Predicting Recipe Health Scores: A Machine Learning Approach

Osman Bulut[1]

[1]Electrical and Computer Enginerring, State University of NewYork at Albany, NY, USA

November 2023

Abstract

- **Background** In an age where individuals are increasingly health-conscious and the internet serves as a vast resource for culinary inspiration, this project addresses a fundamental question: the healthiness of online recipes. Leveraging machine learning, feature engineering, and data processing, we aim to predict the health scores of internet recipes based on their ingredient information. The study focuses on machine learning for prediction, feature engineering for model enhancement, and healthiness evaluation in internet recipes, emphasizing unit normalization and text vectorization for data consistency and accuracy. This work intends to offer a valuable tool for algorithmic meal recommendations and informed dietary choices within the realm of online culinary diversity.
- **Objective** This paper has a threefold objective. Firstly, it seeks to create structured data by amalgamating the RecipeKG[3] knowledge graph with additional web scraping from allrecipes.com and employing feature engineering, unit normalization, and ingredient vectorization. Secondly, it endeavors to apply Classification and Regression techniques to the data prepared within this study for the purpose of predicting the health scores of online recipes. This predictive model can serve as a benchmark for potential food recommendation systems. Lastly, this paper strives to underscore the significance and influence of ingredient usage amounts on food healthiness and the accuracy of ML analysis in this domain.
- **Methods**

  This study leverages the publicly available RecipeKG knowledge graph, comprising 77,835 recipes from Allrecipes.com, to investigate and predict the health scores of online recipes. Two health score systems, USDA and FSA, have been employed, using internationally recognized dietary guidelines. Additional data, particularly the 'number of servings,' was obtained through web scraping from allrecipes.com, and unit normalization was applied to enhance data consistency. The unit conversion process was performed, aiming to standardize units into "cup," "ounce," or "piece." Furthermore, ingredient vectorization was conducted through various hot encoding techniques, facilitating the alignment of ingredients with machine learning models for health score predictions.
- **Results**

  In our comprehensive machine learning analysis, we tackled the challenging task of predicting USDA and FSA health scores for recipes, employing a range of classification and regression models. To enhance the evaluation of our models, we introduced alternative accuracy metrics, including "1 error included accuracy" and "2 error included accuracy," which considered the proximity of predicted scores to actual scores. Remarkably, the Histogram-Based Gradient Boosting model, especially when using Value Hot Encoding, outperformed other classifiers with an accuracy rate exceeding 93.3%, emphasizing the importance of incorporating ingredient usage amounts. In the regression tasks, Histogram-Based Gradient Boosting with Value Hot Encoding achieved the best results, with a 69.4% R-squared score and 0.54 MAE for USDA(6 scores), and a 69.0% R-squared score with 0.74 MSE for FSA(9 scores).
- **Conclusion**

  In conclusion, this study provides a more nuanced approach to assessing the nutritional quality of recipes, demonstrating the potential to predict meal healthiness with greater accuracy by analyzing individual ingredients. The development of open-source datasets and machine learning models aims to empower users to make informed dietary choices and contribute to ongoing research in this field. The significance of considering ingredient usage amounts has been highlighted, emphasizing the potential for more informed culinary decisions and healthier lifestyles.

Keywords— web and public health, feature engineering, text vectorization, machine learning, classification, regression, hot encoding, online recipes

1 Introduction

1.1 Background and Motivation

In an era where individuals are increasingly conscious of their dietary choices, the internet has become a treasure trove of culinary inspiration. Recipes abound on websites and platforms, offering a wealth of options to satisfy diverse palates and dietary preferences. However, in this sea of gastronomic delights, one critical question often lingers: how healthy are these recipes?

This project embarks on a journey to answer this question by harnessing the power of machine learning, feature engineering, and data processing techniques. We delve into the realm of recipe healthiness, aiming to predict the health scores of internet recipes based on their ingredient information.

The health scores utilized in this study are calculated by evaluating the nutritional information of recipes. This information, vital for health-conscious consumers, serves as the foundation for our predictive model. As we explore the nuances of healthiness in internet recipes, we navigate various challenges, including unit normalization, text vectorization, and effective feature engineering.

The key objectives of this project are as follows:

Machine Learning and Prediction: We employ machine learning algorithms to create a predictive model capable of assigning health scores to internet recipes based on their ingredients. This model contributes to a better understanding of the nutritional aspects of recipes and their impact on health.

Feature Engineering: To enhance the performance of our prediction model, we engage in feature engineering, a critical process in selecting, transforming, and creating features from the available data. This optimization aids in the accurate prediction of recipe health scores.

Healthiness of Internet Recipes: We explore the concept of healthiness in the context of internet recipes and establish a framework for evaluating it. By doing so, we provide a foundation for further research in the field of culinary health.

Unit Normalization and Text Vectorization: Achieving consistency in ingredient units is a complex yet crucial task. We normalize units, ensuring that ingredients are accurately measured, and leverage text vectorization techniques to extract meaningful information from ingredient descriptions.

This project aims to fill a void in the realm of semantic-based algorithmic meal plan recommendation and individual ingredient substitution that explicitly incorporates healthiness into the recommendation process. By providing a reliable system for evaluating the healthiness of internet recipes, we empower individuals to make informed dietary choices while exploring the delightful world of culinary diversity.

1.2 Related Studies

The evaluation of the healthiness of online recipes using machine learning models has been explored in a limited number of previous studies. [11] compares algorithmic nutritional estimation with human-provided estimates, analyzing the most influential features used by humans and machines to determine the healthiness of online recipes. In the context of our research, this study involves the prediction of FSA health scores based on the composition of ingredients in recipes.

[10] introduces a novel machine learning pipeline designed for fast prediction of nutrient values from unstructured recipe text. The pipeline utilizes domain-specific embeddings and incorporates external domain knowledge for clustering and model training, ultimately achieving significantly improved accuracy compared to baselines. Additionally, the importance of data in predictive modeling is highlighted, emphasizing the need for a representative training dataset that covers expected variations in deployment data.

Lately, there has been a growing focus on incorporating health considerations into food recommendation systems[7][12]. [9] mobile recommender system addresses the challenge of promoting healthier food choices amid evolving diets and lifestyles, integrating user preferences and health considerations to provide recipe recommendations, and this paper presents the human-computer interaction design, health-aware recommendation algorithm, and initial user feedback. [13] introduces a comprehensive framework for daily meal plan recommendations, offering a unique approach that simultaneously considers nutritional and user preference aspects, integrating multi-criteria decision analysis and optimization techniques, and demonstrates its effectiveness through a case study.

