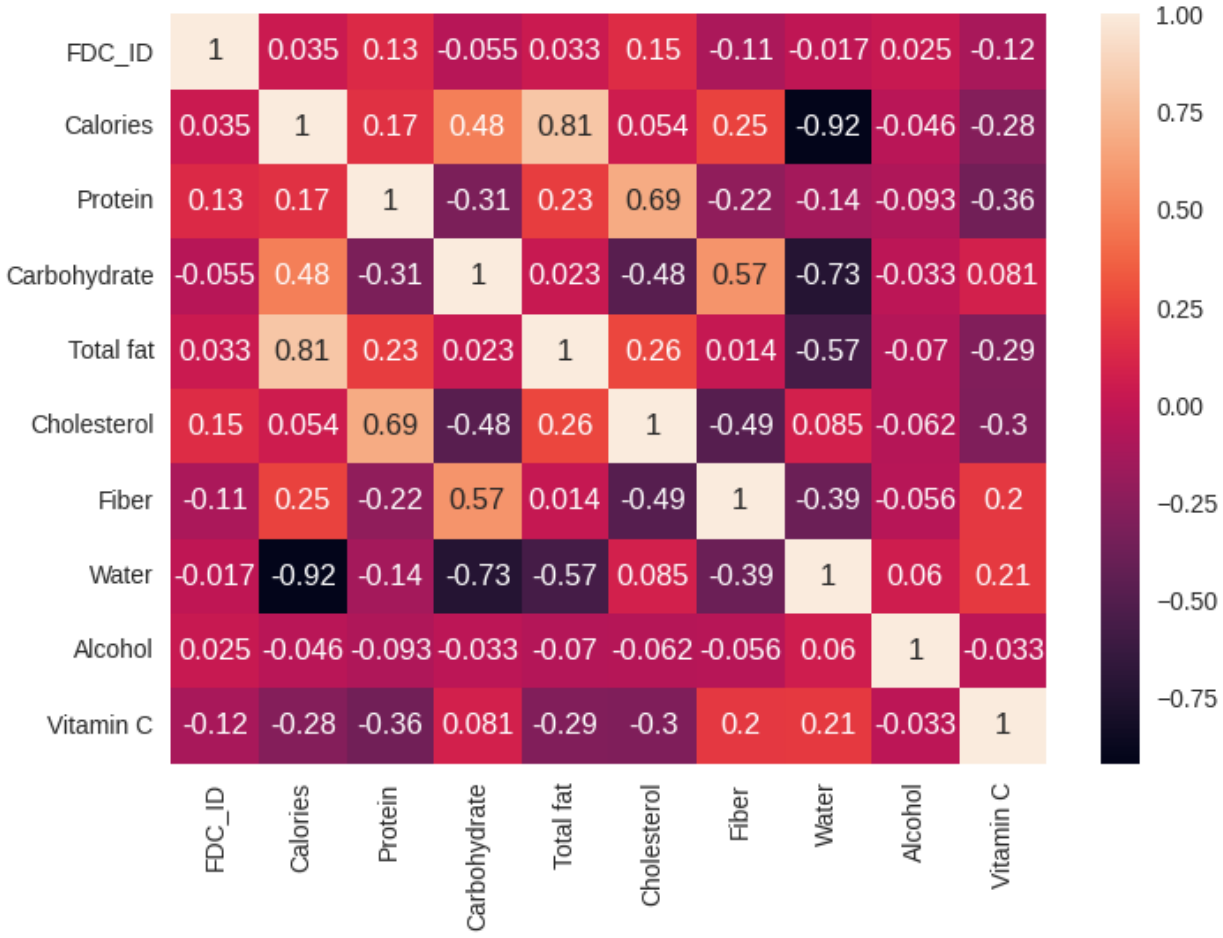


Fig 1: Heatmap for Correlation

Visualization of Preprocessing Results

Several visualizations were produced to illustrate the effects of EDA and preprocessing:

- Box Plots: Utilized to display the effect of outlier handling on the distribution of different nutrients. These plots highlighted the normalization of extreme values which could bias the models.

