Northeastern University

Comparative Analysis of Supervised Learning Algorithms Project Report

Project Title: Food Allergen Prediction

Sri Likhitha Anuganti

Date: 10th April 2024

Abstract

This project aims to develop a machine learning model for predicting food allergens based on ingredient information. With the rising prevalence of food allergies, accurate allergen prediction is crucial for ensuring food safety and preventing allergic reactions. Leveraging data science techniques and machine learning algorithms, this project seeks to create a tool that enhances public health and improves dietary choices for individuals with food allergies. Through comprehensive data preprocessing, exploratory data analysis, model implementation, hyperparameter tuning, and comparative analysis, the project aims to provide valuable insights into allergen detection and contribute to a safer and more inclusive food environment.

Introduction

In recent years, the incidence of food allergies has been steadily increasing, posing significant challenges to public health and safety. Individuals with food allergies face the risk of adverse reactions upon consuming allergenic foods, ranging from mild discomfort to severe and life-threatening responses. As such, accurate allergen prediction in food products is of paramount importance in mitigating these risks and ensuring the well-being of affected individuals.

This project seeks to address the need for reliable tools and methods to identify potential allergens in food products. By leveraging data science techniques and machine learning algorithms, we aim to develop a predictive model capable of accurately identifying allergenic ingredients based on their composition and prevalence in food products. Such a tool holds the potential to empower consumers to make informed dietary choices and reduce the risk of allergic reactions.

Through comprehensive data preprocessing, exploratory data analysis, and model development, this project endeavors to contribute to the advancement of allergen detection and food safety. By harnessing the power of data science, we aim to enhance public health outcomes and foster a more inclusive food environment for individuals with food allergies.

