Machine Learning for Predicting Healthiness Through Ingredient Analysis

SmartSnack
Shubham Roy (B2330040), roy.sonai665@gmail.com,
Parimal Bera (B2330031), parimalbera244@gmail.com

May 1, 2024

Abstract

This paper explores the application of machine learning techniques for predicting the healthiness of food products based on ingredient analysis. We investigate various classification algorithms and evaluate their performance on a dataset containing nutritional information of different food items. With the growing emphasis on healthy eating habits, there is an increasing need for accurate tools to distinguish between healthy and unhealthy food options. In this study, we explore the application of machine learning techniques to predict the healthiness of food products based on ingredient analysis. We investigate a variety of classical machine learning algorithms along with some advanced models to classify food items into healthy and unhealthy categories. These findings contribute to the development of tools that can aid consumers in making informed dietary choices and promote healthier eating habits.

1 Introduction

The rising awareness of the critical role healthy eating plays in overall well-being has sparked an ever-growing demand for innovative tools that empower consumers to make informed dietary choices. In response to this pressing need, machine learning emerges as a beacon of hope, offering unparalleled opportunities to develop sophisticated predictive models capable of evaluating the healthiness of food products based on their intrinsic ingredients. In this paper, we embark on an exhilarating journey into the realm of machine learning, driven by the quest to unlock the potential of ingredient analysis in predicting the healthiness of food items. Through the lens of cutting-edge algorithms and advanced methodologies, we delve deep into the intricate relationships between nutritional compositions and dietary health outcomes.

2 Methodology

In our study, we employed a diverse range of machine learning algorithms to classify food items into healthy and unhealthy categories. Each algorithm offers unique strengths and characteristics, which we explored to determine their efficacy in predicting food healthiness based on ingredient analysis.

2.1 Machine Learning Algorithms

We evaluated a diverse array of machine learning algorithms, each with its own set of characteristics and assumptions. Here's a brief overview of the algorithms we considered:

- **Logistic Regression:** Logistic regression models the probability of a binary outcome using a logistic function. It fits a linear decision boundary to the data and predicts the probability that a given data point belongs to a particular class.
- **K-Nearest Neighbors (KNN):** KNN classifies data points based on the majority class among their k nearest neighbors in the feature space. It relies on the assumption that similar data points belong to the same class.
- **Support Vector Machine (SVM):** SVM constructs hyperplanes in a high-dimensional space to separate data points into different classes, maximizing the margin between classes. It can handle non-linear decision boundaries by using kernel functions.
- **Naive Bayes:** Naive Bayes models the probability of a data point belonging to a particular class based on the conditional probabilities of its features. It assumes independence among features, which simplifies the calculation of probabilities.

Each machine learning model offers unique advantages and may be more suitable for specific types of data or classification tasks. By exploring a diverse range of models, we aim to identify the most effective approach for predicting food healthiness based on ingredient analysis.

3 Experimental result

In this section we are going to see the result of our project.

- **Dataset:** We collected a dataset containing nutritional information for a variety of food products. The dataset includes features such as energy content, protein, carbohydrate, fat, fiber, sodium, iron, cholesterol, and various other nutrients per 100 grams of each food item. Additionally, each food item is labeled as either "Healthy" or "Unhealthy" based on expert judgment. In addition to traditional food products,

we also collected data from fast food sources. This additional dataset comprises nutritional information obtained from ingredient lists provided on fast food packaging. These data supplement our analysis by providing insights into the nutritional profiles of commonly consumed fast food items. Furthermore, we combined two distinct datasets for our analysis. The first dataset was collected from the USA health ministry and contains comprehensive nutritional information sourced from various food products. The second dataset was obtained from Kaggle and includes data on food items and their ingredients. By merging these datasets, we leverage a diverse range of sources to create a comprehensive and representative dataset for our analysis. The combination of datasets from different sources enriches our analysis and enhances the breadth and depth of our findings. By incorporating data from fast food sources and merging datasets from multiple sources, we ensure the robustness and reliability of our analysis, enabling us to draw meaningful conclusions regarding the healthiness of food items across different categories and sources. You can access the dataset from the following link:

`https://drive.google.com/drive/folders/1S1XCTZtUlO9cLMutgUzG3ATKmO-cO1cr?usp=drive_link`

- **Nutritionist Consultation and Threshold Values:** To ensure the accuracy and relevance of our analysis, we consulted nutritionist Soumita Kundu to establish threshold values for key nutrients. Soumita Kundu's expertise in nutrition science provided invaluable insights, guiding us to define threshold values for essential nutrients such as protein, carbohydrates, fats, fiber, sodium, iron, and cholesterol. These thresholds play a pivotal role in our classification framework, aiding in the assessment of nutritional quality. We classify food as healthy or unhealthy for four types of people: normal individuals, those with high cholesterol, those with high fat , and those with high sugar. The established thresholds serve as critical indicators in this classification process. We express our sincere gratitude to Soumita Kundu for her contributions and unwavering support, which have been instrumental in shaping the direction and quality of our work (Kundu, 2024).

- **Imbalance in Dataset and SMOTE:** The dataset used in this study exhibited a significant class imbalance, with a disproportionately higher number of samples in one class compared to the other. This imbalance can lead to biased model performance, where the classifier may favor the majority class and perform poorly on the minority class. To address this issue, Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE is a popular technique used to balance class distribution by generating synthetic samples for the minority class, thus mitigating the imbalance. By oversampling the minority class, SMOTE helps to create a more balanced dataset, leading to improved model performance and generalization. In this study, SMOTE was applied to the dataset to create synthetic samples for the minority class, ensuring

a more equitable distribution of samples across classes. This approach not only enhanced the robustness of the predictive models but also minimized the impact of class imbalance on model outcomes.

