**FindingHealthyFoodML: Predicting the Healthiest Fast Food Meal Options with K-Means Clustering and DBSCAN**

**Name: Tom Le**

**Department of Computer Science**

**Lamar University, Texas, USA**

**Email: letomcs@gmail.com**

**FindingHealthyFoodML: Predicting the Healthiest Fast Food Meal Options with K-Means Clustering and DBSCAN**

**Abstract— Fast food has always been a cheap and fast alternative for people who do not have the time to prepare a meal for themselves. While it is a cheap and convenient option, fast food is known to be unhealthy due to it being processed, which are foods that lack essential vitamins and nutrients. However, not all fast-food items are bad; there are always alternatives to the unhealthy, greasy menu items, whether it is a salad or a grilled chicken sandwich. There are many different aspects to consider when finding the right food choice, such as food options that have the most calories for weight gain and contain much less saturated fat or food options that contain the most protein with less sodium for those who have high blood pressure. The goal of this project is to find the best alternative healthy menu items from many well-known fast-food chains. To find the most optimal fast food menu items from these fast-food chains, K-Means and DBSCAN will be utilized in the machine learning model. With the usage of K-Means, classifying fast food meals from different fast-food chains as healthy or unhealthy based on calorie count, sodium levels, fat percentage, protein content, etc. Along with the usage of DBSCAN, identifying meals or outliers that do not fit well into any K-Means clusters, such as extremely high-calorie meal items, by labeling these outliers as noise instead of forcing them into the cluster. With these algorithms, this project aims to guide people in finding better fast-food options that can help people balance their dietary needs.**

1. **Introduction**

Fast food is known for its convenience due to its fast preparation and lower pricing. When you think of fast food, the first thing that probably comes to your mind is processed burgers, fries, pizza, or fried chicken. According to an article, “Eating fast food frequently had been accompanied by low consumption of fruits, vegetables and milk but high fats, carbohydrates and added sugars intake. Fast food meals are calorie-dense as they provide one-third to one half of daily energy but below quarter of micronutrients.” [6]. What most fast foods have in common is that they usually contain a lot of saturated fats and sugars without beneficial nutrition. However, many people forget that many of these foods are also unhealthy for other reasons, like their high sodium content: “The majority of online recipes tend to be unhealthy — with many, for example, having excess amounts of sodium.” [1]. Too much sodium in general is a contributor to increasing your blood pressure, which can lead to heart disease or worsen those who already have high blood pressure or heart disease. According to a survey, “The analysis considered population-wide data over time and regions. This study found a correlation between increases in sodium intake and hospital admissions over time.” [1].

Besides people looking for healthy alternatives at fast-food restaurants, the majority of people do not go to these places just because of convenience but because they like the taste of these sodium-packed, fatty foods: “They could avoid McDonald’s and foods like that, and go to…Soup Sergeant [a local eatery], because… you can get healthier choices [there]. But most of the time they don’t, they just choose to go to a fast-food restaurant, probably just because it tastes good. Lots of fat and salt.” [4]. So, you cannot prevent people from letting them eat foods that they want. Along with their affordability, they are a general go-to option for people, especially teenagers: “Fast food is significant to teenagers because it is one of the few types of food that teenagers can afford to purchase outside of the home and therefore (ostensibly) beyond the influence of their families.” [4]. Another article also states, “The problem of consuming fast food has dramatically increased among populations, especially adolescents and young adults. It has been implicated as a likely contributing factor to the growing obesity rates worldwide.” [8]. Daily fast-food consumption is not good for a person’s health, and it is especially bad as the number of daily consumers continues to increase among younger people, leading to a rising increase in obesity. However, this article states, “According to the National Restaurant Association, the number of adults who say they are trying to choose healthier items at restaurants is on the rise.” [7]. So, it has been evident that many fast-food restaurants are shifting to healthier options; even if their food menus still contain mostly unhealthy foods, it is still a positive shift with the increase of inclusions of healthier meal options.

With a machine model like the one in this project, it could help persuade consumers from proceeding to pick unhealthy options once they realize the ratio of negative to positive nutritional values. To achieve this goal of finding 'healthy' foods or at least 'moderate' foods, the use of two unsupervised machine learning algorithms, K-Means Clustering and DBSCAN, will be utilized in this project development to achieve this goal. With these algorithms, we would also know which foods to avoid when thinking about ordering from certain fast foods, with these foods labeled under the 'unhealthy' category. For the data collection, there are two datasets that are utilized for this project:

Kaggle Dataset:

<https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition>

USDA Economic Research Service:

<https://www.ers.usda.gov/data-products/food-consumption-nutrient-intakes-and-diet-quality>

The dataset extracted from Kaggle is called ‘fastfood.csv’ and its purpose is to provide data on each individual fast-food item from various fast-food chains, which includes all the necessary macronutrients and micronutrients of each food. The dataset extracted from USDA’s website is called ‘table-8-recommended-density-and-2017-2018-density.csv' and its purpose in this project is to show the recommended nutritional intake per fast-food meal item. With these two datasets, we can find whether each food item from the Kaggle dataset is considered ‘healthy,’ ‘moderate,’ or ‘unhealthy’ based on USDA’s dataset.

