Linah Rusere

Rating immunity-boosting capabilities of different types of foods

Problem Statement

The immune system plays a key role in defending the body against disease causing microorganisms. A weak immune system often makes an individual vulnerable to infection by bacteria, virus, fungi etc. When the body if too weak to fight infections this can lead to serious health complications or death. Vaccinations and antimicrobials against specific pathogens are proven ways of protecting the body against harmful diseases, however, their effectiveness also depends on the strength of the body's main natural defense system i.e. the immune system. Is it possible to boost the immune system? Medical professionals speculate making certain lifestyle changes may help to produce a healthy immune system. Can you boost the immune system by improving your diet? Which vitamins or micronutrients are most important? If dietary supplements are unavailable or cannot be taken, which foods are best at promoting a healthy immune system?

The goal of this project is to identify micronutrients that promote a stronger immune system and use this information to classify food products as either immune-boosting or not by analyzing the proportions of immune-boosting micronutrients in the food product. This project is divided into two parts. The first part of the project involves modeling immune-boosting properties of different micro-nutrients by analyzing composition of white blood cells and blood nutrient levels of the surveyed individuals in the NAHNES dataset provided by the CDC. After identifying the immune boosting micronutrients, the USDA nutritional dataset consisting of various foods and their nutritional content will be used to model the immune boosting capabilities of different foods.

Data Wrangling

The National Health and Nutrition Examination Survey NHANES data obtained from the CDC website was used to identify the most important micronutrients. Nutrition_US dataset used to identify immune boosting foods was downloaded from Kaggle as a CSV fie. This dataset was assembled from USDA website.

NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES continuous dataset consists of demographic data, dietary data, medical and laboratory tests, health-related questionnaires collected between 1999 and 2020 in two-year cycles. The data is provided in multiple tables stored as SAS .xpt files. I manually downloaded the files containing relevant data in 10 different cycles from 1999 to 2018. The survey participants were given unique identifiers "SEQN". The SEQN numbers continues from cycle to cycle. All the data was contained in more than a hundred separate files. I excluded the data taken during the pandemic which was cataloged separately. Each cycle has its own set of data tables. The majority of the data tables that were downloaded contained irrelevant columns that had to be filtered out. Moreover, there were gaps in the data where some data was collected in some cycles but not others, and missing values in some columns. In some cases similar data was labeled differently from year to year.

The relevant columns were extracted from each table using similar code series. In the cases where the same data column was represented with different names in different cycles, the columns were renamed to the same uniform name. The tables were initially concatenated by row, combining all the data from all the cycles. Then the columns were merged using an outer join to form the final data frame. The final data set consisted of 101,316 rows and 52 columns.

Exploratory Data Analysis

Selecting Relevant Selection

The immune system consists of various organs, cell and proteins that work together to protect the body against infection. Because the immune system is complex and not completely understood, it is difficult to accurately quantify. Physicians normally use the white blood cell count and antibodies as a measure to determine the strength of the immune system. A white blood cell (WBC) differential gives the count and/or percentage of the major white blood cell types A low white blood cell count may indicate an white blood cell production issues. It is important to note that the normal ranges vary from lab to lab. There are universally applied cutoff since optimal proportions of WBC in a healthy immune system has not been scientifically determined.
Total white blood cells 4.5 to 12 ug/L
neutrophils: 1.5 to 8 ug/L 40% to 60%
lymphocytes: 1.4 to 5.7 ug/L 20% to 40%
monocytes: 0.3 to 1 ug/L 2% to 8%
eosinophils: 0 to 1 ug/L 1% to 4%
Basophils: 0 to 1 ug/L 0.5% to 1%

