Classifier for Foods: Healthy, Unhealthy, or Should be Consumed in Moderation

Nick Steiner, nsteiner@oakland.edu

ABSTRACT

The goal of this project was to identify the best classifier for categorizing foods into three categories: healthy, unhealthy, or should be consumed in moderation. I chose to solve this problem so that anybody who wanted to could determine how often they should eat any food in question. I found that a K-Nearest Neighbors classifier that implemented Z-score scaling, SNOPE oversampling, and PCA achieved the highest accuracy out of my findings at 90.7%. The output from a confusion matrix tells us that the model performs nearly perfectly on healthy and unhealthy foods, but has some confusion with determining foods that should be consumed in moderation. The high accuracy and low mispredictions for the opposing classes: healthy and unhealthy show that this model can be used in a real implementation through a UI as is, but further research and development can improve it further.

1. INTRODUCTION

Problem Statement: What is the most effective machine learning model for classifying foods as healthy, unhealthy, or suitable for consumption in moderation?

Existing Solutions: No previous research has been done using this approach, but there exists a similarly implemented solution using a Convolutional Neural Network (CNN) model to analyze images and determine if the food is healthy or unhealthy without a third class.

Proposed Solution: In summary, this project makes the following contributions:

- I built a classification model to analyze the nutrition facts of foods and determine if they are healthy, unhealthy, or that they should be eaten in moderation.
- I implemented a K-Nearest Neighbors (KNN) classification model to classify foods; this model uses z-score scaling to scale the differences in scaling amongst macronutrients, SNOPE oversampling to balance the existing dataset, and Principal Component Analysis to reduce the number of features in the dataset.
- The KNN model with its implemented characteristics achieves an accuracy score of 90.7%. This accuracy improves on the existing CNN classifier while also adding a third class for a more descriptive output.

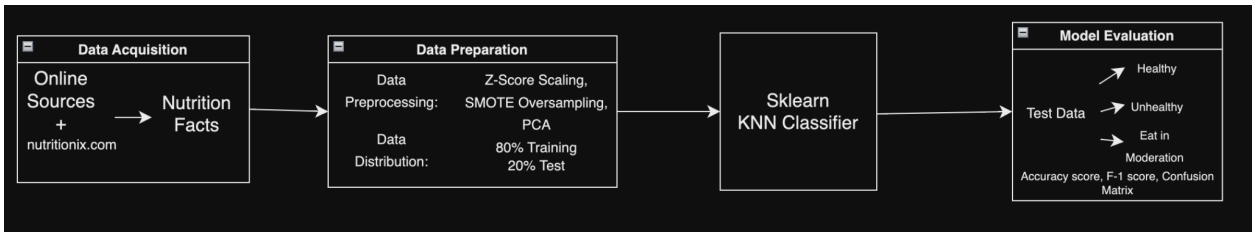
2. RELATED WORK

A similar research task has been completed around this task of classifying foods by if they are healthy or unhealthy. IEEE published a research paper titled *Prediction of Healthy and Unhealthy Food Items using Deep Learning* which details the usage of a Convolutional Neural Network to classify if a food is unhealthy or healthy based on an image of the food which resulted in an accuracy of 90.37%. The differences between my approach and this existing work is the methodology of the end user's data input, the type of model used, and the dataset used for training. This existing CNN requires the user to input a single image of the food, while my approach requires the user to input one text data of every item listed on a nutrition label. The model used in this approach is a CNN neural network, while mine is a simpler K-Nearest Neighbors algorithm. The dataset used in this approach consists of 1270 images, while my dataset consists of 220 sets of foods with their nutrition facts.
My approach achieves a slightly higher accuracy, at 90.7%, processes the input data quicker, and requires minimal hardware compared to what is necessary to run the CNN of the other approach. My approach also adds a third category of "should be eaten in moderation" while this approach only uses the two categories of healthy and unhealthy.

3. MOTIVATING EXAMPLE

Many people across the world are tricked into believing food labels identifying a food as "healthy" when in reality the macronutrients provided on the nutrition facts give a deeper insight into the fact that they are actually unhealthy. According to the Heart and Stroke Foundation of Canada, foods like dried fruit, flavored yogurt, and pretzels are common foods that people think are healthy but actually aren't ["10 foods that sound healthy but aren't."]. On the other hand, some people have preconceived notions on some foods being unhealthy, when in reality they are perfectly fine to consume in moderation and may even be healthy for you.

