Fruitify: A Fruit Classifier

Da Yuan Zhao
Undergraduate at The City College of New York
dzhao000@citymail.cuny.edu

Abstract

Fruits have similar chemical composition with unique amounts of macro and micro nutrients. They can be classified into 7 different categories of fruits that are distinctive in the numerical values of some nutrients. This project aims to develop a machine learning model to classify fruits based on their chemical composition that will be used to create a dietary tool where users can find a type of fruit that fits their best needs. The chemical composition of fruits is obtained through the SR Legacy Food dataset from the USDA FoodData Central and the data is analyzed through several methods using a means and variance approach. Through training, optimization and hyper parameter tuning, a random forest model was developed at an accuracy rate of around 79%.

Table of Contents

1 Introduction

There are many types of fruits and each one has a unique chemical composition of different macro and micro nutrients. Initial plans for this project was for a classification of individual fruits. However, due to the limitations of the data that is required to train the model, the plan was reduced to categories of fruits. This project aims to help people that want to get their fruit intake that suits their nutritional needs down to the digits.

2 Data

2.1 Dataset

The data that is used in this project is from the USDA Food Central database. Concerning fruits, this database contains data for many different forms of fruits such as drinks, concentrates, dehydrated and processed. To maintain standardization of the samples, only raw and unaltered fruits will be used.

There are 5 different dataset in the database. Initial research into this database favored the “Foundations Foods” dataset because there was direct access to raw samples of fruits. However, further examination into this dataset yielded 2 problems: insufficient number of distinct fruits and incomplete dataset. There were only 16 unique fruits which is not enough for the task. Certain samples of some fruits were also missing important nutrients such as carbohydrate, protein, and dietary fiber. The “SR Legacy Foods” dataset was more sufficient with 71 unique fruits but lacked raw samples and there was some missing data. It was ultimately decided that the “SR Legacy Foods” dataset will be used with a total of 87 samples, including different species of some fruits due to its larger number of unique fruits. Initial testing of classifier models with each unique fruit having their own class proved to be futile as the accuracy rate was too low. These samples would be classified into 7 groups. It was decided that fruits would be grouped into 7 classes.

2.2 Classes

These 7 classes are aggregate, berry, citrus, drupe, melon, multiple, pome. A list of all fruits in their respective categories can be found in Appendix A. Examination of these categories of fruits indicates that there is an imbalance of data with many of the samples in the berry category. This may prove to be troublesome in the evaluation of some machine learning models and may require techniques to handle it.

2.3 Data Analysis

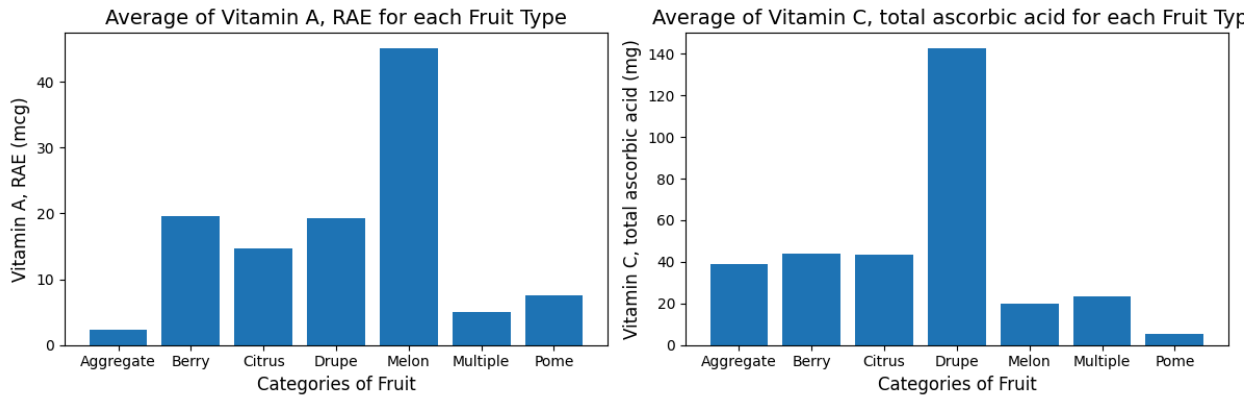


Figure 1: Bar graphs showing the average amount of Vitamin A and Vitamin C for each fruit category. These calculations of averages exclude NaN values.

Through analyses of the averages of each nutrient for each fruit category, it can be concluded that some categories are unique. In Figure 1, the melon category has the highest amount of Vitamin A while the drupe category has the highest amount of Vitamin C. The averages of other nutrients also support this observation and can be found in Appendix B. This finding gives confidence that a model can distinguish between different classes.