1. **Problem Statement**

In what ways can critical thinking improve the feature selection and evaluation process when determining the healthiness of fast-food menu items? When we think about foods to avoid, some people either think about how high the calories might be or how much fat it contains, but it’s not just one or the other; it is both, including other features like sodium and sugar. For instance, including features like saturated fat, sodium, fiber, sugar, and protein can give a much better understanding of a meal’s nutritional value to avoid misleading indicators. For instance, some people may think low-calorie meals are always good when they could be packed with high sodium or lacking in nutrients, or they may also believe that eating just fruits and vegetables is sufficient for your body, but it lacks protein and calcium from meats and dairy, which are needed to maintain a balanced, healthy diet.

What are the potential challenges in scaling a fast-food classification model across multiple restaurant chains? The nutritional content of each food item may change, and the datasets may need to constantly be updated to ensure accuracy, along with not exactly having every food item in the dataset. A challenge when building this model is the cross-implementation between two datasets used in this model. One dataset contains all the nutritional value of each food, and the other contains the necessary nutrition for fast-food intake by the USDA, but the metrics on each dataset differ, which requires additional careful calculations to ensure accurate scaling.

How can identifying outliers with DBSCAN improve the reliability of health-based food recommendations? DBSCAN helps improve the model’s reliability by ensuring that extreme data points don’t distort the overall clustering process. For example, if a single meal item is worth 2,000+ calories with high sodium content, it can shift the centroids of the K-Means clusters. This would cause healthier menu items to be misclassified into a less accurate group. So, with DBSCAN, labeling these extreme meals as ‘noise' separates them from the main clusters, which helps improve the integrity of 'healthy' and 'unhealthy' clusters.

* RQ1: In what ways can critical thinking improve the feature selection and evaluation process when determining the healthiness of fast-food menu items?
* RQ2: What are the potential challenges in scaling a fast-food classification model across multiple restaurant chains
* RQ3: How can identifying outliers with DBSCAN improve the reliability of health-based food recommendations?

This project presents several key contributions:

1. The project utilizes K-Means Clustering to group fast-food items in either ‘healthy,' 'moderate,' or 'unhealthy' clusters.
2. This model preserves clusters integrity with DBSCAN, removing 'noise.’
3. Allows people to find fast-food menu items that best fit their nutritional needs.
4. **Proposed Methodology**

The K-Means Clustering algorithm groups the fast-food menu items based on their feature similarities, such as calories, protein, fat, sodium, and sugar content. By analyzing these clusters, we can classify and distinguish them into groups as 'healthy,' 'moderate,’ or ‘unhealthy,' which is a major factor of this model to find healthy food items.

On the other hand, DBSCAN is an algorithm used in this model to find foods that do not fit into the classification of other items. These outliers might include items with extremely high sodium or calorie levels that could distort the clusters. By treating them as noise, DBSCAN helps preserve the integrity of the clusters and highlights meals that consumers should potentially avoid.

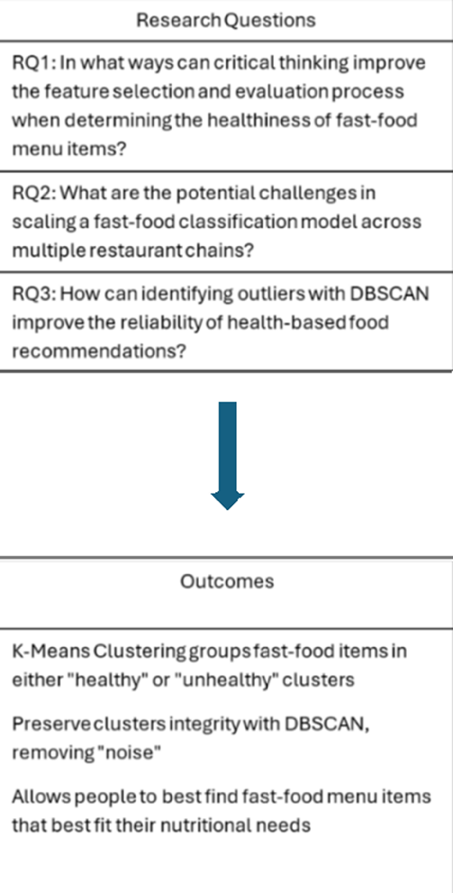


Fig. 1 Research Questions.

The flow diagram gives a visual step of the algorithms used in this project. It begins with downloading and loading the datasets, which include a food dataset and a recommended food intake dataset. This is followed by preprocessing the data to ensure it is clean and ready for analysis. The next step involves scaling the nutritional metrics according to the recommended intake values to standardize the data. After scaling, K-Means clustering is applied to classify foods based on their healthiness. Following the classification, DBSCAN is employed for outlier detection to identify foods that significantly deviate from typical patterns. Finally, the process concludes with displaying the results. The arrows connecting these nodes illustrate the sequential flow of operations, making the entire methodology easy to understand at a glance. The flowchart, as shown in Fig. 2, effectively illustrates the framework of the proposed classification and detection system.

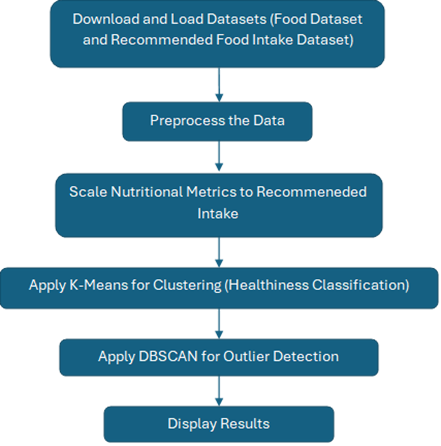


Fig. 2: Flow Chart of the Proposed Methodology.