4. APPROACH



4.1 Data Collection and Preprocessing

Data Sources: The dataset used for this project was gathered entirely by me. I gathered nutrition labels from various sources, compiling them into a spreadsheet to convert into a CSV file. I labeled the dataset using my own discretion but based my labels on what the FDA and other sources say about which macronutrients can make a food healthy or unhealthy. These macronutrients are those listed on a nutrition label, and a mixture of daily value percentages (DV%) and masses were used as features in the dataset:

- Calories (kCal)
- Total Fat DV%
- Saturated Fat DV%
- Trans Fat (g)
- Cholesterol DV%
- Sodium DV%
- Carbohydrates DV%
- Daily Fiber DV%
- Total Sugars (g)
- Added Sugars DV%
- Protein (g)
- Vitamin D DV%
- Calcium DV%
- Iron DV%
- Potassium DV%

Preprocessing Steps: The scales of the different macronutrients had very large differences. For instance, Calories tended to be a lot higher than any other feature, and trans fat tended to be 0 for many data points, and never exceeded 10. To combat this issue I implemented z-score

scaling which uses the mean and standard deviations of each feature to center the values of features around 0. While collecting the data I determined that physically finding data points to balance the dataset lowered the accuracy, so instead I chose to balance the dataset by implementing SMOTE oversampling which creates synthetic data points using k-nearest neighbors and euclidean distance. In order to train the model I also split my dataset into training and testing datasets based on an 80/20 split.

4.2 Model Selection and Training

Algorithms Used: I implemented a k-nearest neighbors (KNN) algorithm. This model is useful because it is great for classification problems and can be used with continuous variables, as all of my features have continuous values. The benefit compared to other algorithms is that it uses euclidean distance and majority voting to determine the classification of new data points. This is useful in my case as I have many features that can cause data points to be very far apart from those in other classes in a high-dimensional space. This is hard for a human to comprehend, but a KNN model has no trouble calculating these distances.

Training Process: There were many different parameters I could tune to find the best accuracy for my model. Before the training I had to find the best amount of components for the Principal Component Analysis (PCA). I did this manually by adjusting the amount of components and recording the accuracy for each number “n”. I also adjusted the number of neighbors used in KNN and whether or not the features are weighted. Knowing that the dimension of the data points is so high and the scale of features are so different I could assume the features should be weighted. After determining this I tested n_neighbors parameters of 1-10 and found that 7 neighbors resulted in the best performance on the test set. The results of this are shown in **Figure 1**.

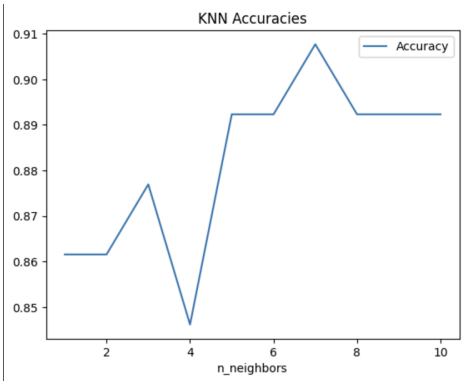


Figure 1

5. EXPERIMENTAL EVALUATION

5.1 Methodology

With this project I was aiming to address the research question: What is the most effective machine learning model for classifying foods as healthy, unhealthy, or suitable for consumption in moderation? My hypothesis is that a simple machine learning model can achieve a higher accuracy on numerical data than a deep learning Convolutional Neural Network.

To test this hypothesis I utilized the ML library scikit-learn to implement Decision Tree, K-Nearest Neighbors, Gaussian Naive Bayes, and Support Vector Machine classifiers, z-score scaling, PCA, SMOTE, train-test-split, and model evaluation metrics. I used the benchmark set by the image processing CNN model in the research previously mentioned: 90% accuracy. I also used confusion matrices to determine which classes were being mispredicted. I did not want 90% accuracy with a lot of mispredictions in the “healthy” or “unhealthy” categories, as this is more dangerous to users than mispredictions in the “should be eaten in moderation” category. The train/test data included all of the nutrition facts gathered split into 80/20 train/test splits. This dataset is realistic to use in this context as the macronutrients specified on a nutrition label are a big indicator on the “healthiness” of a food. The independent variables in the test include the features mentioned previously, and the dependent variable is the category output by the model. The compared accuracies and confusion matrices are presented below in the **Results** section.