2 Methods

2.1 Dataset

2.1.1 RecipeKG

This study is predominantly based on the publicly available knowledge graph known as RecipeKG[4], which encompasses 77,835 recipes sourced from Allrecipes.com[3]. Each entry in the RecipeKG dataset includes critical information, such as Recipe Name, Ingredient Names, Units, Total Quantity, and Nutritional Information. The dataset also incorporates two health scores, each of which is derived from internationally recognized standards for evaluating the healthiness of meals. These standards comprise the 'Dietary Guidelines for Americans,' established by the United States Department of Agriculture (USDA)[5], and the 'Guide to creating a front of pack (FoP) nutrition label,' provided by the United Kingdom Food Standards Agency (FSA)[8].

Consequently, the combined dataset, consisting of RecipeKG data and the scraped 'number of servings' data, serves as the foundation for this study's analysis."

For your reference, we summarize the methods used to calculate the two health scores:

- **USDA Score.** The USDA[5] score is determined based on the presence of seven key macronutrients: carbohydrates, protein, fat, saturated fat, sugar, sodium, and fiber. To create this score, the permissible content in grams for each macronutrient, as a percentage of daily energy intake (e.g., 2,000 calories), is calculated. For each recipe, a point is awarded for each macronutrient for which the recommended content is met. The final score ranges from 0 (indicating an entirely unhealthy meal) to 7 (representing a highly nutritious option).
- **FSA Score.** The FSA[8] Score is designed to generate a "traffic light labeling" score, akin to the three-color-coded nutrition labels found on the front of food packaging, which helps consumers quickly grasp the nutritional value. It uses integer values to rate the levels of fat, saturated fat, sugars, and salt: 2 for green (low), 1 for amber (medium), and 0 for red (high). The scores for each macronutrient are summed to derive a final score, which can range from 0 (very unhealthy) to 8 (very healthy) for each recipe.

It's important to note that the daily recommended values of nutrients vary depending on factors like gender and age[6], which are taken into account in both USDA and FSA calculations. In this study, the calculations use 19-30 years old females as the reference point.

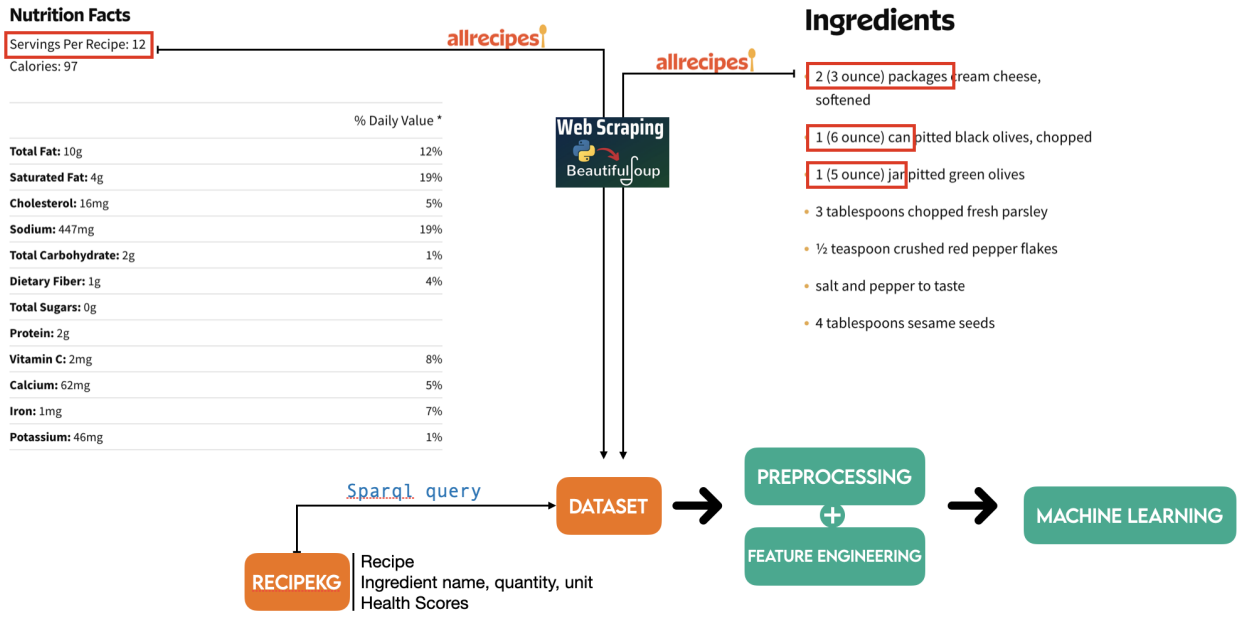

Figure 1: Data Flow and Processing Overview

2.1.2 Web Scraping from allrecipes.com

- **Number of Serving:** In the context of this study, RecipeKG[3] serves as the primary data source. However, the creation of the dataset for machine learning (ML) applications necessitated additional web scraping steps.

The first and foremost of these was the acquisition of "number of serving" information, which was notably absent from RecipeKG.

This particular information is of paramount importance since it corresponds to the quantity of ingredients used in an entire recipe. In contrast, the provided nutritional data is structured for a single serving. This is significant because both the USDA and FSA health scores are predicated on this specific nutritional information.

To ensure data consistency and enable predictive nutritional analysis with respect to the ingredients' usage amounts and health scores, all ingredient quantities had to be normalized by serving size. To achieve this, we conducted web scraping on Allrecipes.com, obtaining the "number of serving" data for each of the 77,537 recipes encompassed within RecipeKG presented in Figure 1.

It is noteworthy that this meticulous process was conducted not only for the purposes of this study but also with an eye toward potential future research initiatives led by different researchers. This additional data was subsequently integrated with the RecipeKG dataset, thus enhancing its comprehensiveness and suitability for our analysis.

- **Conversion from 'Package,' 'Can,' and 'Jar' Units to 'Ounce' Units:** In this study, we emphasize the significance of unit conversion and normalization, which we will delve into in greater detail in subsequent sections. In this particular step, we have already addressed this issue for three significant units: 'Package,' 'Can,' and 'Jar.' These units are inherently ambiguous, with actual amounts varying from one ingredient to another. This ambiguity can potentially affect numerous other units as well. However, the advantage with these three units is that they also have ounce information available on allrecipes.com which also can be seen in Figure 1. Consequently, using their ounce values ensures a higher level of consistency and eliminates ambiguity when dealing with usage amounts. These three units represent a significant portion of the dataset as can be seen in Figure 2, accounting for approximately 27% of the entire dataset. To accomplish this, we obtained the URL information for recipes utilizing these units from RecipeKG and proceeded to scrape the ounce information for these units by visiting each corresponding URL. We utilized the BeautifulSoup 4 library for this purpose. Subsequently, this supplementary data was seamlessly integrated into the RecipeKG dataset, rendering it prepared for the final preprocessing steps before implementing machine learning applications.

cup	119163
teaspoon	67270
piece	56896
tablespoon	49923
can	10056
pound	9428
package	7643
clove	6469
pinch	4137
ounce	3560
head	1042
bunch	903
jar	817
dash	638
quart	445
pint	392
stalk	339
sprig	320
loaf	268
slice	227

Figure 2: Most Used 20 Units in The RecipeKg Data

2.2 Data Preproccesing

2.2.1 Initial Steps

In this section, we will explore the crucial data preprocessing steps undertaken to ensure the quality and reliability of the dataset for subsequent machine learning analysis.