1. **Results and Discussions**

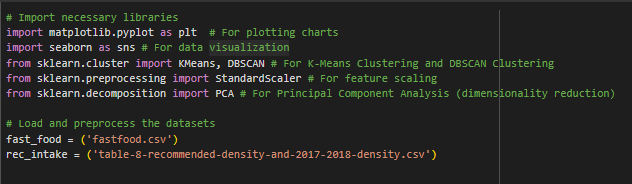
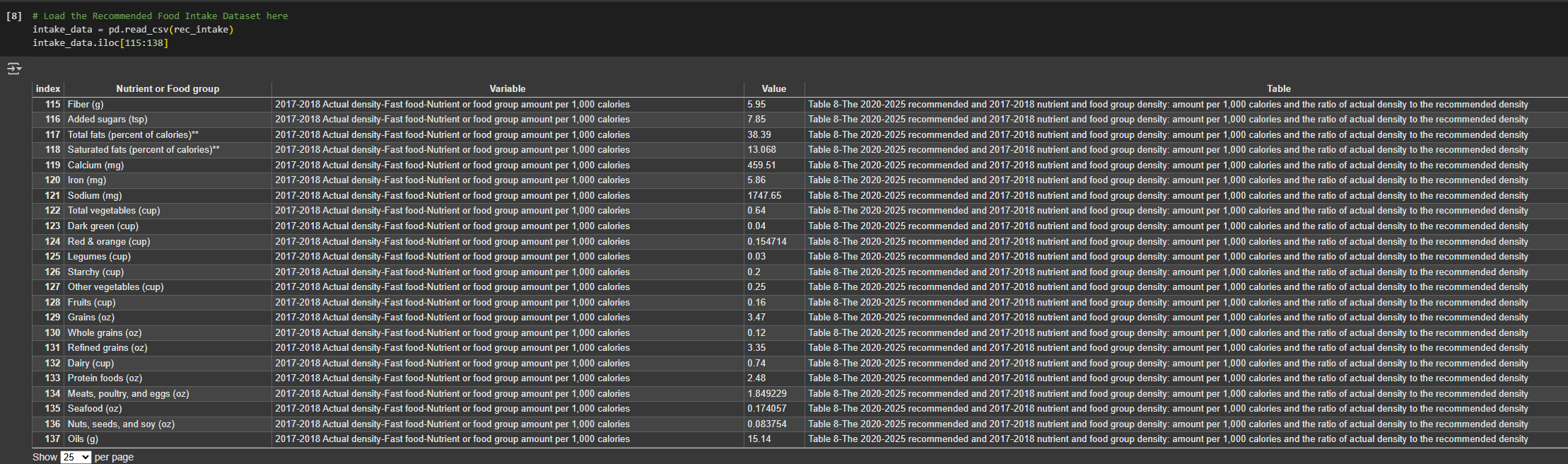
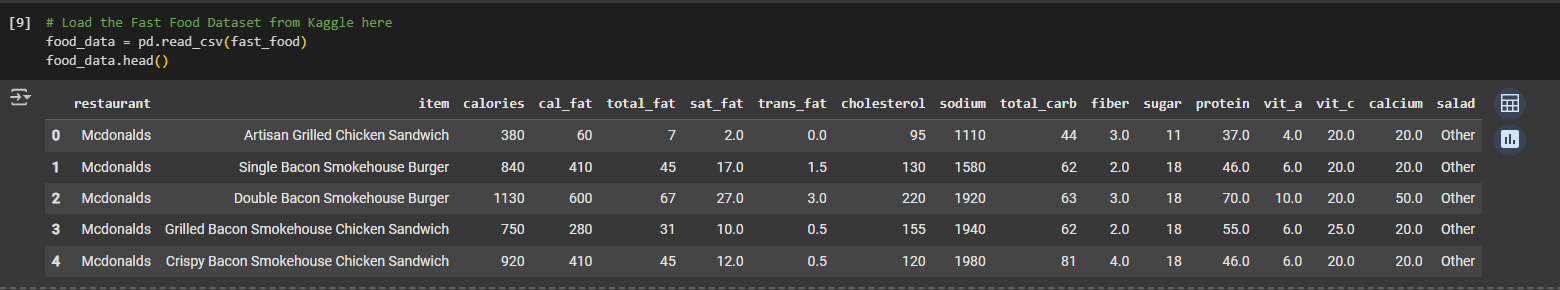


Fig. 3: Importing Necessary Packages and Loading the Datasets

First, we begin with importing the necessary packages as shown in Fig. 3: matplotlib, seaborn, KMeans, DBSCAN, StandardScaler, and PCA. Afterwards, we load the datasets ‘fastfood.csv’ downloaded from Kaggle and 'table-8-recommended-density-and-2017-2018-density.csv' from USDA’s official website.

Fig. 4: Loading the USDA Dataset

Fig. 5: Loading the Kaggle Dataset

In both Fig. 4 and Fig. 5, they are then loading the datasets gathered from Kaggle and USDA.