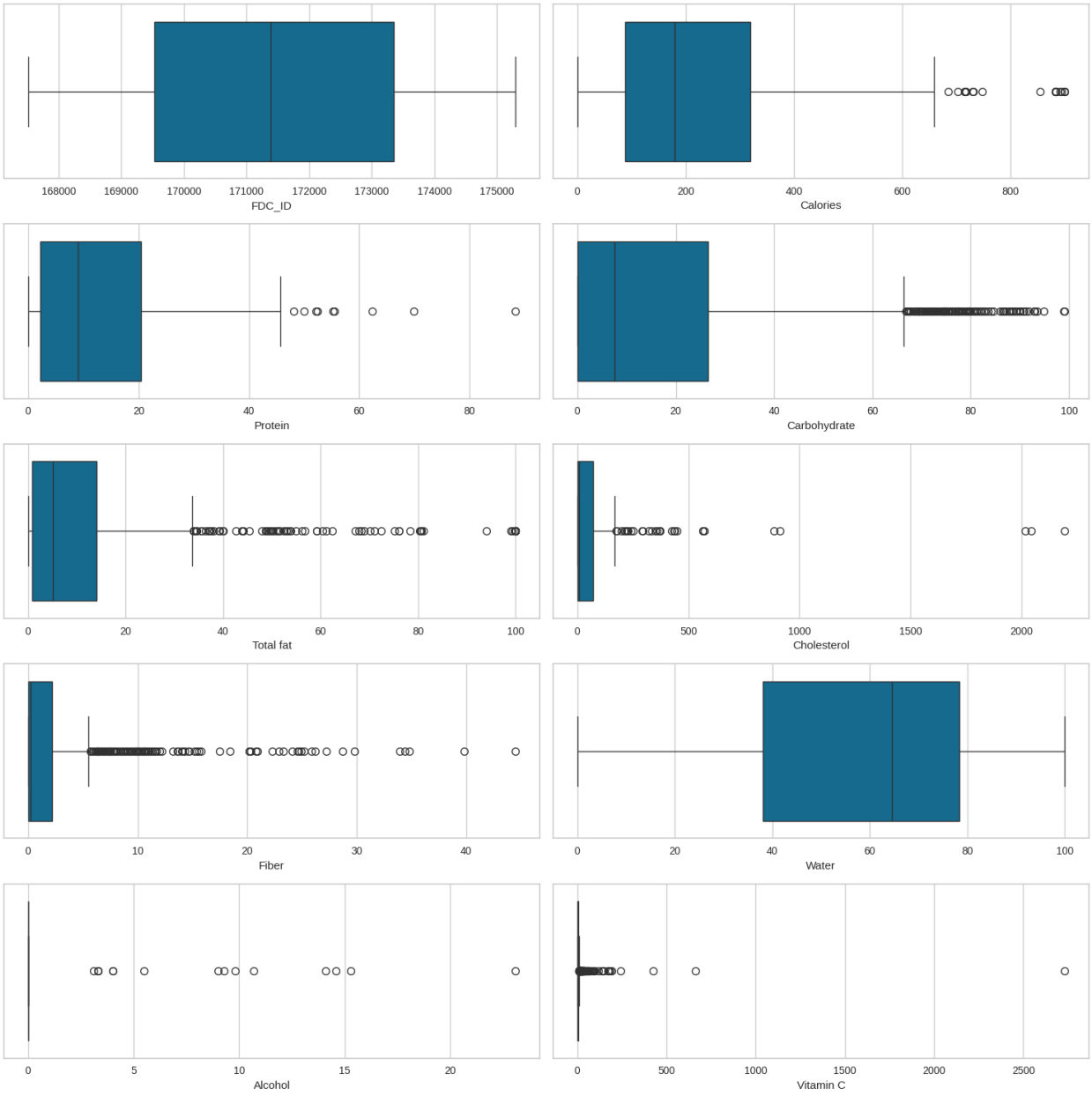


Fig 2: Box Plots of Columns

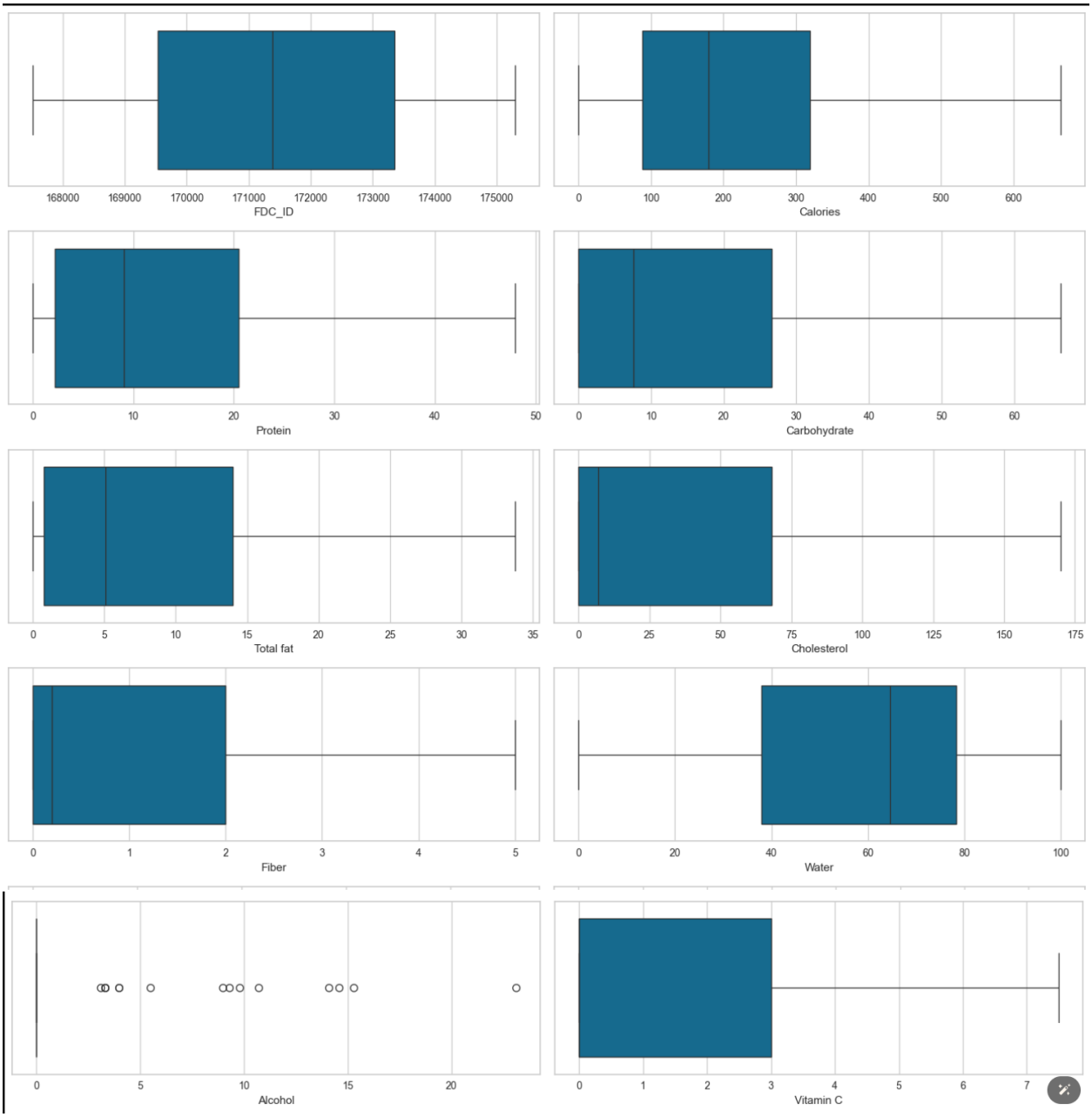
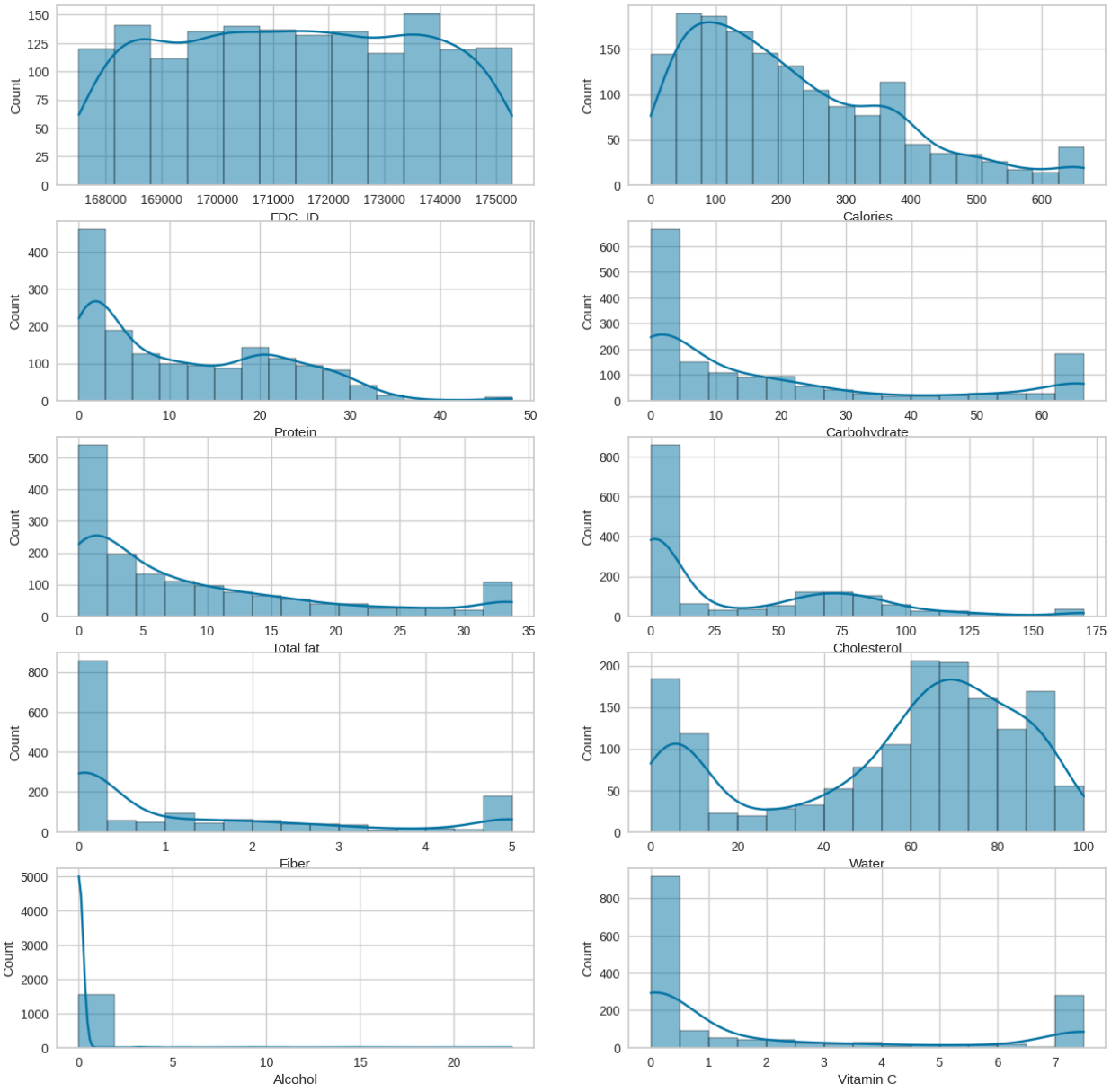


Fig 3: Box plots after outlier treatment

- Histograms and Skewness: Histograms for key features were done to check the distribution along with checking the skewness of the data. From the below result we can see that the data is right skewed with tails on the right side.