- **Conversion to Lowercase:**

In the original dataset, ingredients appeared in various formats, with discrepancies in capitalization. For example, "Sugar," "SUGAR," and "sugar" were treated as distinct ingredients. To address this issue, all strings were converted to lowercase, facilitating uniformity and consistency when dealing with ingredients.

- **Truncation of Recipes without Serving Information:**
  Serving information is of paramount importance as it underpins the calculation of both health scores, USDA and FSA. Nutritional data in the dataset and on allrecipes.com pertains to a single serving. However, ingredient quantities in the dataset encompass the entire recipe, and serving sizes vary across recipes. To facilitate ingredient-based machine learning analysis, all ingredient quantities must be normalized by serving size. Recipes lacking serving size information were subsequently removed from the dataset.
- **Conversion of All Numbers to Float:**
  The dataset contains numerical values represented as strings, particularly in the case of serving sizes and quantities. These values, often expressed as mixed fractions or combinations like "3 1/2 cups," were converted into a numerical data type—specifically, float—using a custom method.
- **Elimination of Recipes with Rare Ingredients:**
  In the dataset, there are approximately 6,000 unique ingredients, and a substantial portion of them is used infrequently, with nearly 3,000 ingredients occurring only once. Conversely, some ingredients, such as "salt," are heavily prevalent, appearing in around 32,000 recipes. The dataset's substantial variability leads to a significant amount of noise and numerous outliers. Roughly 400 unique ingredients are featured in over 93
- **Resolution of Data Inaccuracies:**
  The RecipeKG dataset contained instances of misinformation. For instance, some ingredient and unit combinations were incorrect; for instance, "black pepper" paired with the unit "can" should have been "black bean" instead. The same issue arose with "green pepper" and "green bean." To rectify these inaccuracies, conditional statements were employed, referencing previous data and corresponding allrecipes.com URL addresses. The code for these corrections can be found in the study's GitHub repository.
  6. Removal of Recipes with Uncommon Units
  The dataset exhibited a degree of noise, featuring recipes with units that were not meaningful in the context of RecipeKG, such as "frozen" or "mini." Recipes utilizing these non-standard units were excluded to enhance dataset consistency.
- **Elimination of Recipes with Missing Quantity Data:**
  For specific ingredients like salt, knowing the quantity is vital as sodium significantly influences health scores. An accurate sodium estimation, and subsequently, the overall health score, hinges on having precise quantity data. Therefore, recipes with missing quantity values were removed from the dataset to ensure the reliability of nutritional analysis.
- **Assignment of "Piece" Unit:** Ingredients with a specified quantity but missing unit information were assigned the unit "piece." This applied to ingredients like "1 onion" or "3 eggs," where the quantity was provided but the unit was absent.

In summary, these data preprocessing steps are essential to facilitate meaningful machine learning analysis and accurate health score predictions. By enhancing dataset quality, consistency, and reliability, we pave the way for more robust and accurate insights into the healthiness of internet recipes.

2.2.2 Unit Conversion

In this research, one of the primary challenges encountered was the standardization of units. Units and quantities (e.g., 2.5 teaspoons) play a critical role in accurately estimating the health scores of recipes. However, -despite the initial steps of preprocessing- the dataset contained a wide variety of 31 unique units (even after the first steps of preprocessing) that needed to be normalized to facilitate the application of machine learning techniques.

Although three of the significant units (i.e., package, can, jar) were previously converted to "ounce" units, as depicted in Figure 2, there remained numerous ambiguous units, such as "head" and "bunch."

To address this, the first step involved converting volume units to a common standard, whereby conversions such as 1 cup = 1/16 tablespoon = 1/48 teaspoon were applied, as referenced in [2] and [1]. Additionally, mass units were converted to one another, with 1 ounce equaling 1/16 pound. Subsequently, all units were uniformly converted to "cup" or "ounce."

Furthermore, it was observed that some units were predominantly associated with specific ingredients. For instance, the unit "clove" was predominantly used in the context of "garlic," where 1 clove of garlic equated to 0.2 ounces, as outlined in [2]. Following a comprehensive analysis of each unique unit, certain commonly used ingredients were identified and converted to either "cup" or "ounce." However, this approach was not feasible for all cases,

and those that could not be reliably converted to "cup" or "ounce" were excluded from the dataset. Notably, the "piece" unit was retained as it represented a crucial element within the dataset and was inherently ambiguous in structure, rendering it challenging to standardize (for instance, the distinction between 1 piece of onion and 1 piece of watermelon).

Consequently, through the aforementioned methods, all units were successfully converted into one of three specific units: "cup," "ounce," or "piece." This standardization has effectively reduced the dataset to three distinct unit types. Any further normalization considerations will be addressed during the ingredient vectorization phase.

2.2.3 Vectorization of Ingredients

Ingredient vectorization is a crucial step in our study as it involves the transformation of textual ingredient descriptions into numerical representations, enabling the application of machine learning techniques. This numerical conversion is necessary to perform mathematical operations on ingredients and to align them with machine learning models. Essentially, vectorization empowers machine learning algorithms to process and analyze ingredient data, a vital component for tasks such as predicting the health scores of recipes based on their ingredient compositions.

In our model, we utilize hot encoding in three different variants, as illustrated in Figure 3.

Firstly, for comparison purposes, we employ traditional one-hot encoding. This results in a matrix with 423 columns, each representing a unique ingredient, and 33,523 rows, each corresponding to a unique recipe, as depicted in Figure 4. In this approach, a value of 1 signifies the presence of an ingredient in a recipe, while 0 indicates its absence. Notably, this form of encoding considers the ingredients themselves rather than their usage amounts. It serves to highlight the significance of ingredient quantities in health prediction analyses.

Secondly, concerning ingredient quantities, following the conversion to three distinct units, we employ two variations of hot encoding. In the first approach, we create a unique column for each combination of unit and ingredient. For instance, columns are labeled as "milk in oz," "milk in piece," and "milk in cups," resulting in a total of 1,269 columns (423 ingredients multiplied by 3 units). However, as not all ingredients use all units, we end up with 1,049 distinct columns, as shown in Figure 3. Furthermore, instead of using 1s, this method incorporates the actual usage amounts, as visualized in Figure 4. Like one-hot encoding, a value of 0 indicates that the ingredient is not used.