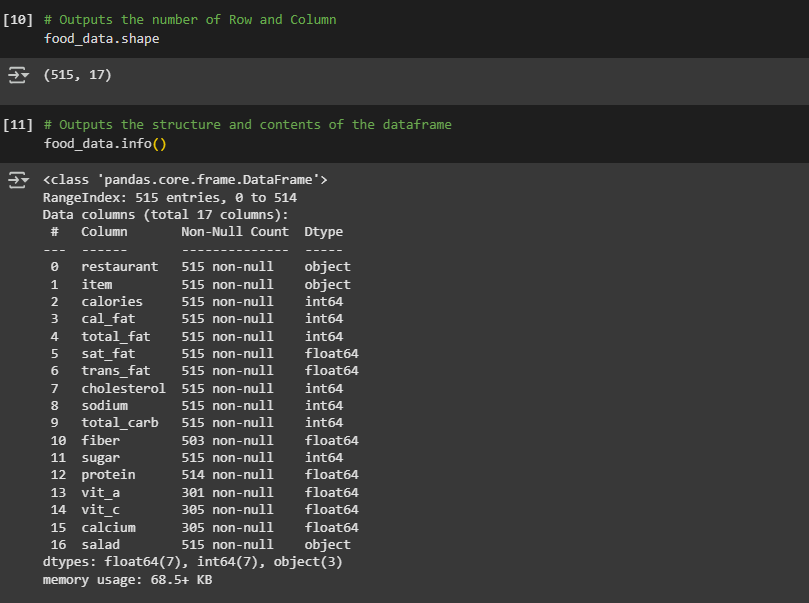


Fig. 6: Beginning of Data Preprocessing

Afterwards, start with basic preprocessing of the data. In Fig. 6, it begins by first finding the size (number of rows and columns) of the dataset. Then outputting the structure and contents of the data frame.

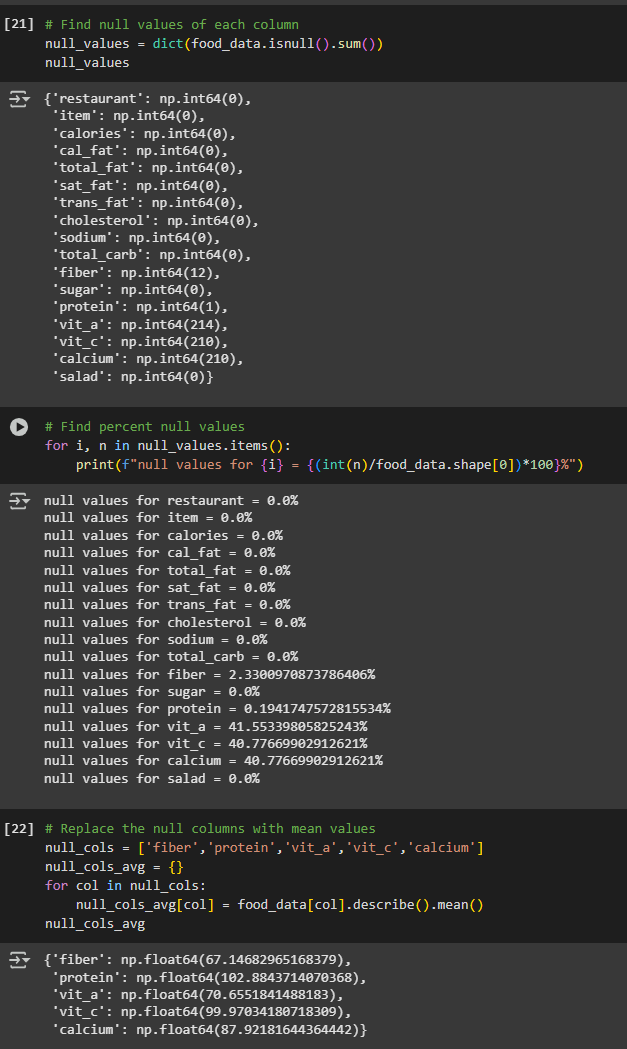


Fig. 7: Finding and Replacing Null Values

The next step in preprocessing this data is to find the null values of the dataset. Then display the percentages of how many null values there are in each column. It is seen in Fig. 7 that the null values present in the dataset are under columns fiber, protein, vitamin A, vitamin C, and sodium, with most of the null values being under vitamin A and vitamin C. To fill in the null values, it is necessary to find the means of each column containing null values.

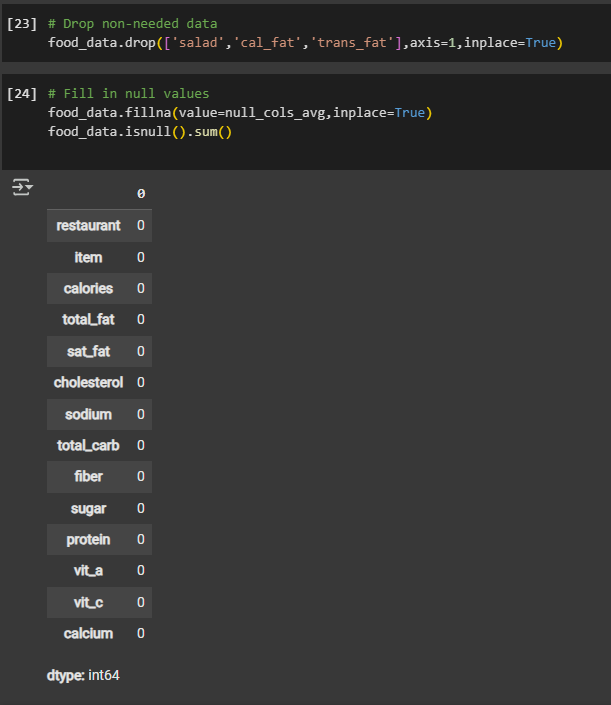


Fig. 8: Dropping Unneeded Data and Filling in Null Values

Here, shown in Fig. 8, it is decided that ‘salad,’ ‘calories from fat,’ and ‘trans fat’ are not needed data for this project, so they have been dropped. Next is to fill in the missing null values for each column.

