

Combating Obesity: Predicting the Health of Recipes

Vivek Srinivasan, vsrinivasan@ucsd.edu

Hacilioğlu Data Science Institute, UC San Diego, 9500 Gilman Drive, CA, 92093.

Submission Date: March 17, 2024.

I. DATASET

Introduction and Purpose

In the United States, obesity and heart disease are continuing epidemics brought on by high-calorie diets. Recipes with excess fats, sodium, and sugars are largely responsible for the high caloric content of the American diet.¹ The recipes and interaction datasets were chosen due to their all-encompassing data regarding recipes, including reviews and nutritional information for each recipe. With detailed preparation times and steps for over 80,000 recipes, a model trained on this data can serve as a powerful predictor for the healthiness of these meals.

This dataset is significant compared to others of its kind, as it is able to detail the nutritional values of each recipe, allowing people to make informed decisions about their health choices. It is well-established that excess caloric intake caused by fatty foods causes the body to store these calories as fat, leading to obesity.² Therefore, it follows that the most direct metrics of recipe health in regards to obesity are the caloric and fat content of each recipe. By classifying the recipes as ‘Healthy’, ‘Moderately Healthy’, and ‘Unhealthy’ based on their fat and caloric content, users can determine their choice of recipe to help them stay in shape or meet prospective goals.

Exploratory Data Analysis

The dataset contained several categorical and quantitative variables, overall detailing the nutritional information of the recipes. Overall, many of the qualitative data columns (‘description’, ‘steps’, ‘tags’) were unnecessary for health analysis, as these

columns varied nominally based on each individual recipe.

Figure 1: Column Names in the Initial Dataset

```
Index(['name', 'id', 'minutes', 'contributor_id', 'submitted', 'tags',  
      'nutrition', 'n_steps', 'steps', 'description', 'ingredients',  
      'n_ingredients', 'rating'],  
      dtype='object')
```

Of the columns, the most significant ones were the nutritional value and recipe complexity (‘n_steps’ and ‘minutes’) columns. First, only recipes with a reasonable caloric content (<10000 cal) and a single-day preparation time were kept. Recipes that take reasonable times are necessities in the average household, and the ideal model is able to predict plans for everyday use, rather than special occasions. The nutrition column then needed to be extracted to gain the individual nutritional values in their own columns. The submission dates of each recipe were also changed to datetime objects from strings to observe patterns over the years.

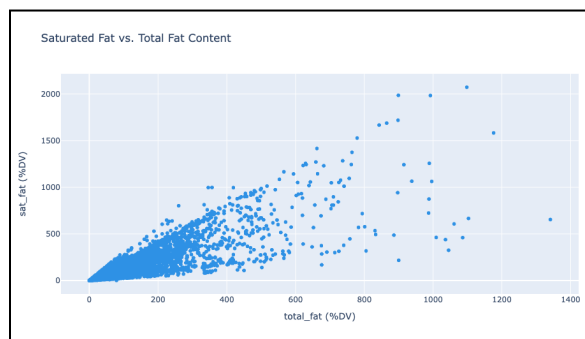
The final dataset included mostly quantitative features, with basic descriptive information (‘name’, ‘id’, ‘submitted’) serving as identification rather than being included in features. Due to the large number of samples present, rows with missing values were removed, rather than being imputed.

Interesting Findings

Initial exploration of the data revealed information about the nutritional values and complexity of each recipe. Total fat and calories were determined to be the most directly related to instances of obesity.² However, values like saturated fat and carbohydrates were explored further in order to identify multicollinearity in

any of the columns. Interestingly, saturated fat was less related to total fat content than expected.

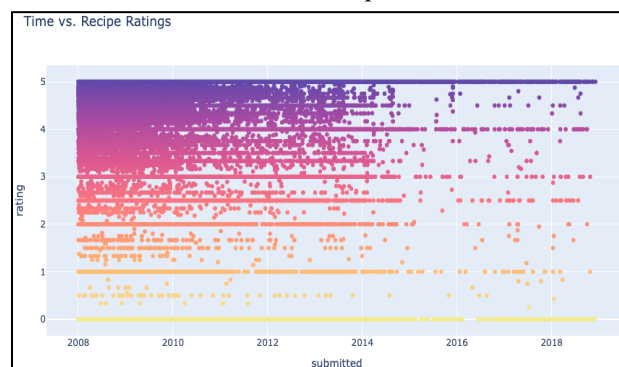
Figure 2: Relationship between total (x) and saturated fat (y). The relationship is more deviated than the linear correlation expected.