The other method involves a conversion between the three remaining units. This approach, while not entirely consistent and requiring big approximations, uses a conversion function (e.g., 1 cup = 1/2 piece = 1/8 ounce) to convert "piece" and "ounce" units to the "cup" unit. Consequently, this method results in a single unit representation. In this case, the one-hot encoding matrix is 33,523 rows by 423 columns, where the values represent quantities when an ingredient is used and 0 when it is not, as illustrated in Figure 3.

In summary, these three encoding methods, along with their respective matrices as shown in Figure 4, are prepared for use in classification and regression applications.

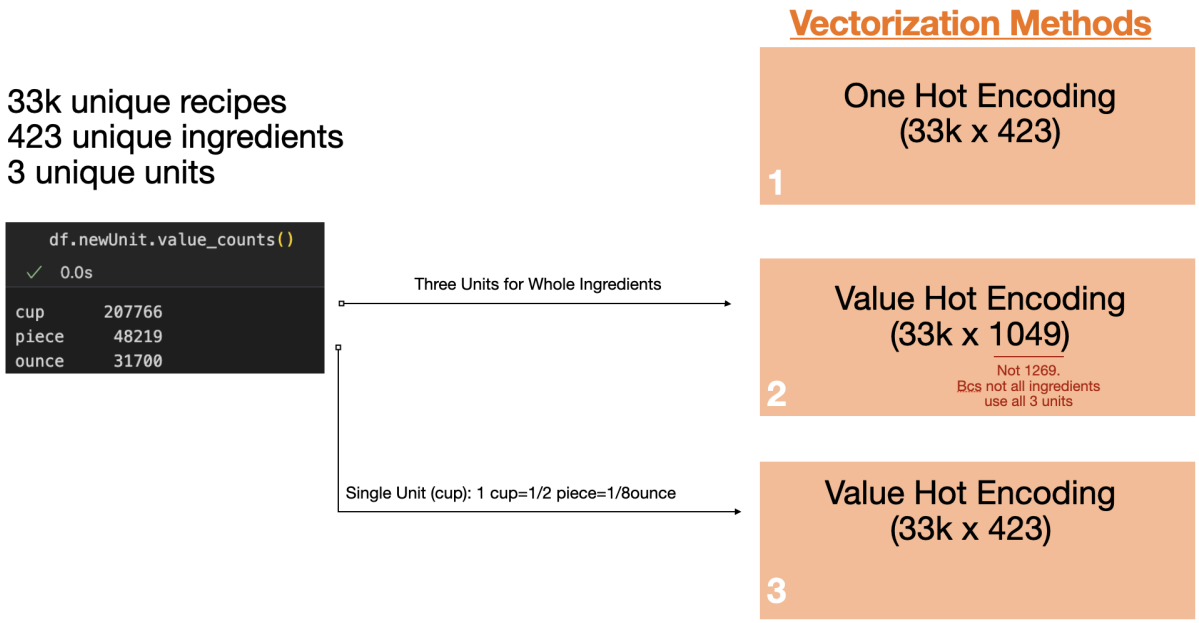

Figure 3: Two different usage of three unique units and one hot encoding

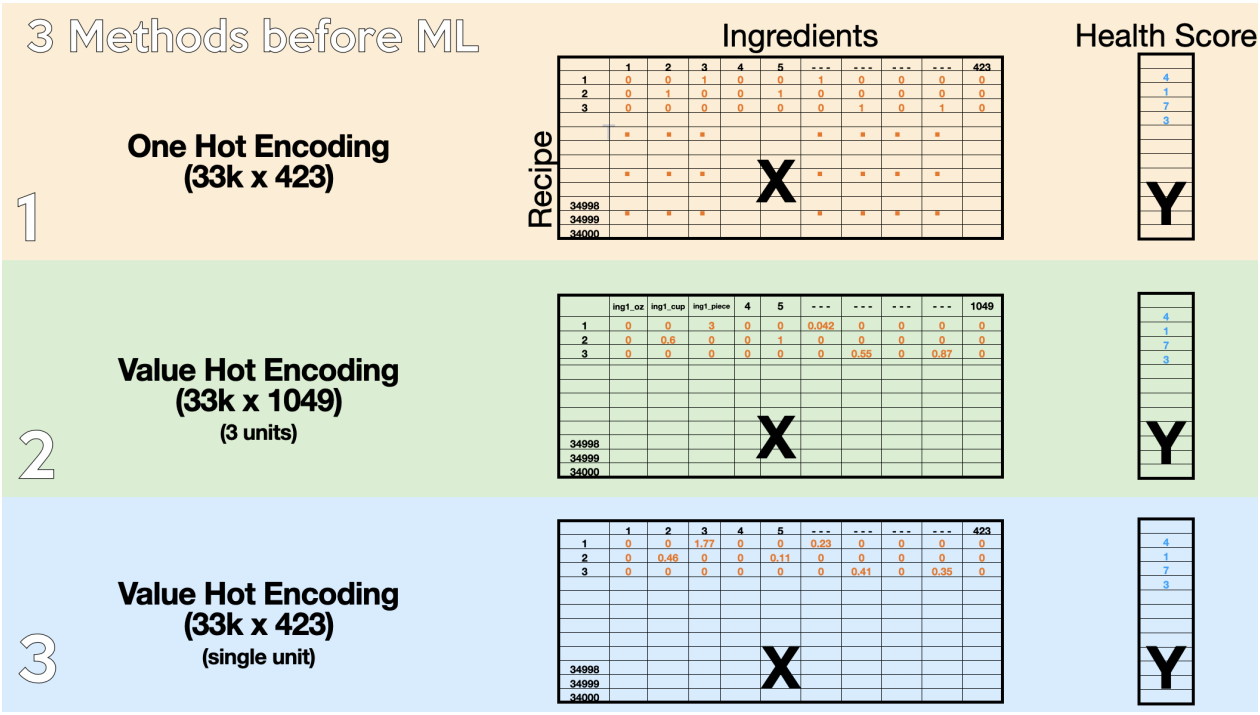

Figure 4: The structure of three different data matrix for ML applications

3 Results

In our machine learning analysis, we employ three distinct matrices, which are illustrated in Figure 4. Our target variables encompass USDA health scores, which consist of 6 classes within a range of 1 to 6, and FSA scores, involving 9 classes with a range of 0 to 8. For the purpose of classification accuracy evaluation, basic accuracy metrics are employed. However, it is acknowledged that basic accuracy alone may not be the most suitable measure for assessing the performance of a health score prediction model. To address this, we introduce alternative accuracy metrics that consider the proximity of the predicted health scores to the actual scores.