FDC_ID	0.001764
Calories	0.904673
Protein	0.642002
Carbohydrate	1.153826
Total fat	1.191444
Cholesterol	1.033586
Fiber	1.201637
Water	-0.679180
Alcohol	14.949544
Vitamin C	1.270426

Fig 4: Histogram and Table for Skewness values

Overview of Model Performance

The study evaluated several machine learning models to classify foods into three nutritional categories: Nourishing, Indulgent, and Balanced. The performance of each model was measured using accuracy, precision, recall, and the F1 score. The models tested included Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Classifier.

	Model	Misclassifications	precision validation	recall validation	accuracy train	accuracy validation	f1 validation	auc validation
0	Base Logistic Regressioin with CV	79	0.74	0.75	0.77	0.75	0.74	0.9054
1	Base Decision Tree	5	0.98	0.98	1.00	0.98	0.98	0.9863
2	GridSearch Decision Tree	5	0.98	0.98	1.00	0.98	0.98	0.9894
3	Base RandomForest CLassifier	3	0.99	0.99	1.00	0.99	0.99	0.9999
4	Base Gradient Boosting	0	1.00	1.00	1.00	1.00	1.00	1.0000

Fig5: Table for efficiency of different models

- Base Logistic Regression: Provided a baseline performance with an accuracy of 91%, F1 score of 0.74.
- Base Decision Tree: Showed good performance with an accuracy of 97%, with an F1 score of 0.97.
- Grid Search Decision Tree: Showed good performance with an accuracy of 98% an F1 score of 0.98.
- Random Forest: Exhibited robust performance with an accuracy of 98% and an F1 score of 1.
- Gradient Boosting: Achieved the best performance with an accuracy of 99% and an F1 score of 1.

Visualization of Results

The performance of the models can also be visualized through the following plots:

- Confusion Matrices: Provided for the top-performing models (Random Forest and Gradient Boosting), these matrices illustrate the true versus predicted classifications, highlighting the models' strengths and weaknesses in identifying each category.

- ROC Curves and AUC Scores: These plots demonstrate the trade-offs between sensitivity and specificity achieved by the Gradient Boosting and Random Forest models, underlining their capacity to manage type I and type II errors effectively.

Following is the Confusion Matrix and ROC curve for Gradient Boosting rest are demonstrated in the ipynb file.

Comparative Analysis

A comparative analysis reveals that while all models performed adequately, Gradient Boosting and Random Forest offered the most promising results. Gradient Boosting, in particular, provided the highest accuracy and F1 scores, which indicates its effectiveness in balancing precision and recall, a crucial aspect in the nutritional classification of foods. Random Forest also showed high scores but was slightly less effective than Gradient Boosting in terms of overall accuracy and F1 score.