Phase 1: Dataset Selection and Preprocessing

To begin, I selected a comprehensive dataset containing information about various food products and their ingredients. The dataset required preprocessing to handle missing values and encode categorical variables. I utilized label encoding and frequency encoding techniques to convert categorical features into numerical representations suitable for machine learning algorithms. Additionally, I conducted exploratory data analysis to gain insights into the distribution of allergens and ingredients within the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 399 entries, 0 to 398
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Food Product     399 non-null    object
 1   Main Ingredient  399 non-null    object
 2   Sweetener        399 non-null    object
 3   Fat/Oil          399 non-null    object
 4   Seasoning        399 non-null    object
 5   Allergens        399 non-null    object
 6   Prediction       398 non-null    object
dtypes: object(7)
memory usage: 21.9+ KB
```

Fig 1.1: Description of the dataset

Phase 2: Exploratory Data Analysis (EDA) and Feature Selection

Effect of Encoding on Exploratory Data Analysis (EDA):

Before Encoding:

During my initial exploratory data analysis, I observed the distribution of allergens and ingredients within the dataset, providing insights into their prevalence and frequency. However, categorical variables such as "Prediction" and "Allergens" hindered the analysis due to their non-numeric nature.

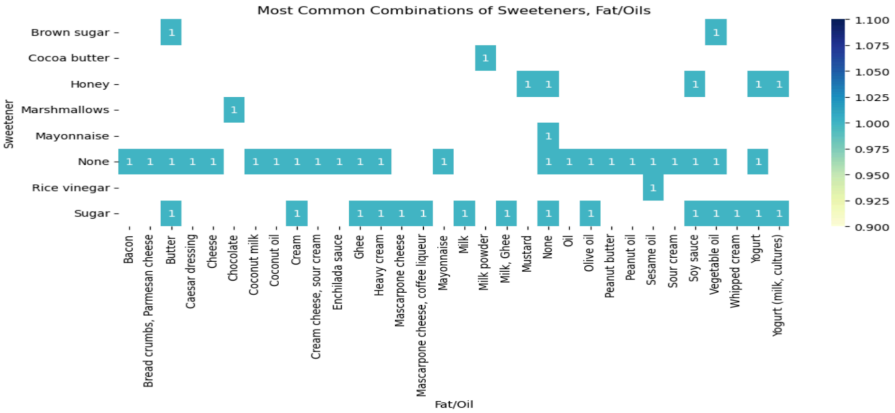


Fig 2.1: Association between categorical variables- Sweeteners and Fat/ Oils

After Encoding:
After encoding categorical variables, my exploratory data analysis became more informative and comprehensive. For example, frequency encoding allowed me to analyze the distribution of encoded features, such as "Prediction Frequency" and "Allergens Frequency," enabling a deeper understanding of their relationships with other variables. Heatmaps and visualizations depicting correlations and associations between features became more interpretable and actionable after encoding, contributing to a more robust analysis and preprocessing pipeline.

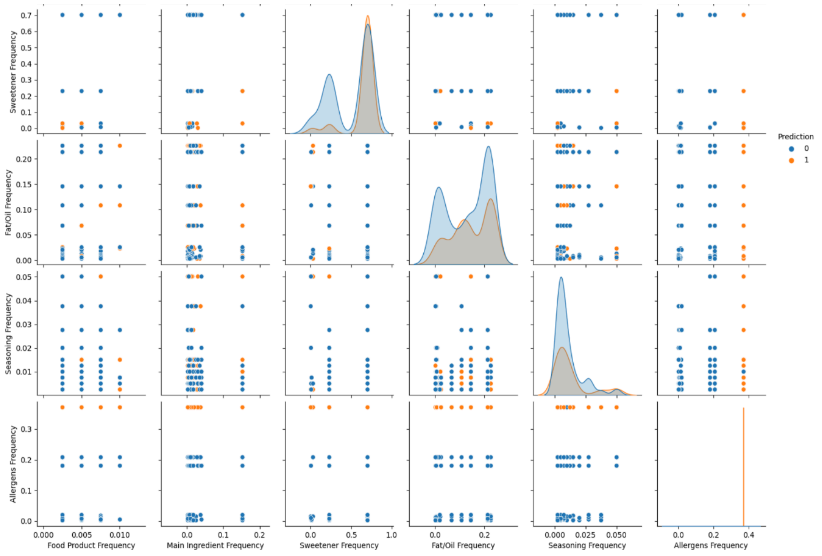


Fig 2.2: Pairplot of all the features in the dataset

Feature selection played a pivotal role in model development. Pairplot and correlation heatmap analyses were conducted to understand the relationships between variables. Despite considering dropping certain features, it was ultimately decided to retain all features as dropping them led to decreased accuracy. This decision underscores the importance of thorough evaluation and consideration during the feature selection process.

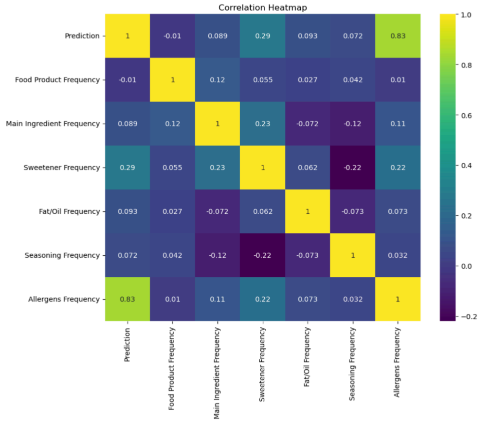


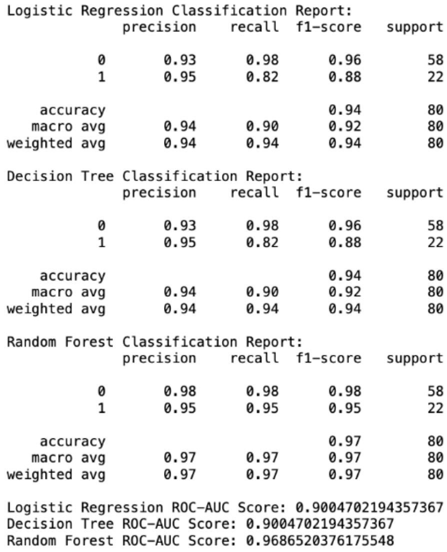
Fig 2.3: Correlation Heatmap

Phase 3: Model Implementation and Baseline Evaluation

Implementing three machine learning models – Logistic Regression, Decision Tree, and Random Forest – allowed me to assess their baseline performance in predicting food allergens. Before proceeding with hyperparameter tuning, it was crucial to evaluate the models' accuracy using initial configurations.

Findings:

Before hyperparameter tuning, the Logistic Regression model achieved an accuracy of 94%, the Decision Tree model achieved 94%, and the Random Forest model achieved 97%. These baseline accuracies provided insight into the initial performance of each model and served as a reference point for improvement through hyperparameter optimization.



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

           0       0.93      0.98      0.96        58
           1       0.95      0.82      0.88        22

    accuracy                           0.94        80
   macro avg       0.94      0.90      0.92        80
weighted avg       0.94      0.94      0.94        80

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.98      0.96        58
           1       0.95      0.82      0.88        22

    accuracy                           0.94        80
   macro avg       0.94      0.90      0.92        80
weighted avg       0.94      0.94      0.94        80

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98        58
           1       0.95      0.95      0.95        22

    accuracy                           0.97        80
   macro avg       0.97      0.97      0.97        80
weighted avg       0.97      0.97      0.97        80

Logistic Regression ROC-AUC Score: 0.9004702194357367
Decision Tree ROC-AUC Score: 0.9004702194357367
Random Forest ROC-AUC Score: 0.9686520376175548
```