By identifying that total and saturated fats were not directly correlated with each other, multicollinearity of these two features was ignored, and saturated fat was able to be used in prediction. In a similar fashion, sugar and total carbohydrates were both identified as adequately separate, and were therefore both used for prediction analysis.

The second major exploration came when looking at feature relationships with time ('submitted'). As the years went on, data seemingly grew more sparse, suggesting that fewer recipes were being submitted into the 2020s as compared to the 2000s and 2010s. This could pose problems for the prospective model's accuracy as the recipes used for training become more out-of-date. Additionally, analysis of ratings showed that recipes were experiencing fewer interactions as time passed, leading to potential concerns regarding the model's permanence.

Figure 3: Relationship between recipe rating and date submitted. The lack of variability between recipes indicates lower rates of interaction as time passes.



II. PREDICTIVE TASK

Identification of Task

The predictive task identified for the model was to predict the "health ranking" of each recipe. Based on the identified problem of obesity in America, recipes were ranked A-C, with group A representing healthy recipes, group B representing moderately healthy recipes, and group C being unhealthy recipes. The groups were assigned based on the total fat and caloric content, with group A recipes being in the lowest (0–33.3) percentile, group B recipes in the middle (33.4–66.7), and group C recipes being in the top percentile (66.8–100) for each category. Splitting of the columns by their percentiles was done due to the right-skewed nature of each column. Since total fat and calories were the features used to predict nutritional value, they were omitted from the model's prediction. The identification of calories and total fat as the columns of analysis stems from the idea that excesses of these factors result in increased likelihoods of obesity in Americans.

This task was done as a classification problem rather than a regression analysis due to the multivariate nature of the response. Health as a general concept is difficult to measure on its own, as it looks different for every person. Furthermore, classification leads to simplicity in

both design and application. Classifying recipes was much simpler than trying to design a quantitative metric for obesity from these features, and it allows users to easily make the same classification of high- or low-calorie, low-fat recipes. By assigning recipe ranks in categories, users can make informed decisions regarding the meals they choose to make.

Metrics of Analysis

Once the classification task had been identified, evaluation was done via the accuracy and F1 scores for the predicted recipes. Because the recipes were divided into three categories, a weighted average was used for the F1 score in order to favor highly-occurring predictions. The moving average of the F1 score was to easily identify consistently odd predictions in the model if they were to occur. Accuracy was naturally chosen to determine how correctly the model predicted, while precision and recall (in the form of the F1 score) were included to ensure the model was not predicting a single value only.

Baseline Models

After establishing the metrics of analysis for the task, two baseline models were developed to compare to the final model. The first model selected was logistic regression. This was used as a true baseline; multiclass logistic regression is effectively the most basic model for predictions, meaning it serves as a solid indicator of improvement, regardless of the final model choice. The only drawback associated with the logistic regression was its assumption of linearity between the features and response. This was a minor issue, as the features were strongly related, though not necessarily linearly.

The second model chosen was a multinomial Naive Bayes' classifier, which was chosen due to its ability to work well with multi-class features. Since the health rank of each recipe was dependent on both calorie and

fat content, a Naive Bayes' model served as an indicator of the reliability of the final model. The main issue with this model was the assumption of independence between the columns. Despite the fact that any moderately linearly-correlated features were removed from the data, many nutritional values are dependent, and the complexity and time of recipes are often related. This resulted in lower ability to predict health rankings, as the columns were related.

The final model was able to improve on these two baselines due to its increased complexity and ability to fine-tune parameters. Indeed, the final model's flexibility as compared to these baselines stemmed from the large assumptions that allowed these models to be simple (linear correlation for logistic classification and independence for Naive Bayes'). Without these assumptions about the input data, the model was not limited to these considerations, and was thus an improvement over the baselines.

III. MODEL

Initial Model and Feature Representation

The model chosen for this task was a random forest classifier (RFC) with a limited maximum depth. Initially, the model was chosen to be a simple RFC with no adjusted parameters, but the size of the dataset resulted in inclusion of a maximum depth to avoid overfitting and reduce the time required for prediction.

The features selected for this task were quantitative. After the nutritional values were extracted, there were no qualitative columns that required one-hot encoding or numerical representation. Indeed, the 'ingredients' column provided similar data to the nutritional values, and the 'steps' column was unnecessary due to the presence of 'n_steps', which detailed the number of steps rather than the steps themselves. Missing values were removed from the data, and any unique features like 'ID' were also removed due to their complete independence from the

metrics being predicted. Figure 4 shows the final columns used to make the data predictions.