5.2 Evaluation Metrics

I used the most common evaluation metric for classification problems, accuracy, to determine how well my model was performing and compare against other models I could possibly use for this solution. To determine the distribution of mispredictions I used a confusion matrix. As mentioned above it was important for most of the mispredictions to occur in the “should be consumed in moderation” class. F1-score is important in assessing my machine learning model because it ensures that the model performs well in both identifying positive cases (recall) and minimizing false positives (precision).

5.3 Results

I originally tried implementing a Decision Tree but did not get the desired accuracy with updated versions of my dataset. I then implemented a K-Nearest Neighbors (KNN) algorithm, which had an accuracy of 90.7% and an F1-score of 90.57%. I implemented a Support Vector Machine

(SVM) model which had an accuracy of 83.08% and an F1-score of 83.03%. Finally, I implemented a Gaussian Naive Bayes model, which did not perform well, as it had 64.6% accuracy and an F1-score of 63.13%. The accuracies are found below in **Figure 2** and the F-1 scores are found in **Figure 3**. Confusion matrices for all models are found in **Figure 4**. The resulting accuracies, f1-scores, and confusion matrices indicate that KNN is the best model to use for this problem and dataset. The accuracies and f1-scores are clearly higher than all of the other three models, and the confusion matrix shows what I wanted to see: the mispredictions are mostly happening with the 3rd class - should be consumed in moderation.