Specifically, in addition to conventional accuracy, we include two supplementary accuracy metrics in our result tables. The first is the "1 error included accuracy," which allows for a prediction to be deemed correct if the model's score is within 1 point of the actual score. For instance, if the actual health score is 5, the model's predictions of 4, 5, or 6 would all be considered correct. Essentially, this approach reduces the number of USDA classes from 6 to 4, facilitating a more lenient evaluation. Similarly, for FSA scores, we employ the "2 error included accuracy," permitting predictions within 2 points of the actual score to be deemed correct. This effectively reduces the number of FSA classes from 9 to 5, thereby accommodating the inherent complexity of the FSA scoring system.

We introduce these alternative metrics to provide a more comprehensive evaluation of the models. The result tables include basic accuracy, 1 error accuracy, and 2 error accuracy, affording flexibility in assessment. Furthermore, we employ truth tables for both classification and regression models to enable a more detailed evaluation of the predictive models.

3.1 Classification

In our classification tasks, we utilized several machine learning algorithms, including Logistic Regression with l2 penalty and lbfgs, Random Forest with Gini Impurity, and Histogram-Based Gradient Boosting with a learning rate of 0.2. Additional models such as AdaBoost, Decision Tree, SVM, and MLPC were also experimented with, although their results did not meet our expectations and are therefore not presented here.

The results of Logistic Regression, Random Forest, and Histogram-Based Gradient Boosting are illustrated in Figures 5 to 10. Remarkably, Histogram-Based Gradient Boosting emerged as the top-performing model for classification, as demonstrated in Figures 9 and 10. With a classification threshold of 1 point proximity for USDA and

2 point proximity for FSA, the Histogram-Based Gradient Boosting Classifier achieved an accuracy rate exceeding 93.3% when employing Value Hot Encoding.

A notable observation in the results is the significance of incorporating ingredient usage amounts. The accuracy score of Value Hot Encoding surpasses that of One Hot Encoding by approximately 10-15%. This underscores the importance of collecting serving-per-recipe information and incorporating it into the dataset, as it significantly enhances the predictive capabilities of the analysis.

The Feature Importance Chart for the Random Forest Classifier and the Correlation Map of the features are both displayed in Figure 11.