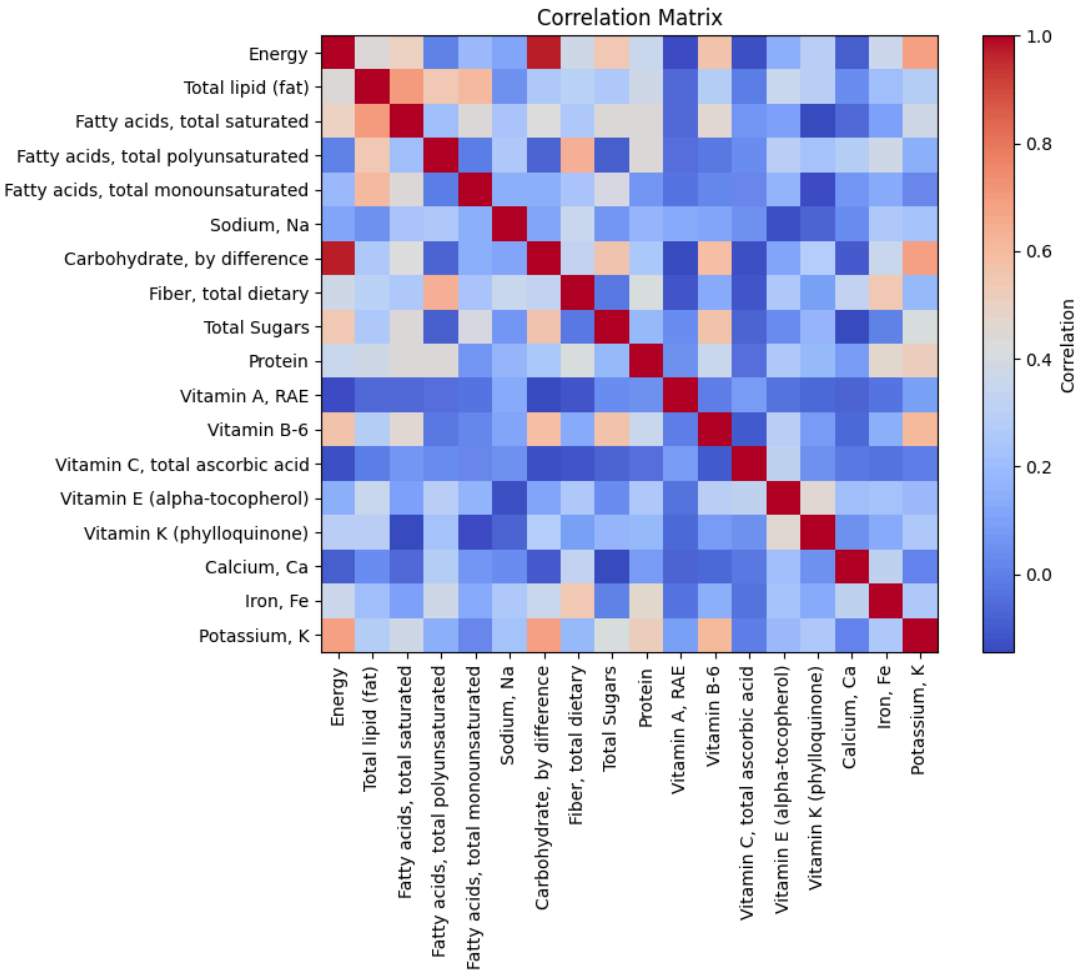


Figure 2: Heat map showing the correlation coefficient between all nutrients. The calculations of these values exclude NaN values.

Figure 2 shows some surprising correlations between some nutrients. Carbohydrates and calories (energy) are highly correlated followed by potassium and energy and potassium and carbohydrates. Although these findings are not considered when building the model, it helps to build a better understanding of data.

2.4 Missing Data

The dataset is not complete and has some missing values. The average missing data is around 11.7%. There are several ways to deal with missing data: replace with averages, replace with 0, replace using values predicted by KNN, or ignore it.

Method 1 (replacing NaN values with averages) may be misleading because it distorts the true nature of the data that may include outliers. Thus, training a model on this method may be vulnerable to inaccuracies when presented with anomalies. However, previous findings may mitigate this issue since some categories of fruits are unique in terms of their nutrients.

Method 2 (Replacing NaN values with 0), in theory, is a bad practice since the introduction of artificial data, not supported by anything, is very misleading. This method is used because it was not clear if the missing data was a result of an absence of the nutrient in a sample.

Method 3 (Replacing NaN values using values predicted by k-nearest neighbor imputation) or method 4 (Ignore the missing data) are the more reliable methods as it preserves the raw data.

3 Model

3.1 Type of Classifier

There are many types of classifier models in machine learning. For this project, simple classifiers like decision trees, random forest, support vector machine (SVM), stochastic gradient descent (SGD), and k-nearest neighbors (KNN) are used. A train/test split was not used to train the data due to the size of the dataset being too small. Instead, a 4 stratified folds cross validation is used to calculate the accuracy of all tests in this project.