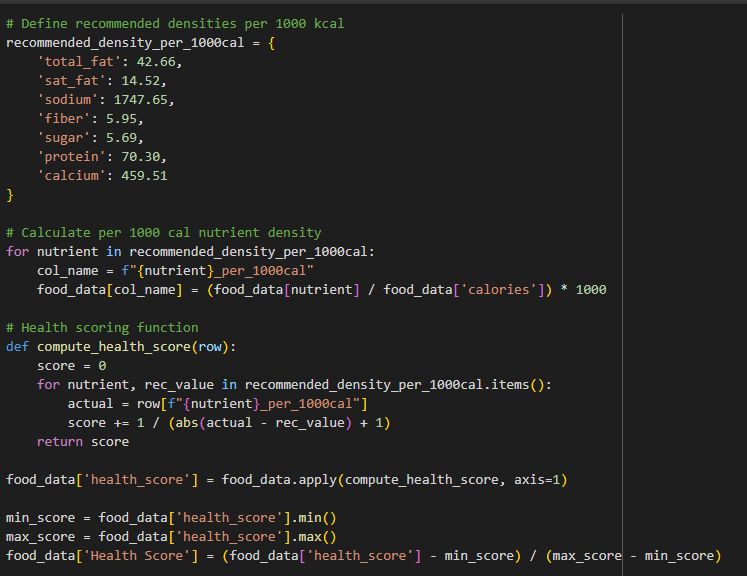


Fig. 9: Finding Health Score

This is where the real development begins after preprocessing the data. At the start of Fig. 9, it first shows the recommended density per 1000 calories. This is the information extracted from the USDA’s dataset, which displays the recommended number of macronutrients and micronutrients per 1000 calories of a meal. With this information, we can scale each food’s nutritional value from the food dataset based on USDA’s dataset to see if each food meets or is close to meeting the recommended fast-food nutritional intake. After scaling each nutritional value of a food item, we can then calculate the health score using this equation:

Health Score = 1 / ((actual value – recommended value) + 1)

Returning the results of this will give us the needed health score to determine if it is close to meeting the recommended fast-food nutritional intake. Afterwards, to make the health score more understandable for users, the score is scaled down to between 0 and 1, with 0 being ‘Not Ideal’ and 1 being ‘Ideal’. So, if a food health score is closer to 0, it is nowhere near the recommended fast-food nutritional intake, while if a food health score is closer to 1, it means that it is near the recommended fast-food nutritional intake.

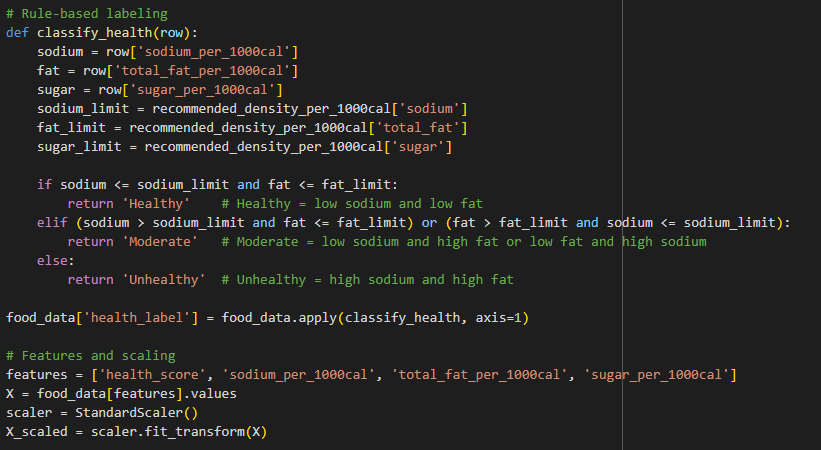


Fig. 10: Rule-Based Labeling and Feature Scaling

As shown in Fig. 10, we can finally classify a fast-food item to be either ‘healthy,’ ‘moderate,’ or ‘unhealthy’. To do this, we start by initializing the sodium, fat, and sugar, since these rows determine the classification of a food item. Then, initialize the ‘sodium limit,’ ‘fat limit,’ and ‘sugar limit’ to give a comparison between the food’s negative nutritional value and the recommended nutritional value. Since fats and sodium are the most impactful value in determining the food’s classification, the sugar value has been left out to give a better overall classification. To determine if a food is ‘healthy,’ it would have less than or equal to the fats and sodium of the recommended fast-food nutritional value. For a food to be ‘moderate,’ it would need to either have fats or sodium less than the recommended fast-food nutritional value. For a food to be classified as ‘unhealthy,’ both the fat and sodium values would have to exceed the recommended fast-food nutritional values. Following this, the implementation of the features ‘health ‘health score,’ ‘sodium per 1000 calories,’ ‘total fat per 1000 calories,’ and ‘sugar per 1000 calories’ is needed for clustering, and to properly cluster the food items based on these features, we need to scale them using the StandardScaler method.