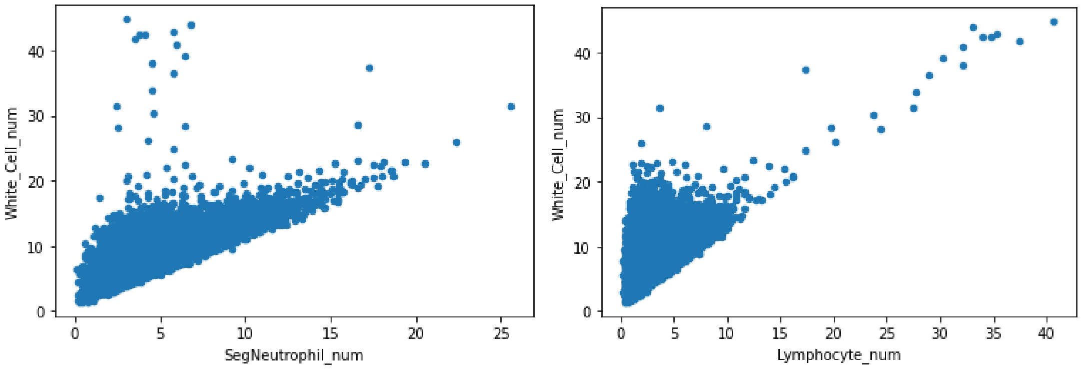

Figure 1. Relationship between total white blood cell count and neutrophils and lymphocytes.

The most important of these are the Neutrophils which act as the first line of defense during an infection and therefore must always be present in the normal quantities; and Lymphocytes(Tcells and Bcells) which activate the rest of the immune system and induce more production of immune cells and antibodies. Low levels of the other immune cells do necessarily indicate the state of the immune system since these cells are normally produced during an infection. The strong correlation between the total WBC count and the lymphocytes and neutrophils further suggests the high weighted importance of these cell types.

The distributions of the white blood cell numbers and the percentage proportions relative to their cutoff values were significantly different. Thus, I used the combined cutoffs for total white blood cell count, neutrophils and lymphocytes to label the survey participant's immune system strength as “High” or “Low”. This gave 62% “Low” and 38% “High” value counts. I did not impose the upper limit since high levels of WBC that exceed the upper limit can be a result of an acute infection which does not necessarily indicate the state of the immune system.