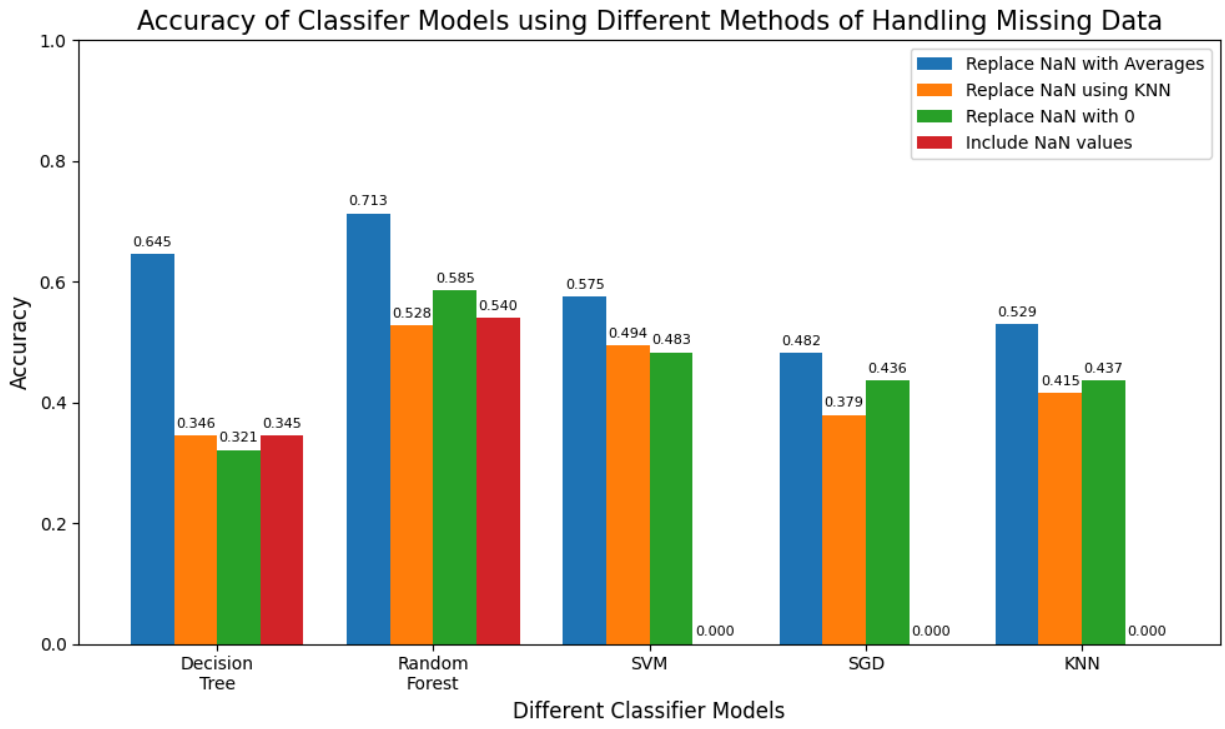


Figure 2: Bar graph shows the accuracy rate of different classifier models using default parameters for each method of handling NaN values. Note that Method 4 wasn't tested for SVM, SGD, and KNN due to the nature of these models.

On average, Method 1 performed 14.6% better than other methods. For decision tree and random forest, Method 1 performed 16.2% and 30.7% better than the other methods, respectively. The performance of Method 1 suggests that the model is better at predicting classes based on their averages and may be sensitive to outliers. The random forest model had the highest accuracy rate and the behavior of it works well with imbalance data. Going forwards, the random forest model with method 1 will be used.

3.2 Sample Subsets

The dataset has a total of 87 samples with 70 unique fruits. Four subsets of the dataset were created to be evaluated with the random forest model.

The first subset contains all fruits and their species. Given our small number of samples, this subset exists to maximize the number of data to train the model. The second subset contains all species and all fruits excluding avocados because they are the largest outlier concerning total fats as shown in Appendix C. Avocados from California are the biggest outlier followed by avocados from Florida. The third subset contains all fruits with their most commonly consumed species. The last subset is the intersection between the second and third subset.

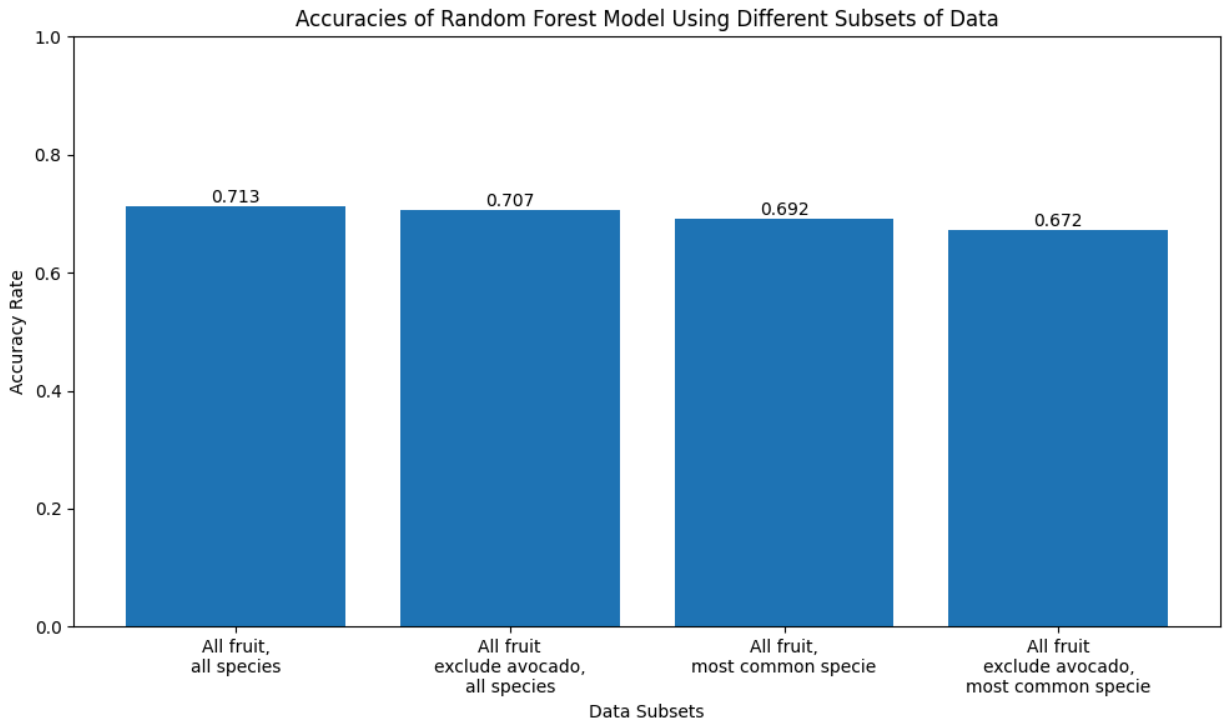


Figure 3: Bar graph shows the accuracy rate of the random forest model, using default parameters, on different subsets of the samples.