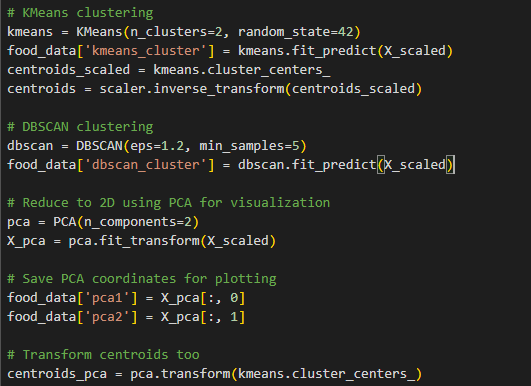


Fig. 11: Applying K-Means Clustering and DBSCAN Algorithms

Here, displayed in Fig. 11, we get to finally apply the two algorithms K-Means Clustering and DBSCAN. With K-Means Clustering, it displays clusters into two initial groups, which can either be ‘healthy’ or ‘unhealthy’ based on features. With the DBSCAN algorithm, it will display the main cluster, separating the outliers, which are considered noisy data. Next, since we want to consider all the features for ‘x’ and ‘y’ instead of two elements, PCA is implemented to reduce the number of dimensions when trying to graph the data. Lastly, the inclusion of centroids is not forgotten for the K-Means clusters to show the center points of each cluster.

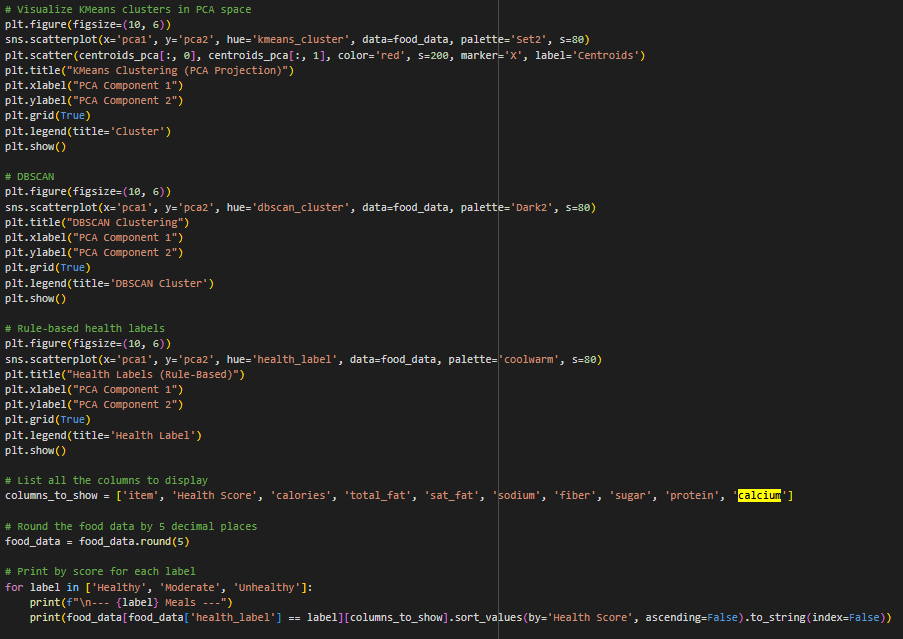


Fig. 12: Visualizing the Applied Algorithms

Now we get to visualize all the algorithms of this project with Seaborn and plot the points with Matplotlib. Besides K-Means Clustering and DBSCAN, the rule-based labeling system is also displayed for the most accurate classification of each food item, which shows all the foods under the ‘healthy,’ ‘moderate,’ and ‘unhealthy’ categories.