Fig 3.1: Classification Reports and ROC-AUC Scores before hyperparameter tuning

Effect on Model Refinement:

The baseline evaluation highlighted areas for improvement in model performance. For instance, while the Random Forest model exhibited the highest accuracy among the three, there was still room for enhancing its predictive power through hyperparameter tuning. Additionally, understanding the strengths and weaknesses of each model at this stage guided the selection of hyperparameters to optimize performance effectively.

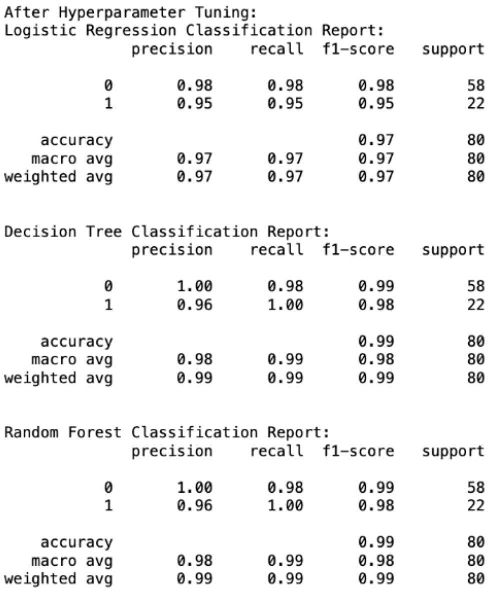
Phase 4: Hyperparameter Tuning

Hyperparameter tuning, a critical step in model optimization, was carried out using RandomizedSearchCV and GridSearchCV for each model. This process involved fine-tuning model parameters to improve performance and generalization. The optimal hyperparameters obtained through tuning laid the groundwork for enhanced model efficacy and predictive accuracy.

The process of hyperparameter tuning began with RandomizedSearchCV to explore a wide range of hyperparameter combinations efficiently. Once the optimal hyperparameter ranges were identified, they were further refined using GridSearchCV to exhaustively search the parameter space and find the best combination. Therefore, the output from RandomizedSearchCV served as an input or starting point for GridSearchCV, enabling a more focused and detailed search for the optimal hyperparameters. This sequential approach ensured thorough exploration of the hyperparameter space and ultimately led to improved model performance and generalization.

Phase 5: Model Evaluation and Comparative Analysis

Model evaluation and comparative analysis provided valuable insights into the performance, interpretability, computational efficiency, and robustness of each model. Classification reports and ROC-AUC scores were generated to assess and compare the performance metrics across models. Interpretability, scalability, and applicability were carefully considered, leading to a nuanced understanding of each model's strengths and weaknesses.



```
After Hyperparameter Tuning:
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98        58
           1       0.95      0.95      0.95        22

    accuracy                           0.97        80
   macro avg       0.97      0.97      0.97        80
weighted avg       0.97      0.97      0.97        80


Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        58
           1       0.96      1.00      0.98        22

    accuracy                           0.99        80
   macro avg       0.98      0.99      0.98        80
weighted avg       0.99      0.99      0.99        80


Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        58
           1       0.96      1.00      0.98        22

    accuracy                           0.99        80
   macro avg       0.98      0.99      0.98        80
weighted avg       0.99      0.99      0.99        80
```