Health Label	**Count**
Healthy	7338
Unhealthy	428

Table 1: Health Labels for Normal Person

- **Experimental settings:**
  - For each type of individual, specific criteria were defined to label food items as healthy or unhealthy based on dietary guidelines and nutritional recommendations tailored to each group's health needs.
  - **Normal People:** Food items were labeled as healthy if they met general nutritional guidelines for a balanced diet, with moderate levels of fat, sugar, and sodium. This criteria ensures that food options align with the dietary requirements of the general population.
  - **People with Cholesterol Issues:** Food items were labeled as healthy if they were low in cholesterol and saturated fats, and high in fiber and unsaturated fats. This classification aims to support individuals with cholesterol concerns by promoting foods that contribute to heart health.
  - **People with High Fat:** Food items were labeled as healthy if they were low in saturated fats and trans fats, and high in unsaturated fats. This categorization caters to individuals with high-fat intake.
  - **People with High Sugar:** Food items were labeled as healthy if they were low in added sugars and high in fiber. This labeling strategy targets individuals with high sugar consumption, prioritizing foods that aid in sugar regulation and overall health.
- **Experimental results:**
  - **Normal People:**
    We evaluated the performance of each algorithm using accuracy, precision, recall, and F1-score metrics. Additionally, we conducted cross-validation to assess the robustness of the models. Normal people, representing a broad demographic, often have diverse dietary habits and nutritional requirements. Hence, it was essential to ensure that the models could accurately classify food items as healthy or unhealthy for this group.

Model	Accuracy
Logistic Regression	0.9761
K-Nearest Neighbors	0.9761
Support Vector Machine	0.9142
Naive Bayes	0.99

Table 2: Accuracy of Machine Learning Models for Normal Person

– **People with Cholesterol Issues:**
For people with cholesterol issues, we rigorously assessed the effectiveness of each algorithm by evaluating key performance metrics including accuracy, precision, recall, and F1-score. Furthermore, to gauge the robustness and reliability of the models, we conducted cross-validation procedures. Individuals with cholesterol concerns often require specific dietary considerations due to their health condition. Therefore, it was imperative to verify that the models could reliably distinguish between healthy and unhealthy food items tailored to this demographic, considering their unique dietary needs and restrictions.

Model	Accuracy
Logistic Regression	0.9582
K-Nearest Neighbors	0.9690
Support Vector Machine	0.9176
Naive Bayes	0.9414

Table 3: Accuracy of Machine Learning Models for People with Cholesterol Issues

– **People with High Fat:**
For individuals with high fat intake, we meticulously evaluated the efficacy of each algorithm by scrutinizing essential performance metrics such as accuracy, precision, recall, and F1-score. Additionally, to ascertain the resilience and consistency of the models, we conducted comprehensive cross-validation analyses. Individuals with high fat intake often necessitate particular dietary attention due to their nutritional requirements and potential health implications.

Model	Accuracy
Logistic Regression	0.9761
K-Nearest Neighbors	0.9761
Support Vector Machine	0.9142
Naive Bayes	0.9996

Table 4: Accuracy of Machine Learning Models for People with High Fat

– **People with High Sugar :**
For individuals with high sugar intake, we meticulously examined the performance of each algorithm by evaluating crucial metrics such as accuracy, precision, recall, and F1-score. Furthermore, we conducted thorough cross-validation procedures to ensure the reliability and consistency of the models. Individuals with high sugar intake often require specialized dietary attention due to potential health concerns associated with excessive sugar consumption. Therefore, it was imperative to verify that the models could accurately classify food items as healthy or unhealthy for this demographic, considering their specific dietary habits and nutritional requirements.

Model	**Accuracy**
Logistic Regression	0.8926
K-Nearest Neighbors	0.9886
Support Vector Machine	0.9557
Naive Bayes	0.9158

Table 5: Accuracy of Machine Learning Models for People with High Sugar

4 Summary

Our project aimed to develop and evaluate machine learning models for predicting the healthiness of food products based on ingredient analysis. With the increasing emphasis on healthy eating habits, there is a growing demand for accurate tools to distinguish between healthy and unhealthy food options. To address this need, we embarked on an extensive exploration of machine learning techniques applied to nutritional data.

We collected a diverse dataset containing nutritional information for various food products, including traditional items and fast food options. Additionally, we merged datasets from multiple sources to create a comprehensive dataset for analysis. Through collaboration with nutrition experts, we established threshold values for key nutrients to guide our classification framework.

Our study focused on four prominent machine learning algorithms: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes. We evaluated the performance of these models using metrics such as accuracy, precision, recall, and F1-score. Cross-validation procedures were conducted to assess the robustness and generalization capabilities of the models.

Each algorithm was tested on different demographic groups, including normal individuals, those with cholesterol issues, high-fat intake, and high-sugar intake. By tailoring our analysis to specific dietary needs and health concerns, we aimed to provide practical insights into the effectiveness of machine learning models in classifying food items for diverse

populations.

Overall, our project contributes to the development of tools that can empower consumers to make informed dietary choices, promote healthier eating habits, and address the growing challenges of nutrition-related health issues.

References

[1] BDA. Big data analytics. http://cs.rkmvu.ac.in/academics-msc-in-big-data-analytics-data-science/, 2016.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir Naumovich Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5 – 32, 2001.

[4] Soumita Kundu. Nutritionist at rsv hospital, tollygunge, kolkata, 2024.

[5] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.

[6] Frank Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Laboratory, 1961.

[7] Shubham Roy and Parimal Bera. My data of food. https://drive.google.com/drive/folders/1S1XCTZtUlO9cLMutgUzG3ATKmO-cO1cr?usp=drive_link, 2024.