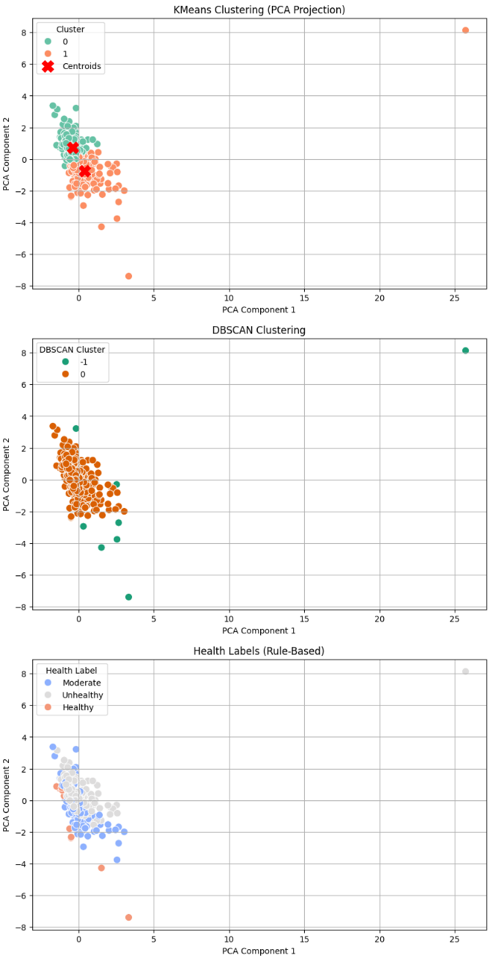


Fig. 13: K-Means Clustering, DBSCAN Clustering, and Health Labels Results

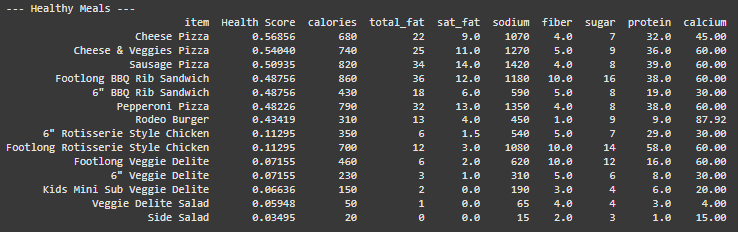


Fig. 14: Sample Healthy Foods and Health Scores Output



Fig. 15: Sample Moderate Foods and Health Scores Output

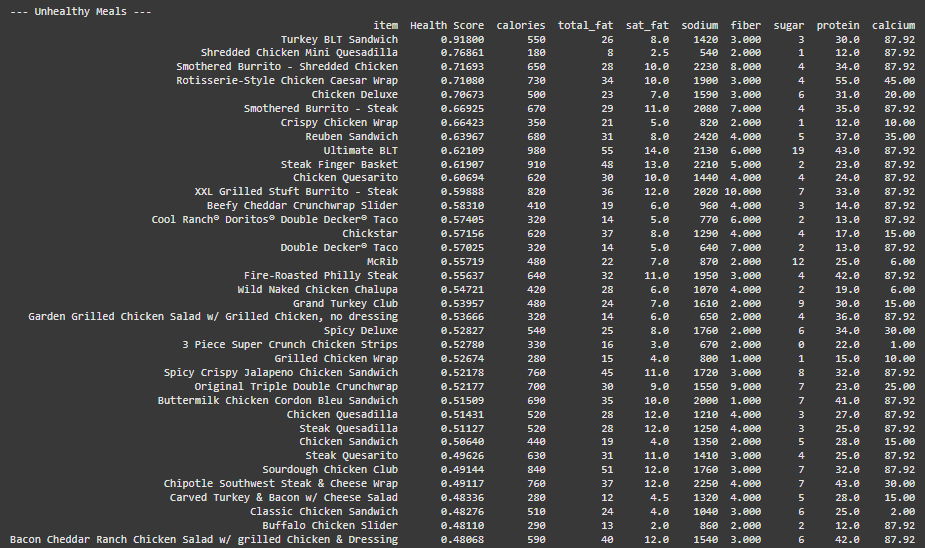


Fig. 16: Sample Unhealthy Foods and Health Scores Output

1. **Software and Hardware Requirements**

In designing this project, the necessary software includes the latest Python version and an integrated development environment (IDE) such as Visual Studio Code. There are not many hardware requirements at all. All that was needed was a computer that can run Python or is able to open Google Colab on a browser.

1. **Development Challenges**

There were many challenges within the development of this project. The first challenge was inexperience, since there was no prior experience with machine learning algorithms. It was quite a hurdle to get started on the implementation with K-Means Clustering and DBSCAN. So, it took extra time trying to learn the syntax and understand how each algorithm can be applied to the machine learning model, especially working with unsupervised datasets.

The second challenge was to figure out how to scale the food items in the ‘fastfood.csv’ dataset with USDA’s recommended nutritional intake dataset, for the metric units in USDA’s dataset are not in either grams or milligrams, which required an extra step of conversion. Converting each micronutrient to grams or milligrams by hand and initializing them in the code became the easiest approach to work with.

The third challenge was trying to figure out how to determine how to classify whether a food is either ‘healthy’ or ‘unhealthy,’ which turned out to be a big challenge. The initial approach was to classify food based on saturated fats, sodium, and sugar, since fast food is known to be high in these amounts, contributing to unhealthy diets. However, it was an issue with menu items like salads not being classified as healthy because they could contain a high ratio of sodium, which counteracts the other beneficial micronutrients (protein, fiber, and calcium) if the consideration was just the negative nutrients and not the positive nutrients. The second approach to this issue was determining ‘healthy’ and ‘unhealthy’ foods based on overall nutritional contents, meaning considering both positive and negative nutrients. How this is done is basically, if the food item’s positives outweigh the negatives, then it would be grouped with the ‘healthy’ cluster, but if its negatives outweigh its positives, it would be grouped in the ‘unhealthy’ cluster. This way, a food item is not disregarded as ‘unhealthy’ due to its few drawbacks when it delivers more positives, but you should still consider the amount of saturated fats, sodium, and sugar of the item before considering consuming it. For example, if a person with diabetes were to pick out a food item within the ‘healthy’ cluster, then they would look at the sodium and sugar amount first before making the decision to choose that item. So the implementation of both methods in the machine learning model to classify foods as being either ‘healthy,’ ‘moderate,’ or ‘unhealthy’ based on their total fats and sodium levels along with displaying the overall nutritional value by adding a ‘Health Score,’ which is supposed to show how close the food item is to a recommended fast food meal. This means that food like salads will usually have low scores because salads usually lack nutrition like protein and fats. While foods such as the ‘chicken soft taco’ from Taco Bell will have a perfect score because it fulfills every nutritional requirement, its only drawback is its high sodium content, which is why it would be classified under ‘moderate’ and not under ‘healthy.’ Following this issue is trying to figure out how to plot the ‘x’ and ‘y’ labels for visualization, since there were three micronutrients to consider rather than just two. This caused the plot points to be very messy and made the plot points look like a lot of individual points rather than cluster groups. This is due to many food items varying in amounts of sodium, saturated fats, and sugars. For example, a salad is typically healthy but could contain very high amounts of sodium, or a burger that is very well within the recommended nutritional intake could contain high amounts of saturated fats. This issue was solved through the implementation of the PCA (Principal Component Analysis), which helped with the ‘x’ and ‘y’ labels because it reduced the number of dimensions from sodium, saturated fats, and sugar to just two principal components that capture the most significant variance in the data. PCA not only made the clusters more distinct by separating the clusters into groups more clearly, but it also made the individual messy points more interpretable.