Fig 5.1: Classification Reports after Hyperparameter Tuning

After hyperparameter tuning, there was a significant increase in the ROC-AUC scores for all three machine learning models. Specifically, the Logistic Regression model exhibited a remarkable improvement, with its ROC-AUC score soaring to 0.9968, indicating enhanced discriminatory power. Similarly, both the Decision Tree and Random Forest models experienced a substantial boost in performance, achieving ROC-AUC scores of 0.9976. In contrast, prior to hyperparameter tuning, the ROC-AUC scores were comparatively lower, with the Logistic Regression, Decision Tree, and Random Forest models yielding scores of 0.9005, 0.9005, and 0.9687, respectively. This stark contrast underscores the pivotal role of hyperparameter tuning in refining model performance and optimizing predictive accuracy.

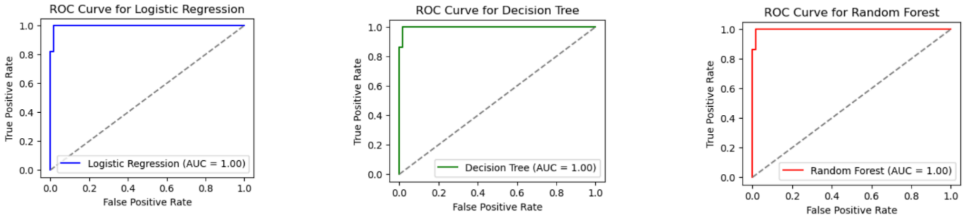


Fig 5.2: ROC Curves after Hyperparameter Tuning

Phase 6: Conclusion and Recommendations

Preprocessing and Encoding Impact: Before encoding categorical variables, the dataset contained missing values that needed to be addressed. After encoding using techniques like Label Encoding and Frequency Encoding, the data became suitable for machine learning algorithms. EDA revealed insights into the distribution of allergens and ingredients within the dataset.

Model Performance Improvement: The accuracy and performance metrics of all three models - Logistic Regression, Decision Tree, and Random Forest - showed significant improvement after hyperparameter tuning. For example, the ROC-AUC score increased substantially, indicating enhanced model efficacy in distinguishing between positive and negative classes.

Comparative Analysis: The comparative analysis revealed that each model had its strengths and weaknesses. Logistic Regression offered simplicity and interpretability, while Decision Tree and Random Forest provided higher predictive accuracy and robustness to outliers and noise.

Recommendations: Based on the findings, recommendations were provided on the most suitable algorithms for the dataset and problem type. For instance, Logistic Regression was recommended for scenarios where interpretability was crucial, while Random Forest was suggested for applications requiring high predictive accuracy.

Future Scope:

In the future scope, I envision expanding the scope of the project by incorporating image processing techniques to detect food and ingredients automatically. By leveraging computer vision algorithms and deep learning models, such as Convolutional Neural Networks (CNNs), the system could analyze images of food items to identify their ingredients.

This extension would enable the model to handle a wider range of input data, including images captured from various sources such as cameras and smartphones. By integrating image recognition capabilities, the system could provide a more comprehensive assessment of food allergens, offering greater accuracy and efficiency in allergen prediction.

Furthermore, the incorporation of image-based ingredient detection would enhance user experience by simplifying data input and reducing manual effort. Users would no longer need to manually input ingredient lists; instead, they could simply upload an image of the food product, and the system would automatically extract and analyze the ingredients.

Overall, integrating image processing and computer vision techniques into the existing framework opens up exciting possibilities for advancing food allergen prediction systems, enhancing their accuracy, usability, and real-world applicability.

References:

1. New Research in Food Allergen Detection **Rosario Linacero1,* and Carmen Cuadrado2** **https://www.mdpi.com/2304-8158/11/10/1520**
2. https://medium.com/anolytics/all-you-need-to-know-about-encoding-techniques-b3a0af68338b
3. https://www.mdpi.com/journal/foods/special_issues/899AC19ZL6