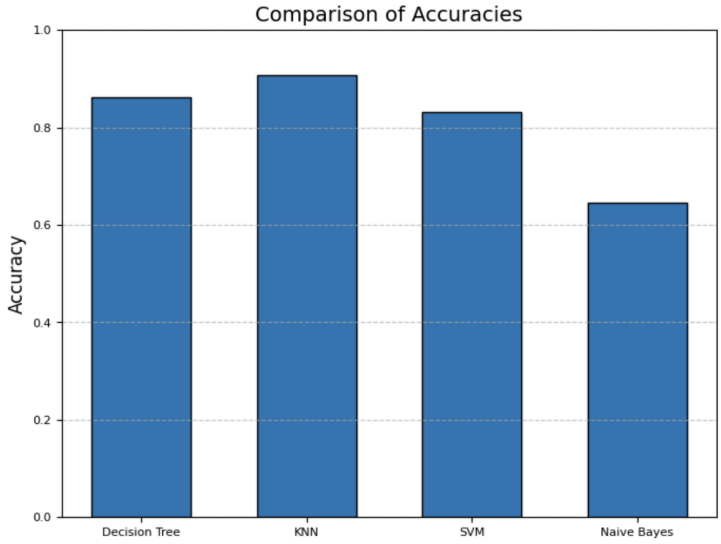


Figure 2

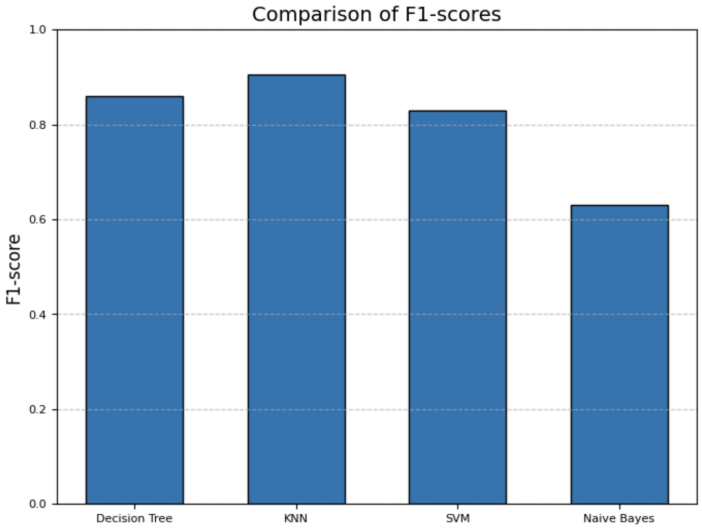


Figure 3

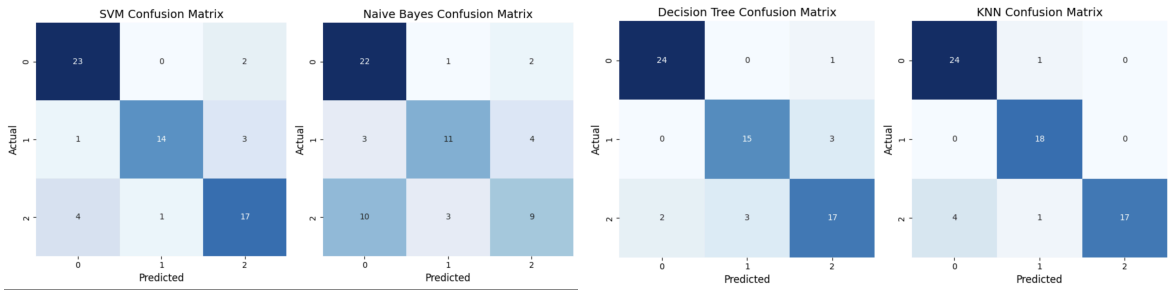


Figure 4

5.4 Discussion

My hypothesis was found to be true, as I was able to slightly exceed the CNN model's accuracy while implementing a third category for classification. These results signify the strength of my approach over the CNN model: a third class is added for more precision. As expected going into this hypothesis testing, my weakness still exists: it is easier to input the desired data for the CNN model than it is for my approach. This is the case because of the fact that I use many more features that come directly from the user: all 15 nutrition facts, while the CNN model only requires one "piece" of data from the user: an image of the food.

6. LIMITATIONS

The largest area of improvement for this project is the discretion made when I labeled the dataset. Since I gathered this dataset myself I had to determine based on some sources' information what category each food fell into. The accuracy of the model could likely be improved if a professional nutritionist was consulted in the process of collecting this dataset. Another large limitation of this project is the assumptions I made regarding the end users. My classifications were made based on the FDA's recommendations for daily value intakes based on a 2,000 calorie diet. Many end users' would not fall into this dietary recommendation. The last limitation of my project is that many indicators of food healthiness are not reflected in the nutrition label, such as unhealthy food additives, preservative usage, and the preparation of a food.

7. CONCLUSIONS AND FUTURE WORK

This project has resulted in a tool that can be used by many if implemented through an application or simple online UI on a website such as HuggingFace. The 90% accuracy benchmark gives better results than what can be determined by most people who do not have background knowledge in nutrition. To increase this accuracy further and provide an even more useful tool I would consult a nutritionist to help me understand what makes a food healthy or unhealthy, or let them label the dataset themselves. I would also turn this model into a full-fledged application with more features that make it easier to use the model, such as a nutrition label reader through an image. This application would also allow the user to input more information about themselves and the model would be able to use this information to tailor the classification more towards the specific user, as well as change over time based on users' eating habits. In the future I'd also implement more features into the model to take into account more than just what is presented on nutrition labels.

8. DATA AVAILABILITY

This dataset and model code are available through GitHub: https://github.com/nsteiner/CSI4130_Project/

9. ACKNOWLEDGEMENTS

My dataset was labeled using information from the FDA ["How to Understand and Use the Nutrition Facts Label."]. I also used Nutritionix.com to gather a majority of the nutrition label information ["Nutritionix Database."].

10. REFERENCES

"Daily Value on the Nutrition and Supplement Facts Labels." *FDA*, 5 March 2024, https://fda.gov/food/nutrition-facts-label/daily-value-nutrition-and-supplement-facts-labels. Accessed 8 November 2024.

"50 Foods That Are Super Healthy." *Healthline*, https://www.healthline.com/nutrition/50-super-healthy-foods#fa-qs%20,%20https://www.nutritionix.com/food. Accessed 8 November 2024.

"How to Understand and Use the Nutrition Facts Label." *FDA*, 5 March 2024, https://www.fda.gov/food/nutrition-facts-label/how-understand-and-use-nutrition-facts-label. Accessed 8 November 2024.

"Nutrition Education Resources & Materials." *FDA*, 27 September 2023, https://www.fda.gov/food/nutrition-food-labeling-and-critical-foods/nutrition-education-resources-materials. Accessed 8 November 2024.

"The Nutrition Facts Label." *FDA*, 5 March 2024, https://www.fda.gov/food/nutrition-education-resources-materials/nutrition-facts-label. Accessed 8 November 2024.

"Nutritionix Database." *Nutritionix - Largest Verified Nutrition Database*, http://nutritionix.com. Accessed 8 November 2024.

"US Dietary Guidelines." *Dietary Guidelines for Americans: Home*, https://www.dietaryguidelines.gov/. Accessed 8 November 2024.

Zamarripa, Maria. "20 Healthy Condiments (And 8 Unhealthy Ones)." *Healthline*, 20 August 2020, https://www.healthline.com/nutrition/list-of-condiments. Accessed 8 November 2024.