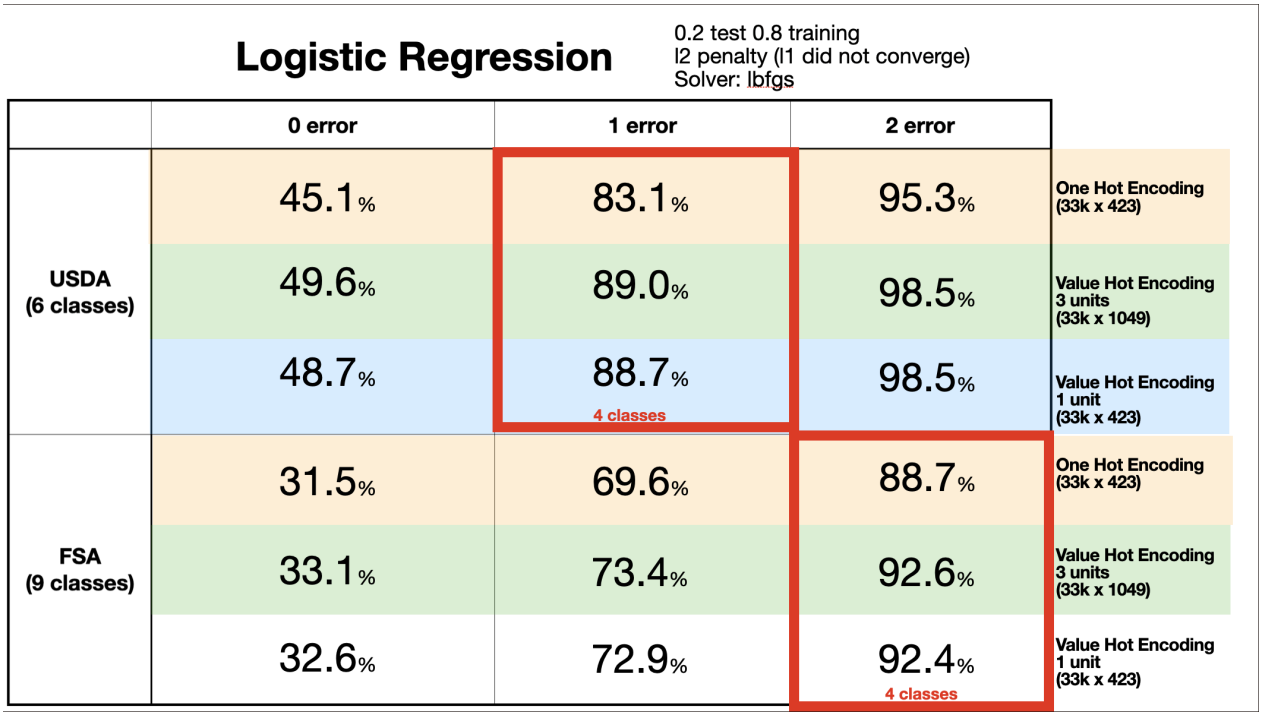

Figure 5: Logistic Regression Accuracy Scores

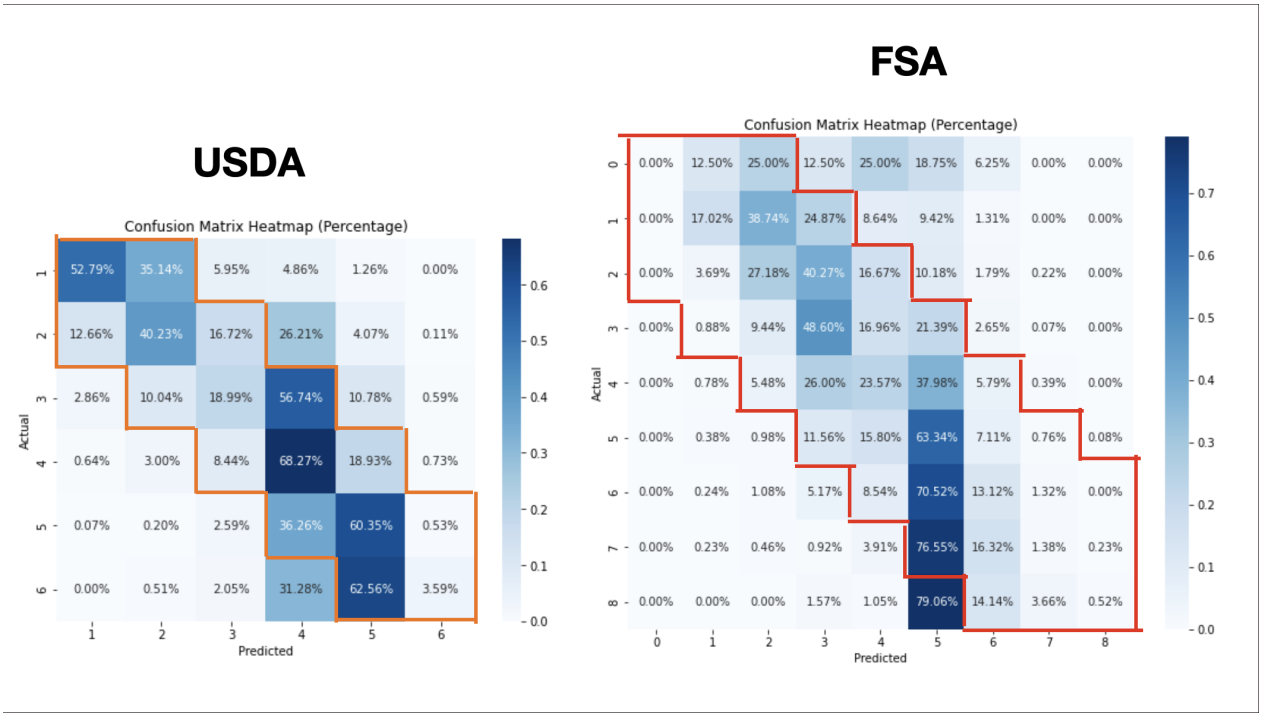

Figure 6: Logistic Regression Truth Table Heat Map

Random Forest

0.2 test 0.8 training
Gini impurity (> entropy or logloss)

	0 error	1 error	2 error	
USDA (6 classes)	45.4%	82.6%	94.8%	One Hot Encoding (33k x 423)
	56.7%	91.5%	99.0%	Value Hot Encoding 3 units (33k x 1049)
	57.5%	91.7% 4 classes	99.2%	Value Hot Encoding 1 unit (33k x 423)
FSA (9 classes)	30.4%	66.8%	87.3%	One Hot Encoding (33k x 423)
	40.1%	79.3%	94.5%	Value Hot Encoding 3 units (33k x 1049)
	41.4%	79.8%	94.5% 4 classes	Value Hot Encoding 1 unit (33k x 423)