Figure 4: Columns used as features for data predictions.

```
[ 'minutes',
  'n_steps',
  'n_ingredients',
  'rating',
  'sugars (%DV)',
  'sodium (%DV)',
  'proteins (%DV)',
  'sat_fat (%DV)',
  'carbs (%DV)']
```

Incorporating the nutritional values, rating, and metrics of preparation complexity into the data was an imitation of the qualities recipes are often based around. Simple and quick recipes are heavily encouraged in a society where people have little time to spend on their meals, while low-sodium, low-sugar recipes are often marketed as healthy to consumers. Highly rated recipes were arguably one of the most important features to include, as the model serves no purpose if rating is not considered and all the recipes ranked as healthy are unappetizing, as no incentive is provided to want the healthy foods.

Improvements on the Model

Once the established maximum depth of the model was determined to be necessary, analysis with gridsearchCV was used to confirm the best depth for the model, and a depth of 5 was then selected. This is an ideal depth, being large enough to allow the decision trees in the model to make complex decisions, while still being small enough to avoid the high variance commonly associated with decision trees and random forests.

All other parameters were tuned well for the data by default, and therefore did not need to be included as adjusted hyperparameters during the gridsearchCV. The default of a single

instance being the minimum leaf size and 2 samples being required for splitting were both acceptable, as the inclusion of a specified maximum depth was able to effectively prune the trees contained in the forest. Using the Gini index to determine where and when to split allowed the model to lower this index, as the distribution of recipes into the three categories was relatively equal, with group C having the lowest number of samples ($n_A = 33,264$, $n_B = 27,961$, $n_C = 21,935$). The decision to use the Gini impurity over entropy considerations was done to reduce the complexity of the model and allow the decision trees to operate quickly, as there was little difference in the performance of the models.

Drawbacks of the Model

Overall, the model was optimized to avoid issues with overfitting or inaccurate predictions. The optimization of the model was significant in terms of its interactions with other datasets. Splitting the data into training and test sets offset some of the prospective variability, but there was no method to compare to other datasets with international recipes prepared with different ingredients and steps. In this way, the model's prediction is limited to the United States. The other issue with the model comes from the data-gathering process; there is not always a way to ensure nutritional values are included in the recipe data, so the preparation and ratings become even more important when a lack of nutritional information is provided. Overall, the main issue stems from increased variance rather than any biases present in the model, and much of the optimization served to reduce this variance.

IV. LITERATURE

Due to its unique, all-encompassing column structure, this dataset has been used for a wide variety of food and recipe-related predictions. This dataset was chosen based on

use in a previous class; though the application was different, the dataset was used in DSC80 at UCSD to develop understanding of regression and prediction problems.³ In this way, there have been dozens of simple linear and logistic regression performed on the dataset to predict recipes and nutritional values. Furthermore, the time of submission ('submitted' column) allows for time series and moving average analysis, though these have not been performed due to their complexity in comparison to linear regression. Additionally, the ingredients, ratings, and steps columns serve as powerful tests of the bag-of-words and text feature classifications, as used in the original study with this dataset.⁴ By predicting the recipe information from its ingredients, new recipes were generated and compared with existing recipes.

In comparison with previous regression and text feature analysis, this problem was a classification, and therefore required the columns to be in new formats for proper prediction. The combination of features used as the response was also not something that was done during simple regression or bag-of-word analysis; these were all univariate problems coming from a single dataset column, while this task required the use of two columns.

Because of its novelty, similar problems and results specific to this dataset are not currently published. However, in fields of health and food sciences, classifying recipes in terms of their health is a large part of ensuring meals that accommodate all needs. Research is also being performed on the idea of healthy eating; the psychology behind labeling food as 'healthy' or 'unhealthy' and nutrient profiling are subjects of intense study. These studies have shown that healthy food is not directly translatable to 'good' food, and unhealthy food is not always 'bad'. There are mixes of healthy and unhealthy aspects in many recipes, and this balance ultimately contributes to health in the long term.⁵ Overall, the continuing necessity of healthy

options and diverse foods means that datasets with recipe and nutritional information have novelties that do not wear off, making them important subjects of study regardless of their popularity.

V. RESULTS

Once the data was ready for analysis, a testing size of 20% was used to separate the dataset into train and test sets. The training set had a size of 62,370 recipes, and the test set had a size of 20,790 recipes. After the train-test split, the data was ready to be fit for the baseline and test models.