3.3 Feature Subsets

There are 18 features or nutrients, to be considered when training the model. Due to the nature of the random forest model, it can handle and perform better than other models when dealing with high dimensional data. However, it is not a bad idea to perform feature selection to maximize the accuracy rate. Thus, several methods of feature selection are used: selection by random forest, selection by variance, selection using kBest, selection by FDA required nutrients.

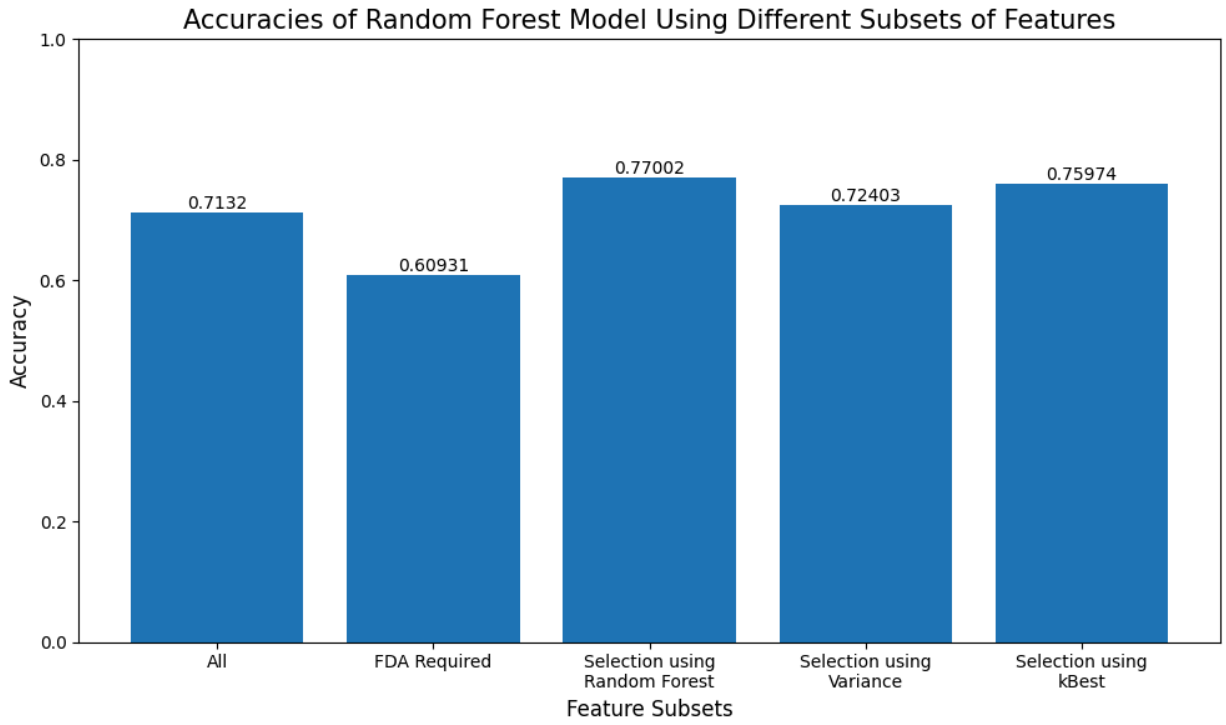


Figure 4: Bar graph shows the accuracy rate of the random forest model, using default parameters, on different subsets of features.

3.3.1 Feature Selection Using MDI

The random forest model from the Sci Kit Learn library provides a means to rank features based on the mean decrease impurity (MDI). This measure is used to calculate the change in impurity when a split occurs for that feature. The higher the MDI, the greater the importance of the feature is to the model.