1. **Conclusion**

This project aimed to develop a machine learning model to predict the healthiest fast-food options based on macro and micro nutritional content. It is known that many fast-food options are known to be unhealthy because of their high contents of saturated fats, sodium, or sugar. Throughout the development of this machine learning model, many issues were encountered, from inexperience to finding the best methods to cluster each food menu item. The machine learning model used two algorithms, K-Means Clustering and DBSCAN, to filter and classify these fast-food items based on their nutritional content along with the inclusion of a rule-based label to give further accurate findings. This project resulted in not only finding all the ‘healthy,’ ‘moderate,’ and ‘unhealthy’ foods from the provided dataset, but it also listed the ‘Health Scores’ of each food, showing how far each meal is from a complete nutritional meal. In the end, the goal of the project was achieved, but there are still many improvements to be made.

1. **Future Work**

There was no nutritional expertise when working on this project, so it can be further improved with better calculations for improved accuracy in determining healthy and unhealthy foods. An improvement idea is to factor in more micronutrients, such as vitamins A and C, to further improve the food predictability.

Implementing a recommender system for users with specific dietary needs is another addition to be made. For example, recommend foods to people who are looking to gain weight, lose weight, diabetics, fulfill daily nutritional needs, etc. This study illustrates two types of recommender systems with “The first category uses similarity measures to recommend recipes, which are most like the meals that user likes.” [3]. This type of system recommends foods based on users’ preferences but is not exactly ideal for those finding healthier options at fast-food chains. The second recommender system suggests, “For the second category, both user’s likes and dislikes are important, this category focuses more on the user’s nutritional needs.” [3]. With this recommender system, it suggests that some unhealthy foods be replaced with healthier options. This is the ideal type of recommender system that this project aims for further future improvements.

We could also further improve accurate food intake tracking with wearable devices: “These devices can track parameters such as chewing patterns, swallowing frequency, and even digestion-related data to estimate food intake. Some examples include smartwatches, neck-worn sensors, and ingestible sensors that can monitor the ingestion and digestion of specific food items.” [2]. With technology like this, it can greatly improve nutritional intake tracking from fast-food items, which saves time for people tracking their meals or, in general, allows the average users who do not track their meals to know the impacts on their health from the kind of food they are consuming.

Additionally, the implementation of image detection would also further improve classification and nutritional tracking. An example of this sort of implementation is a mobile application called ‘SnapCalorie’. The main purpose of this application is to find an accurate approximate nutritional value and calories just by taking a picture of food. One of the main issues with the development of this project is that many foods naturally have imbalances in the dataset, which made it more complicated to classify foods based on nutritional value. However, this study suggests a method that combines visual recognition and multi-stage transfer learning: “This is accomplished by iteratively clustering visually similar food items and updating the Convolutional Neural Network (CNN) model through multi-stage transfer learning during the training phase. During the validation and testing phases, the trained model in the multi-stage transfer learning stage is applied directly for food image classification.” [5]. By integrating CNN-based image recognition, the model’s clustering classification can be improved through nutritional characteristics from visual features, leading to more accurate testing, validation, and classification of fast-food meals.

References

[1] D. Zhu, Z. Lian, "Parking robot based on fuzzy reasoning and parking big data" Academic Search Complete. <https://research-ebsco-com.libproxy.lamar.edu/c/2rtsa2/search/details/fmfvcbbnxb?db=a9h&isDashboardExpanded=true&limiters=None&q=Parking%20robot%20based%20on%20fuzzy%20reasoning%20and%20parking%20big%20data> (accessed Feb. 14, 2025)

[2] H. Canlı, S. Toklu, "Design and Implementation of a Prediction Approach Using Big Data and Deep Learning Techniques for Parking Occupancy" Academic Search Complete. <https://research-ebsco-com.libproxy.lamar.edu/c/2rtsa2/search/details/iaqppqbzvn?db=a9h&isDashboardExpanded=true&limiters=None&q=Design%20and%20implementation%20of%20a%20prediction%20approach%20using> (accessed Feb 14. 2025)

[3] D. Zhu, "Deep learning over IoT big data-based ubiquitous parking guidance robot for parking near destination especially hospital" Academic Search Complete. https://research-ebsco-com.libproxy.lamar.edu/c/2rtsa2/search/details/uhsparkxv5?db=a9h&isDashboardExpanded=true&limiters=None&q=deep%20learning%20over%20iot%20big%20data (accessed Feb. 14, 2025)