Figure 7: Random Forest Accuracy Scores

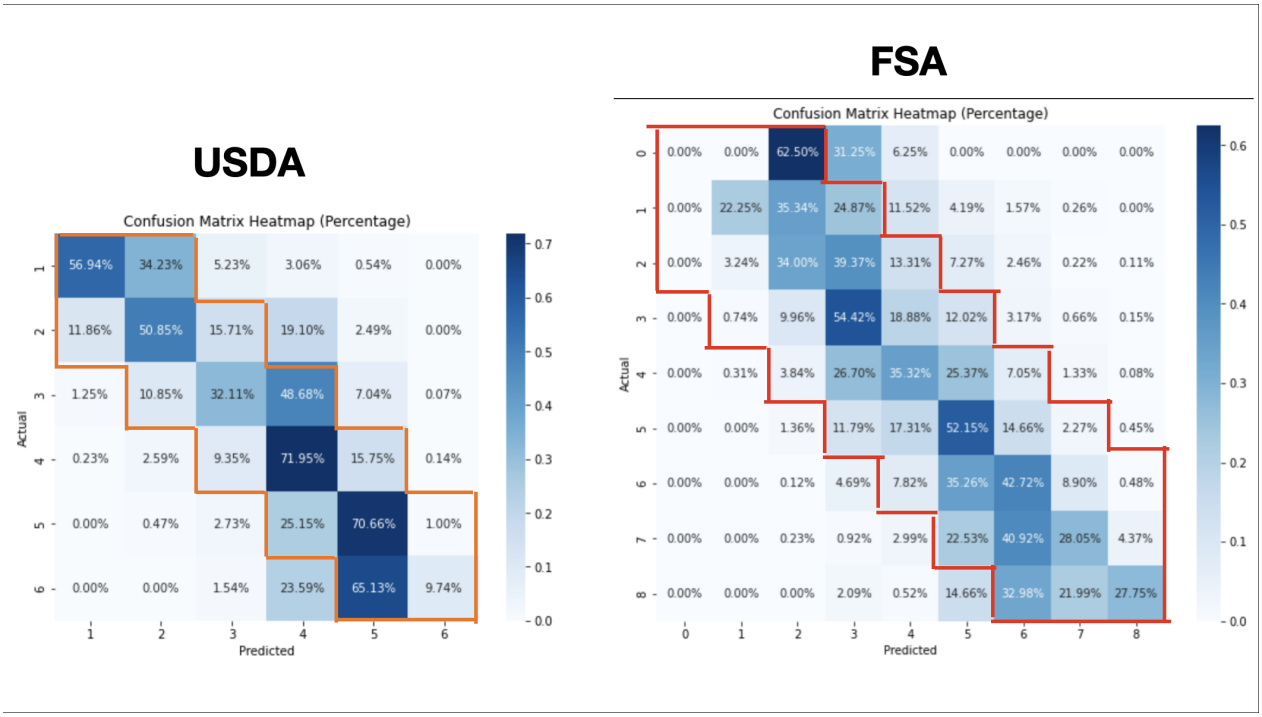

Figure 8: Random Forest Truth Table Heat Map

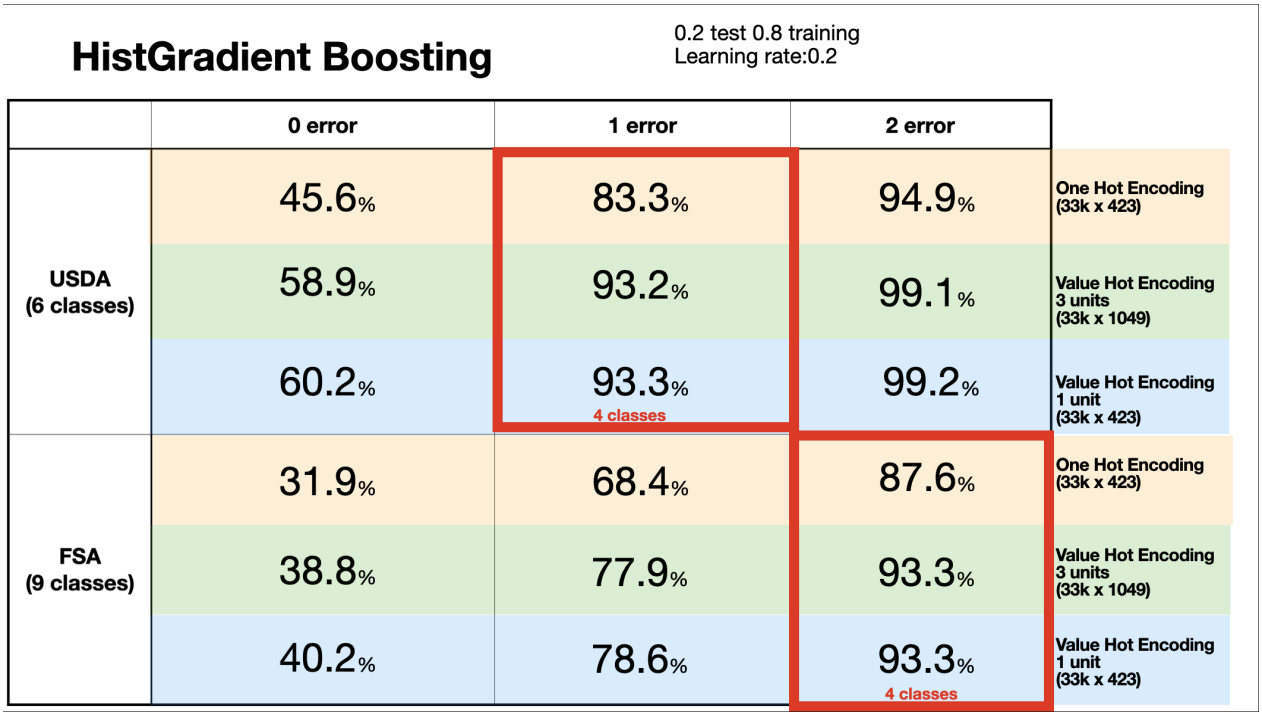

Figure 9: Histogram Based Gradient Boosting Accuracy Scores

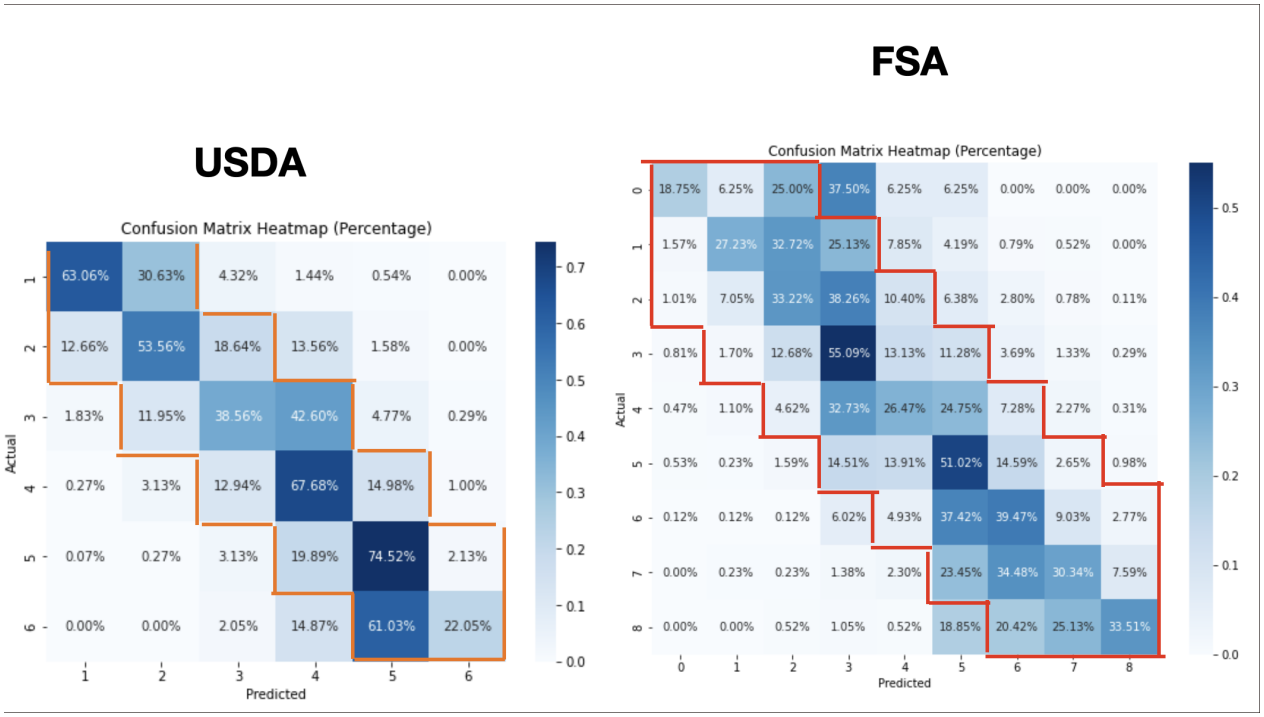

Figure 10: Histogram Based Gradient Boosting Truth Table Heat Map

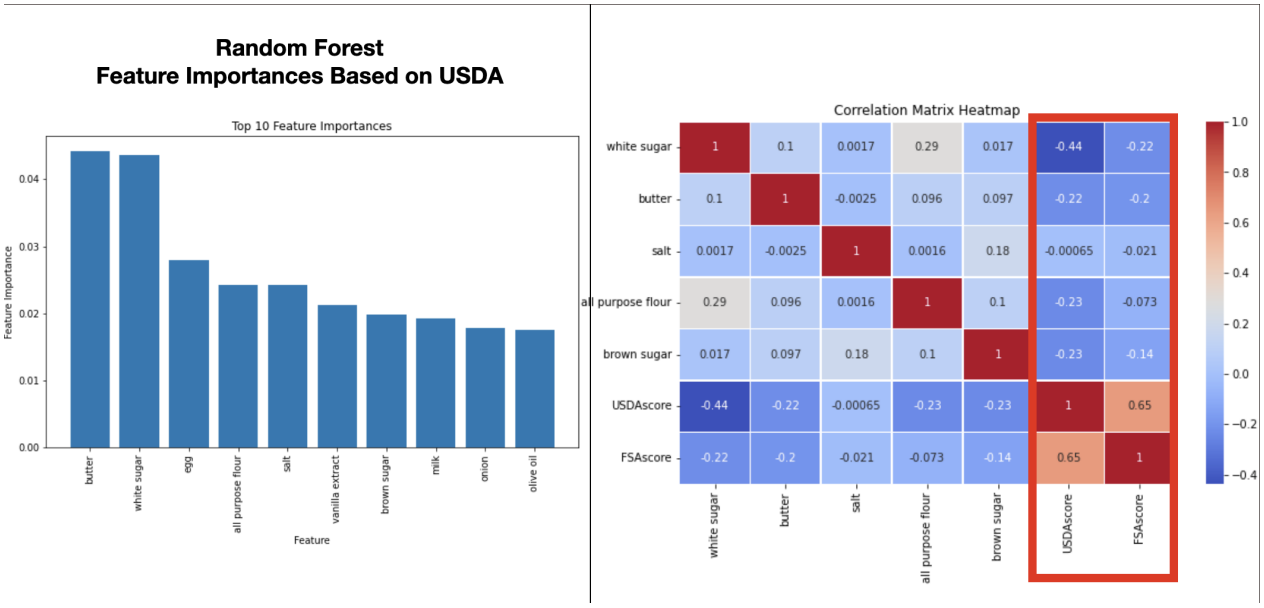

Figure 11: Feature Importance Chart for Random Forest and Correlation Map of Features