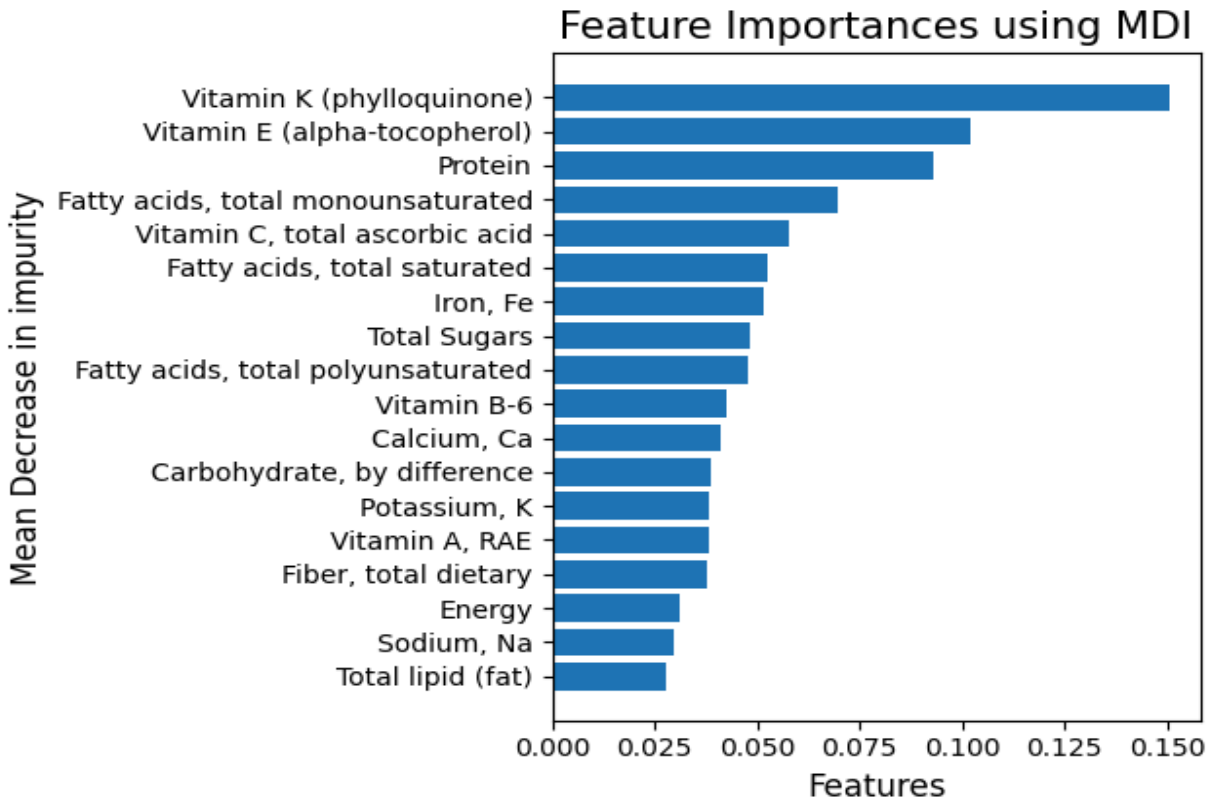


Figure 5: Bar graph shows the ranking of each feature by their importance.

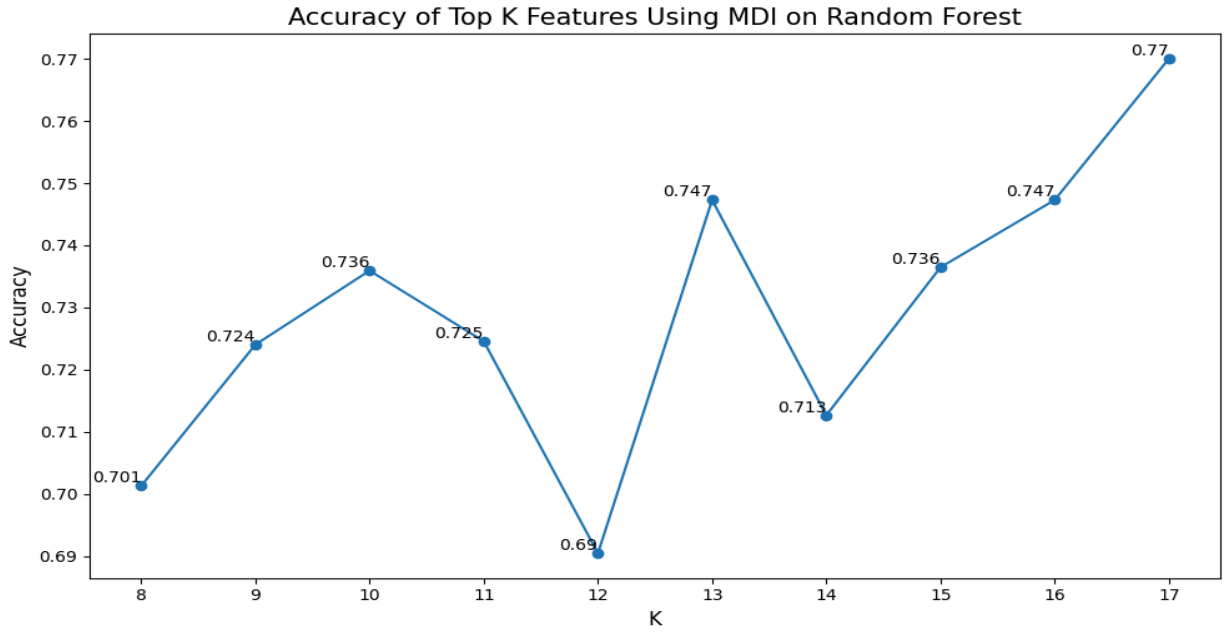


Figure 6: Line graph shows the accuracy rate of the top k features using the MDI on a random forest model.

Figure 5 illustrates that vitamin k is the most important while total fats is the least important feature while training the random forest model. Using this information, an algorithm was conducted on several subsets of features using the top k most important feature. Figure 6 demonstrates that training the random forest model on the top 17 features yielded the highest accuracy at around 77%. This method of feature selection includes every feature except the total amount fat.

3.3.2 Feature Selection Using Variance

This method selects features whose variance is greater than the threshold at 0.8. Using this method favors features with greater outliers than features that don't. This method of feature selection includes 12 features[1]. It is unsurprising that total fat is included in this method due to the extremities of avocados.

3.3.3 Feature Selection Using kBest

Using scikit-learn's f_classif score function, this method is similar to feature selection using variance in the sense that both methods utilize variance. However, the f_classif score function considers the variance in relationship to the classes. It computes the ANOVA F-value between the feature and the classes. This method selects the top k feature with the highest f_ckassif score.

[1] Energy, total fat, monounsaturated fat, sodium, carbohydrate, dietary fiber, total sugars, vitamin A, vitamin C, vitamin K, calcium potassium