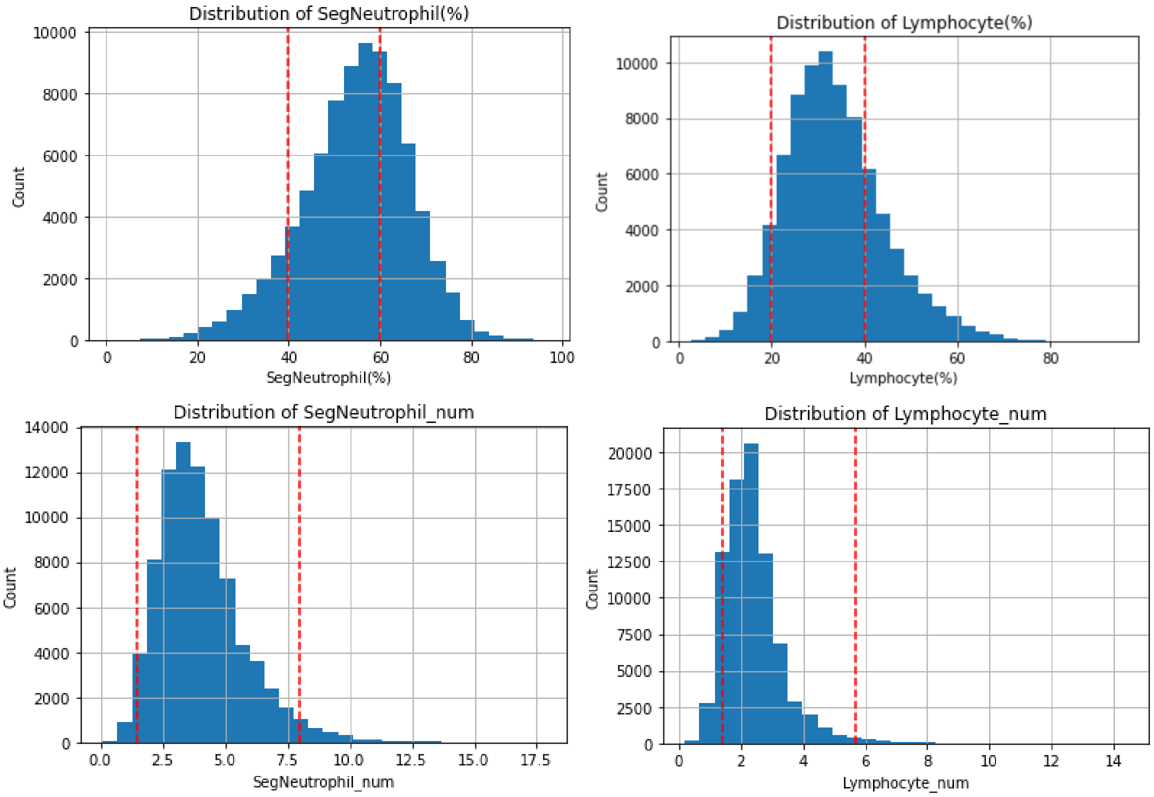

Figure 2. Varying distribution of neutrophils and lymphocytes

The features used in this study were carefully selected based on previous studies and literature reviews that have suggested there may be direct or indirect impact of these features on the health and immunity of individuals. Certain lifestyle changes have been suggested to boost the immune system, however, there are no scientifically proven direct links between lifestyle and enhanced immune function. Various aspects that are thought to affect the immune system include diet, weight, age, race, exercise, smoking, alcohol, sleep, stress, vaccines and diseases of the immune system. Some vitamins and minerals e.g vitamins A B C D E and folic acid and iron are believed play a crucial role in WBC formation. Low levels of selenium, copper, and zinc are suspected to play a role in WBC production although not yet proven scientifically. A blood test can identify whether these nutrients are low. The extent to which there nutrients contribute to increased levels of WBC has not been scientifically determined. Pregnancy and Smoking can cause lower levels of WBC. Alcoholism can also result in these chronic malnutrition and can therefore also be a cause of low WBC. Weight and BMI can lead to increased or decreased WBC count. Viral infections that last for several months (or indefinitely) and diseases that affect the immune system such HIV can causes low white blood cell count. Cancers that impact the bone marrow can cause low WBC counts, as most white blood cells are produced in the bone marrow. Most cancer treatments also cause a drop in WBC counts. WBC counts would be extremely low in these situations, not just mildly below the normal range. Age has also been understood as a general determinant of immune system health. Generally elderly people tend to have weaker immune responses. Race can also have an impact of an individuals WBC count. For example people of African origin can have a strong immune system even when their base level of white blood cell count are lower than normal.