3.2 Regression

In our regression tasks, we employed various machine learning algorithms, including Random Forest Regression with squared error loss, Keras Sequential with ReLU activation function and Adam optimizer. We also explored additional models such as Support Vector Regressor, Ridge, and Lasso, though their results did not meet our expectations and are thus not presented in this section. For evaluation, three metrics were utilized: R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). As depicted in Figure 12, Histogram-Based Gradient Boosting with Value

Hot Encoding achieved the best results, attaining a 69.4% R-squared score and 0.54 MAE for USDA and a 69.0% R-squared score and 0.74 MSE for FSA.

Furthermore, similarly to the classification models, the significance of ingredient usage amounts and the impact thereof on regression models are evident when comparing the R-squared scores of one-hot encoding and Value Hot Encoding.

As a final point, we explored using nutrient values as the target variable in the Histogram-Based Gradient Boosting regressor, and the results are presented in Figure 13.

		Random Forest Regressor			Keras.Sequential (Relu,mse,0.001)			HistGradient Boosting Regressor			
		R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	
USDA (6 classes)		36.1%	1.01	0.78	31.5%	1.08	0.82	41.8%	0.92	0.77	One Hot Encoding (33k x 423)
		64.8%	0.56	0.55	64.7%	0.56	0.58	68.2%	0.50	0.55	Value Hot Encoding 3 units (33k x 1049)
		65.8%	0.55	0.55	67.0%	0.52	0.56	69.4%	0.48	0.54	Value Hot Encoding 1 unit (33k x 423)
FSA (9 classes)		30.6%	2.06	1.13	29.6%	2.12	1.15	37.6%	1.85	1.09	One Hot Encoding (33k x 423)
		57.9%	1.25	0.85	62.2%	1.12	0.79	67.5%	0.97	0.76	Value Hot Encoding 3 units (33k x 1049)
		59.9%	1.19	0.82	61.1%	1.16	0.81	69.0%	0.92	0.74	Value Hot Encoding 1 unit (33k x 423)

Figure 12: Regression Results of Different Models

	HistGradient Boosting Regressor								
	Carbonhydrate	Protein	Fat	Sat Fat	Fiber	Sodium	Sugar		
R^2	36.5%	55.1%	29.3%	41.6%	50.4%	4.1%	50.5%		One Hot Encoding (33k x 423)
	71.6%	67.5%	63.9%	70.1%	66.3%	14.8%	83.1%		Value Hot Encoding 3 units (33k x 1049)
	71.6%	67.1%	63.9%	70.3%	66.4%	20.4%	83.6%		Value Hot Encoding 1 unit (33k x 423)

Figure 13: Regression Results for Different Nutrients

4 Conclusion

In conclusion, this study addresses a critical need in the era of information-driven dietary choices. By delving into the individual ingredients of recipes rather than focusing solely on entire dishes, we have demonstrated the potential to

predict the healthiness of meals with greater accuracy. In doing so, we strive to offer users a more nuanced approach to understanding the nutritional quality of their diets.

Recipe websites and apps are pervasive platforms for culinary decisions, yet they often lack mechanisms to promote healthier choices. Our machine learning model aims to bridge this gap and provide users with the tools to make informed decisions. Through rigorous data collection and extensive data preparation, we have compiled a comprehensive open-source recipe dataset available on GitHub, which will not only support our research but also contribute to the broader scientific community's efforts in this domain.

Furthermore, the importance of considering ingredient usage amounts has been highlighted. This nuance is crucial in predicting the healthiness of recipes accurately. We have made significant strides in the development of classification and regression models to predict health scores, offering a robust foundation for further research in this evolving field.

Our findings shed light on the intricacies of nutritional analysis and the potential for more informed dietary choices. This project is a testament to the power of machine learning in reshaping how we think about and interact with our food choices, ultimately contributing to healthier lives.

The finalized dataset for subsequent machine learning (ML) analysis and all project-related code are accessible on GitHub at the following repository: https://github.com/osmanbulutedu/Ingredient-Based-Recipe-Analysis

References

[1] Cooking and Baking Calculators, https://www.inchcalculator.com/cooking-calculators/, 2023.

[2] Produce Converter,2023,https://www.howmuchisin.com/produce$_c onverters$.

[3] Charalampos Chelmis and Bedirhan Gergin. A knowledge graph for semantic-driven healthiness evaluation of recipes. *Semantic Web Journal*, 2021.

[4] Charalampos Chelmis and Bedirhan Gergin. A Knowledge Graph for Semantic-Driven Healthiness Evaluation of Online Recipes, 2022.

[5] United States. Dietary Guidelines Advisory Committee. *Dietary Guidelines for Americans,2020-2025*, volume 9th Edition. U.S. Department of Agriculture and U.S. Department of Health and Human Services, 2020.

[6] Joint WHO/FAO Expert Consultation. Diet, nutrition and the prevention of chronic diseases. *World Health Organ Tech Rep Ser*, 916(i-viii):1–149, 2003.

[7] David Elsweiler, Hanna Hauptmann, and Christoph Trattner. *Food Recommender Food recommenderSystems*, pages 871–925. Springer US, New York, NY, 2022.

[8] UK FSA. Guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlets. *Food Standards Agency*, 2014.

[9] Mouzhi Ge, Francesco Ricci, and David Massimo. Health-aware food recommender system. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, page 333–334, New York, NY, USA, 2015. Association for Computing Machinery.

[10] Gordana Ispirova, Tome Eftimov, and Barbara Koroušić Seljak. Exploring knowledge domain bias on a prediction task for food and nutrition data. *IEEE*, 2020.

[11] Markus Rokicki, Christoph Trattner, and Eelco Herder. The impact of recipe features, social cues and demographics on estimating the healthiness of online recipes. 2018.

[12] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, and Martin Stettinger. An overview of recommender systems in the healthy food domain. 2018.

[13] Raciel Yera Toledo, Ahmad A. Alzahrani, and Luis Martínez. A food recommender system considering nutritional information and user preferences. *IEEE Access*, 7:96695–96711, 2019.