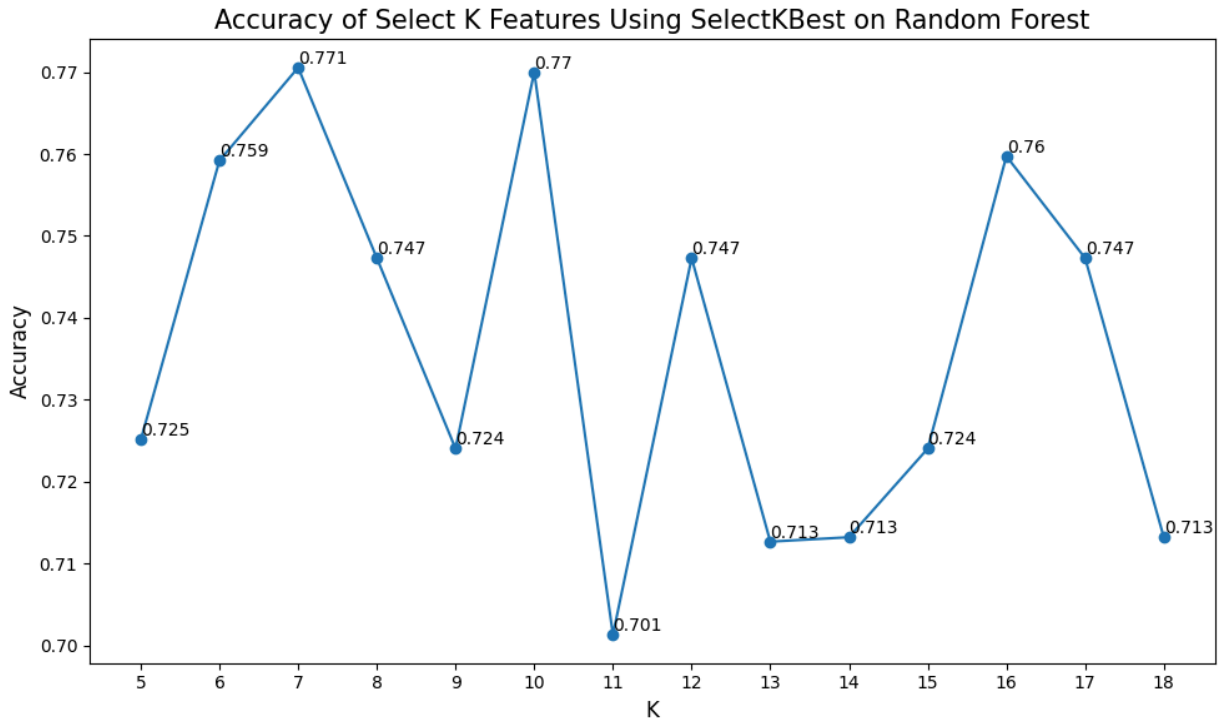


Figure 7: Line graph shows the accuracy rate of the top k features using the f_classif score function.

Figure 7 illustrates that the highest accuracies are at k = 7 and k = 10. However, this number of features is insufficient for the task. Thus, the next highest accuracy at k = 16 is used. This method of feature selection includes every feature besides total fat and vitamin C.

3.3.4 Feature Selection by FDA Required Nutrients

The FDA requires 14 select nutrients, including calories, to be listed on a US Nutrition Facts Label. This selection does not include trans fat, cholesterol, vitamin D

because these nutrients are absent from all of the fruits in the dataset. This method of feature selection includes 11 features[2].

3.4 Optimizing Sample Subsets and Feature Subsets

Combinations of sample subsets and features subsets are tested to find the most accurate when training a random forest model.

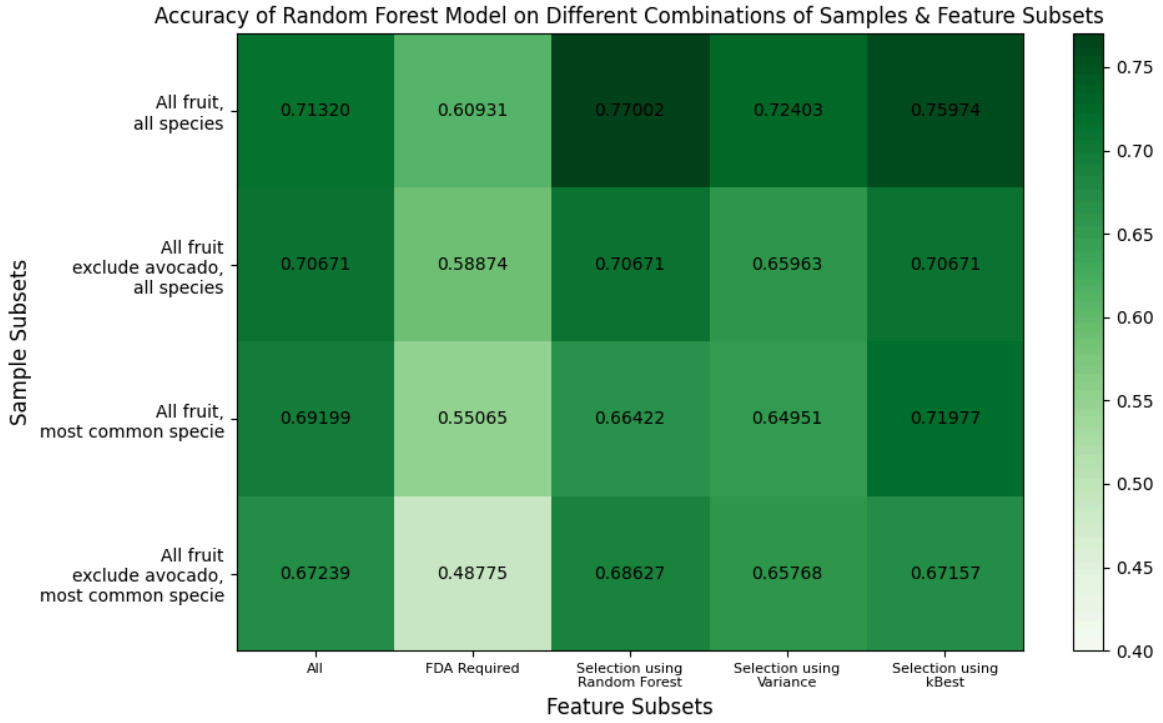


Figure 8: Heat map shows the accuracy rate of different combinations of sample and feature subsets.