Baseline Performance

The baseline performance was acceptable in terms of logistic regression, and worse than expected for the Naive Bayes' classifier. Table 1 details the accuracy and F1 scores for each of their models after being fit to the training data. In comparison to the training data, the similar test values show both models have decently low variabilities, but the Naive Bayes' classifier assumption of independence resulted in much lower accuracy predictions than the logistic classifier.

Table 1: Accuracy and F1 scores for both baseline models.

	Accuracy	F1
Logistic	0.761	0.761
MultiNB	0.625	0.623

Final Model Performance

In comparison, the random forest classifier, the chosen model, outperformed both baselines by non-negligible margins while maintaining the lower variance of the baselines. The training and test set data is detailed in table 2.

Table 2: Training and test-set accuracies and F1 scores for the Random Forest Classifier.

	Accuracy	F1
Train	0.839	0.837
Test	0.831	0.830

As seen, the test accuracy is higher than the logistic regression by about 0.070, and the F1 score is also a 0.069 improvement. The test accuracy of the final model is much higher (0.213) than the Naive Bayes', as is the F1 score (0.213). These scores are extremely similar for the final model, meaning the precision and recall combined as F1 to yield similar results as the overall accuracy. The metrics proved to be appropriate for conducting classification, as the correct and incorrect classifications were both scored through accuracy and F1, providing a comprehensive view of the model's capabilities. Therefore, the model predicted consistently throughout the dataset, meaning its variance is low but its bias is variable depending on the sample. There was little indication of bias in the model, as the data received was unbiased and large enough to represent all three categories.

This model outperformed both baselines due to complexity and set maximum depth. The simpler logistic and Naive Bayes' classification relied heavily on existing assumptions, but there were no assumptions made in any of the decision trees, meaning the model was more flexible in its operation. Flexibility is oftentimes a path to overfitting, but the set depth prevented this from occurring and made the model much faster and more efficient. Because the model already outperformed both baselines, entropy and hyperparameter analysis was deemed unnecessary.

The most useful features were likely the sodium and carbs columns, as these two nutritional values are most connected to obesity (aside from calories and total fat). Minutes were

also likely connected deeply to the recipe health, as the longer ingredients are cooked, the less helpful the nutritional values were due to the differing nature of cooked and raw foods. The least useful column was the recipe rating; the presence of rating in the model is important, but placing it in the features means that it will be used, but not magnified the same way that 5-star ratings are.


In future iterations, the ratings can be removed and used after classification to allow people to see how they are eating and why it is important. Furthermore, the maximum depth and minimum samples required to split the tree can be tuned using gridsearchCV to find the best parameters and adjust the forest to those parameters. Using the Gini impurity yielded similar results as the entropy of the tree, so Gini was used for this dataset (Table 3). In the presence of datasets resulting in large node entropies, the model can be reset to split based on entropy.

Table 3: Accuracy and F1 scores for the model tuned to Gini impurities and entropies.

	Accuracy	F1
Gini	0.831	0.830
Entropy	0.829	0.827

Overall, the high performance of the model confirmed its success in comparison to the baselines, therefore displaying the necessity of the feature column. It also showed that there are non-linear relationships between the feature and response variables, which explained the failures of the baseline models. Development of models such as this to accurately predict recipe health is a crucial step to combating the obesity epidemic in food and health sciences. Thus, continued research is necessary for the common health of the nation.

VI. REFERENCES

- ¹ National Institute of Diabetes, Digestive, and Kidney Diseases. (2018) Overweight and Obesity Statistics,
[https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity#:~:text=More%20than%201%20in%203,who%20are%20overweight%20\(27.5%25\).](https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity#:~:text=More%20than%201%20in%203,who%20are%20overweight%20(27.5%25).)
- ² National Health Service. (2023) Causes - Obesity,
<https://www.nhs.uk/conditions/obesity/causes/#:~:text=Obesity%20is%20a%20 complex%20 issue,in%20the%20body%20as%20fat.>
- ³ Rampure, S. The data science lifecycle , DSC 80. <https://dsc80.com/proj04/>.
- ⁴ Majumder, B. P. et al. (2002) Generating Personalized Recipes from Historical User Preferences,
<https://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19c.pdf>
- ⁵ Davies, S., Lobstein, T. (2008) Defining and labelling 'healthy' and 'unhealthy' food. Public Health Nutrition, DOI: 10.1017/S1368980008002541,
<https://pubmed.ncbi.nlm.nih.gov/18510787/>.