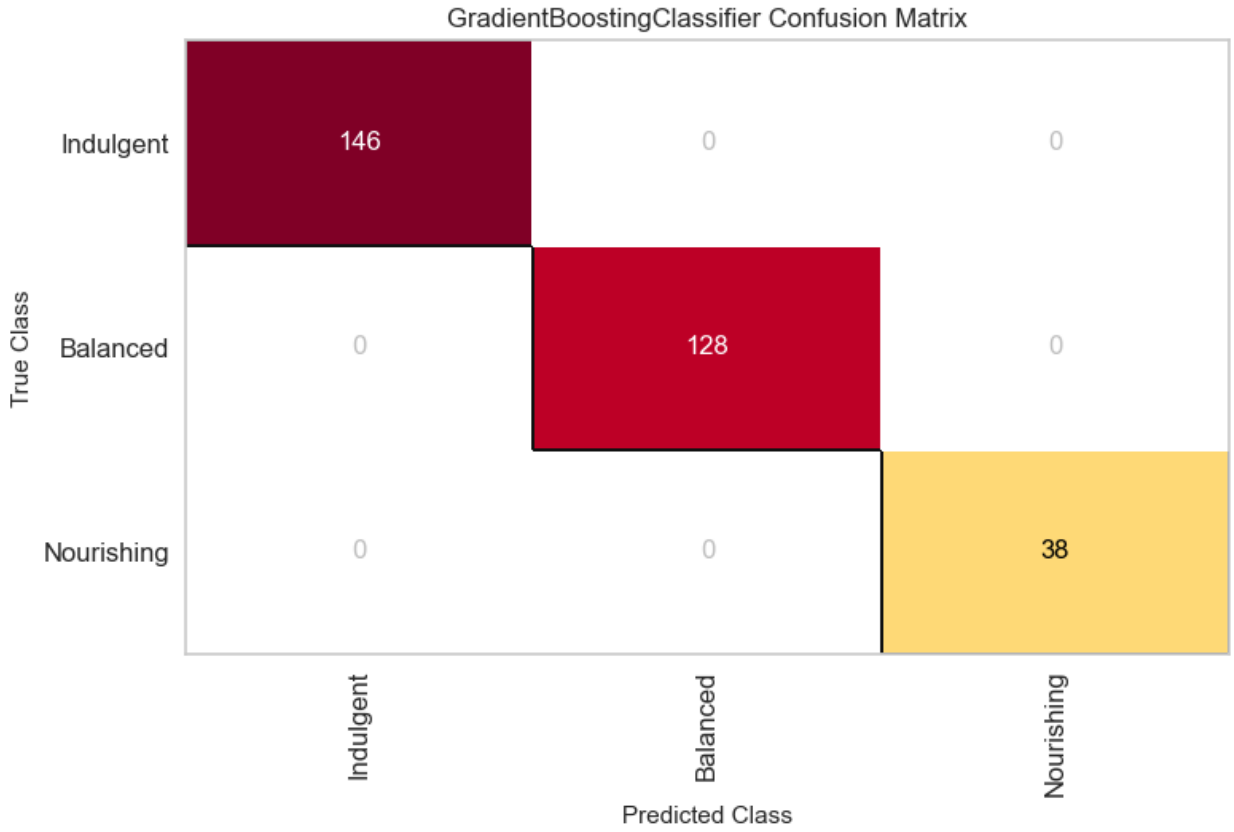


Fig 6: Confusion Matrix for Gradient Boosting Classifier

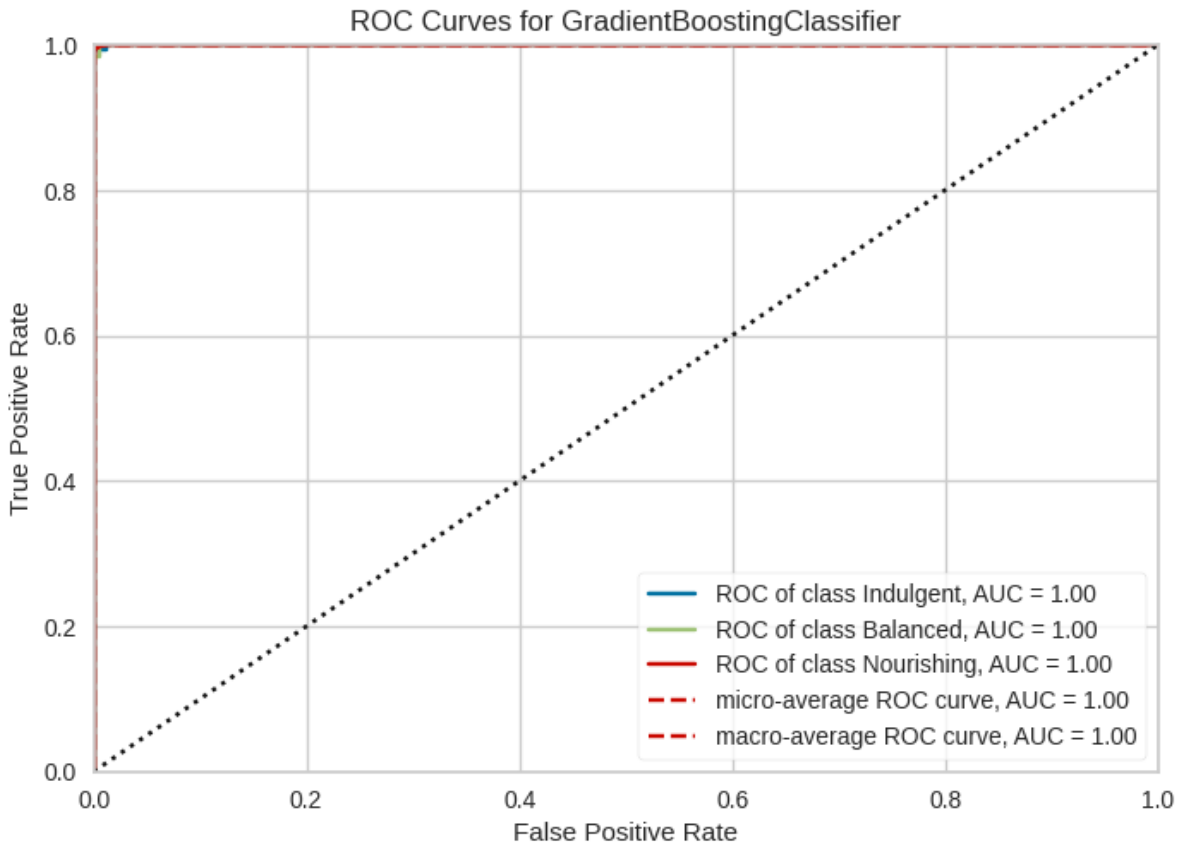


Fig 7: ROC Curves for Gadient Boosting Classifier

Significance of Findings

The findings from this study are significant in several ways:

- Model Selection: Gradient Boosting and Random Forest are validated as effective tools for food classification tasks, with the potential to be integrated into dietary recommendation systems.
- Policy Implications: These results can inform public health policies by providing a scientific basis for more nuanced dietary guidelines.
- Personalized Nutrition: The models have the potential to be used in personalized nutrition plans, helping individuals make informed dietary choices based on comprehensive food categorization.

CONCLUSION

This study set out to revolutionize dietary recommendations by applying machine learning techniques to categorize 7,793 food items into Nourishing, Indulgent, and Balanced categories based on their nutritional content. Through comprehensive exploratory data analysis and rigorous preprocessing—including handling missing values and normalizing data—this research established a robust foundation for model accuracy and reliability. The Gradient Boosting and Random Forest models emerged as particularly effective, with Gradient Boosting displaying superior performance with an accuracy of 98% and an F1 score of 0.98. These models not only validate the use of advanced computational methods in nutritional science but also enhance the precision of dietary classifications, which are crucial for developing tailored nutritional guidelines.

The implications of these findings are significant, extending from individual dietary guidance to public health policy. By offering a detailed, data-driven categorization of foods, the study provides a tool that can adapt dietary advice to individual needs, thereby supporting more nuanced and effective nutritional interventions. This approach has the potential to inform public health initiatives and aid in the combat against diet-related diseases by facilitating more scientifically grounded dietary recommendations. Moreover, the methodology and findings of this research serve as a foundational resource for nutritional advisors and health applications, potentially leading to improved health outcomes at both individual and community levels.

Looking ahead, the project identifies several avenues for further research and application enhancement. Expanding the dataset to include more diverse food items and additional nutritional factors would likely improve the model's applicability and accuracy. Future studies could explore the integration of real-time data processing techniques to develop dynamic nutritional guidance systems. Additionally, investigating the long-term health impacts of adopting diets based on these machine learning classifications could empirically validate and refine the effectiveness of the food categories proposed. By advancing these areas, subsequent research can continue to enhance the intersection of machine learning and nutritional science, driving forward innovations in personalized diet planning and public health strategies.