It is unsurprising that the accuracy rates using FDA required features performed the worst because the features were not selected via a statistical means of measure. As shown in Figure 8, the highest accuracy combination of sample and feature subset uses

[2] Energy, total fat, saturated fat, sodium, carbohydrate, dietary fiber, total sugars, protein, calcium, iron, potassium

all fruits with all species and feature selection using MDI at an accuracy of around 77%. Moving forward, the model will be trained with this combination of data.

3.5 Random Forest Hyper Parameter Optimization

Given the relatively small size of the dataset, a grid search can be performed since the process won't be too computationally expensive.

```
parameters = {'n_estimators' : [100, 150, 175, 200],
              'criterion' : ['entropy', 'gini', 'log_loss'],
              'max_depth': [1, 2, 3, 4, 5],
              'max_features': ['log2'],
              'min_samples_leaf': [1, 2, 3],
              }
```

Figure 9: Hyper parameters tested using grid search.

The best optimized hyper parameters with an accuracy of around 79.33%: RandomForestClassifier(max_depth=5, max_features='log2', n_estimators=175, random_state=11)

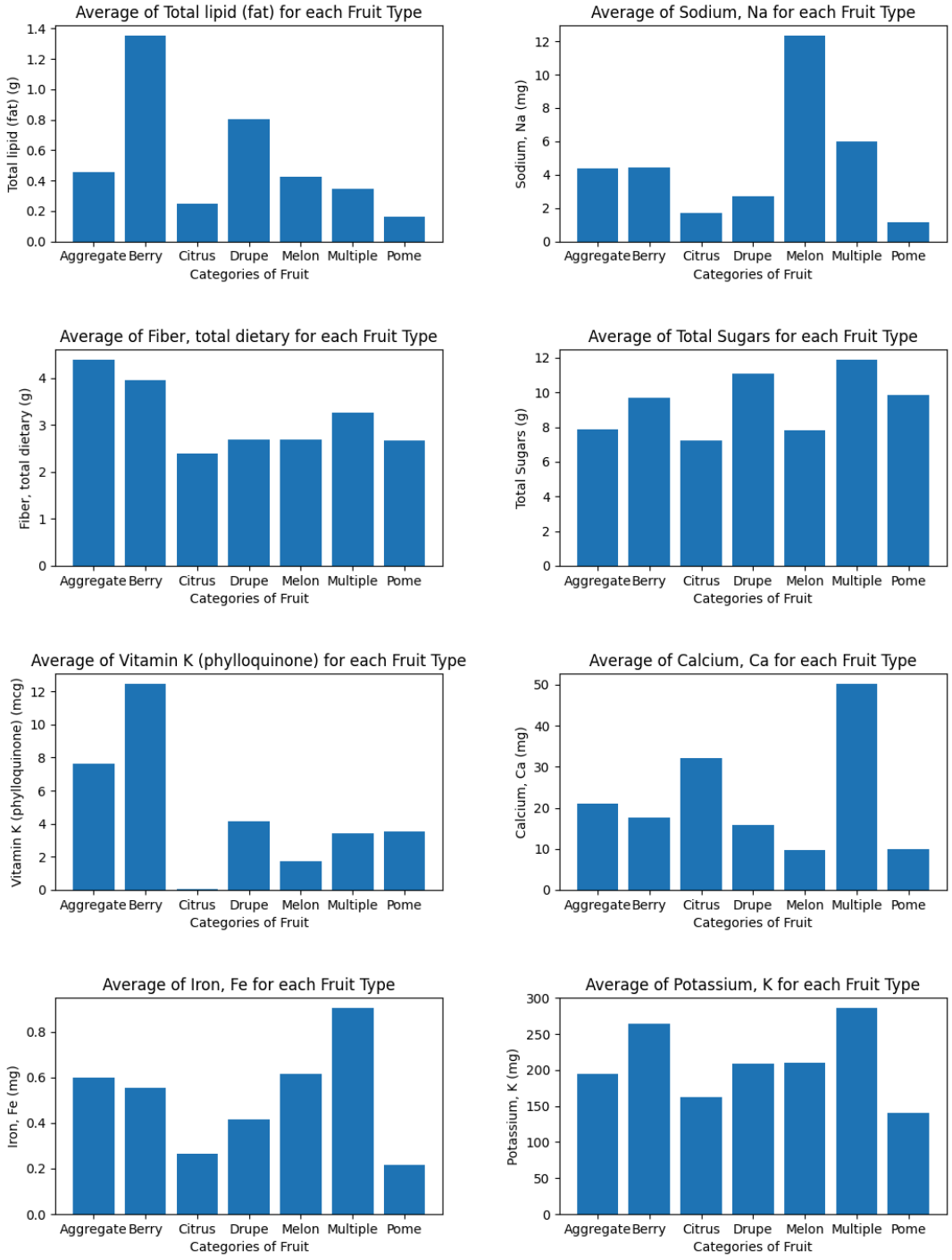
4 Conclusion

The purpose of this project is to develop a machine learning model to predict the category of fruit given macro and micronutrients by the user. Analysis of the samples revealed that certain classes had higher amounts of nutrients than others. Optimization of the data produced a combination of subsets that had the highest accuracy: replacing NaN with averages, using all samples, and selecting features using MDI score. Lastly, hyper parameters were tuned and yielded a classifier model with an accuracy of around 79%. Future research and development can scale this project to classify more broader categories of food such as dairy, meat, vegetables, etc.