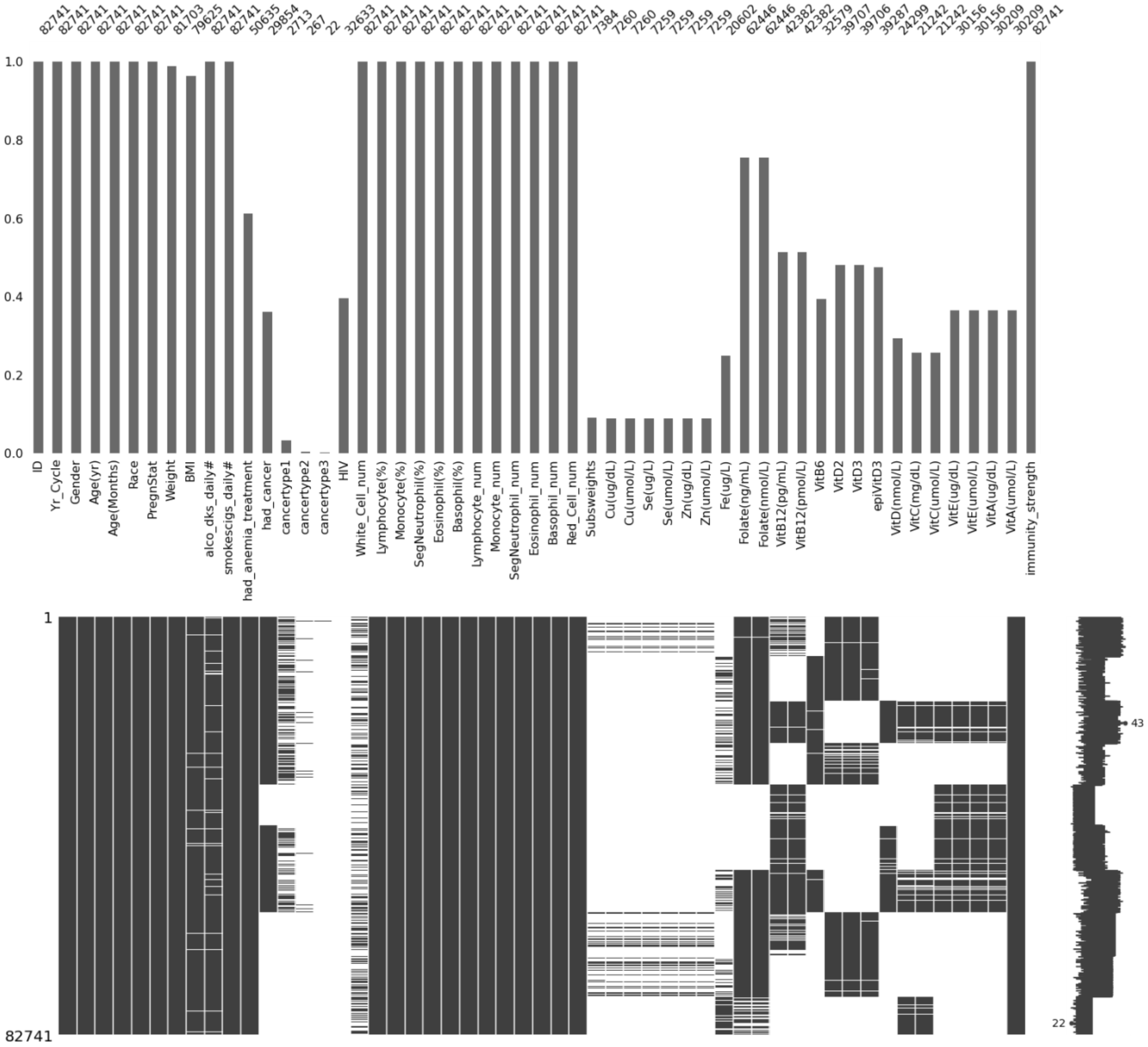

Figure 3. Visualizing missing data. The bar chart shows the relative proportions of missing values by column. The matrix plot shows a color fill for each column where data is present. The white space are the null values.

The data set consisted of a lot of missing values. Since white blood cell counts were used to label the data, I removed all entries missing lymphocyte numbers reducing the data set to 82741 rows. Missing values in the HIV, pregnancy status were assumed to be negative and the number of cigarettes smoked and alcohol drinks was assumed to be none and the data was imputed accordingly. The cancer type columns that had very few data points were removed. There were gaps in the data where some data was collected in some cycles but not others, and missing values in some columns. This was especially concerning since the micronutrients copper, zinc and selenium, I was mostly interested in did not have overlapping data with the vitamins that were also relevant features. Retaining all these columns was crucial for modelling the data and imputing the missing values using the generic methods such as the mean or median would not have been ideal representation of the data due to the large number of missing values.