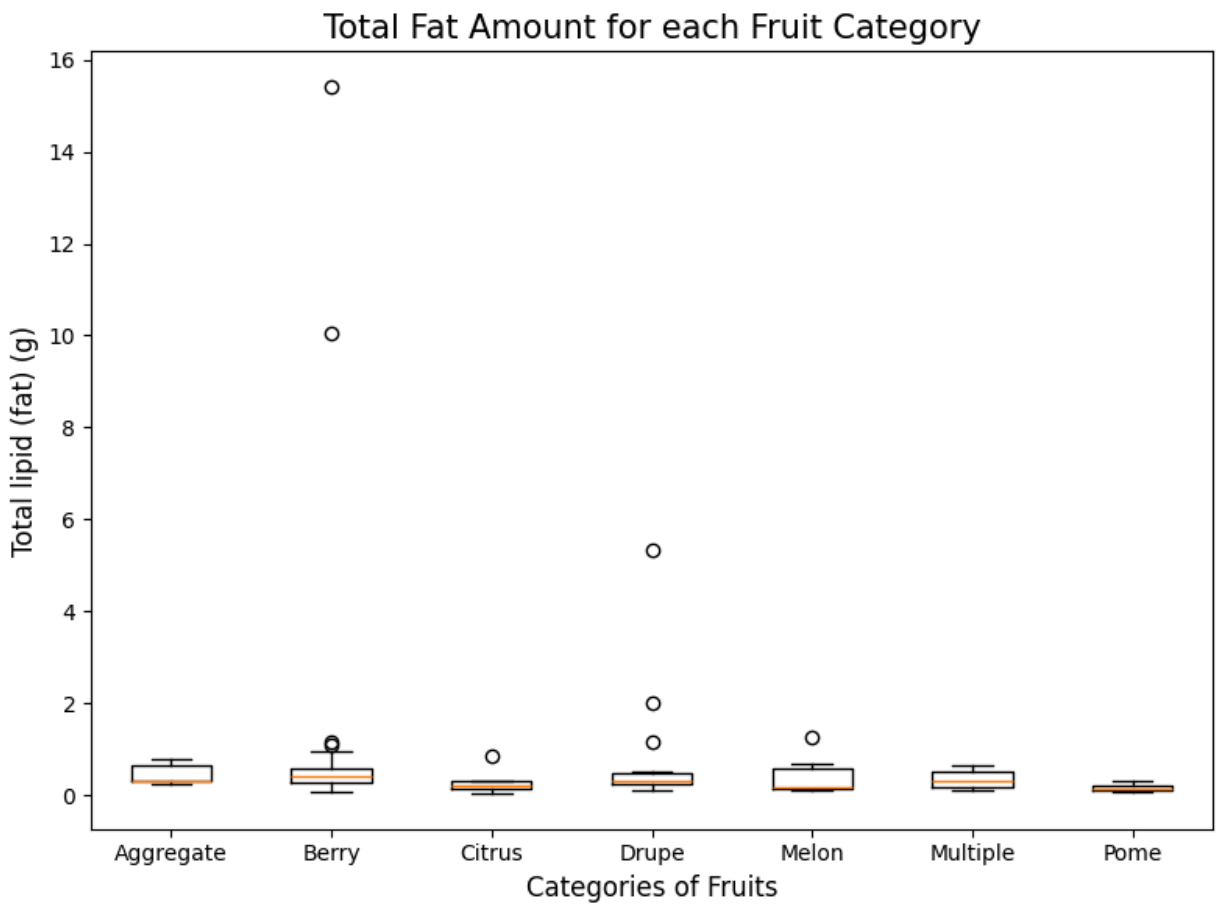
5 Appendix

Appendix A: All fruits in their respective categories. Berries have the highest number of fruits at 27 while melons have the lowest at 6.

Aggregate	Berry	Citrus	Drupe	Melon	Multiple	Pome
- Blackberry - Boysenberry - Cherimoya - Cloudberry - Loganberry - Raspberry - Soursop - Strawberry - Sweetsop	- Avocado (Cali.) - Avocado (Fl.) - Banana - Blueberry - Cranberry - Currant (black) - Currant (red & white) - Elderberry - Feijoa - Gooseberry - Grapes (red or green) - Groundcherry - Guava (common) - Guava (strawberry) - Kiwifruit (green) - Kiwifruit (ZESPRI) - Oheloberry - Papaya - Persimmon (Japanese) - Persimmon (native) - Plantain (green) - Plantain (yellow) - Pomegranate - Rose-apple - Sapodilla - Sapote - Pitanga	- Clementine - Grapefruit (white) - Grapefruit (pink and red) - Kumquat - Lemon - Lime - Orange (California) - Orange (Florida) - Orange (navels) - Tangerine - Pummelo	- Acerola - Apricot - Cherry - Durian - Java-plum - Jujube - Longan - Lychee - Mango - Nance - Nectarine - Opuntia - Peach - Plum - Rowal	- Horned melon - Cantaloupe - Casaba - Honeydew - Watermelon - Passionfruit (Granadilla)	- Abiyuch - Breadfruit - Fig - Jackfruit - Mulberry - Pineapple - Roselle	- Apple (fuji) - Apple (gala) - Apple (red delicious) - Crabapple - Loquat - Pear - Pear (red anjou) - Pear (bartlett) - Pear (bosc) - Pear (green anjou) - Pear (asian) - Quince



Appendix B: Bar graphs showing the average amount of various nutrients for each fruit category. These calculations of averages exclude NaN values.



Appendix C: Box plot showing the total fat for each fruit category.