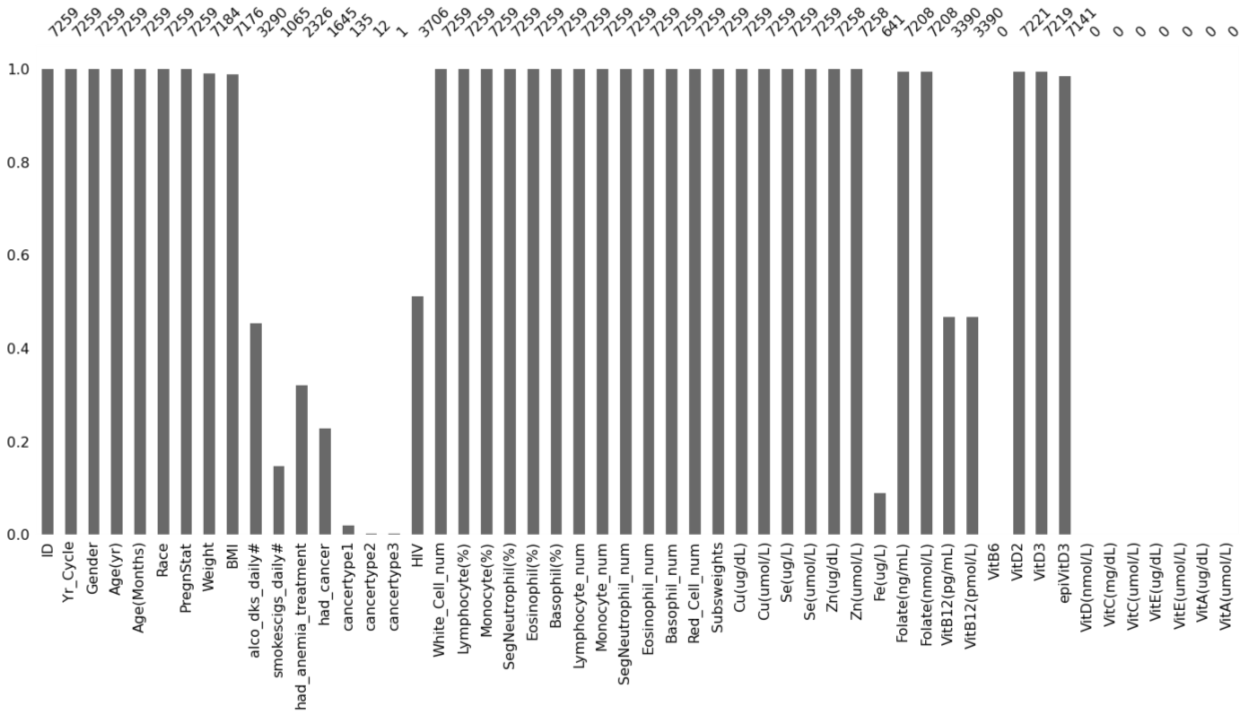

Figure 4. The subset of the data that contains the only copper, zinc and selenium data is missing corresponding data of 4 vitamins

The rest of the null values were imputed using Datawig SimpleImputer. DataWig uses a deep learning algorithm impute missing values of both categorical and numerical columns.

Model Selection

The final dataset with imputed values was used to compute the classification model and calculated feature importance was used to identify the micronutrients that were contributing more to the model. Data modeling was done with Decision Tree classifier and Random forest classifier algorithms with hyperparameter tuning using Gridsearch. Accuracy and precision metrics were used to compare the models Feature importance was computed using two different methods: scikit-learn's permutation based method and the built in Random forest feature importance method, Mean Decrease in Impurity (MDI).

The best model was a Random Forest classifier 94% accuracy and 95% precision.

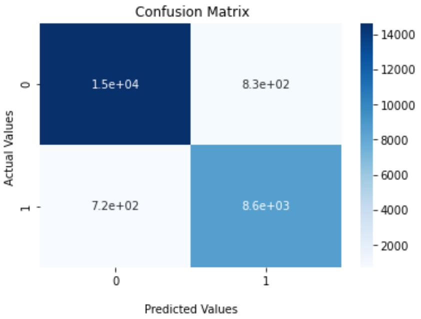

Figure 5. The confusion matrix of the best model

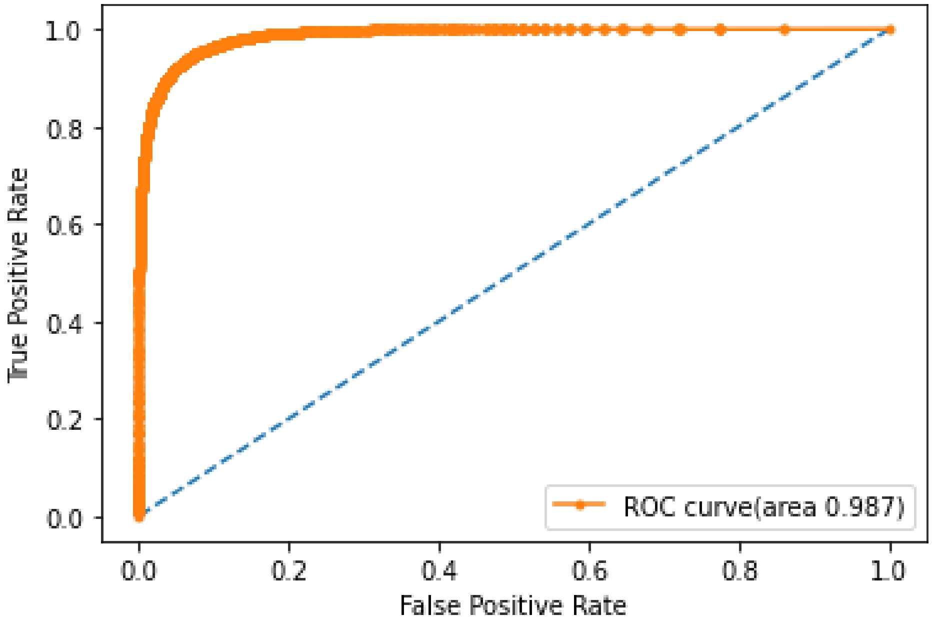

Figure 6. ROC curve for the best RF model

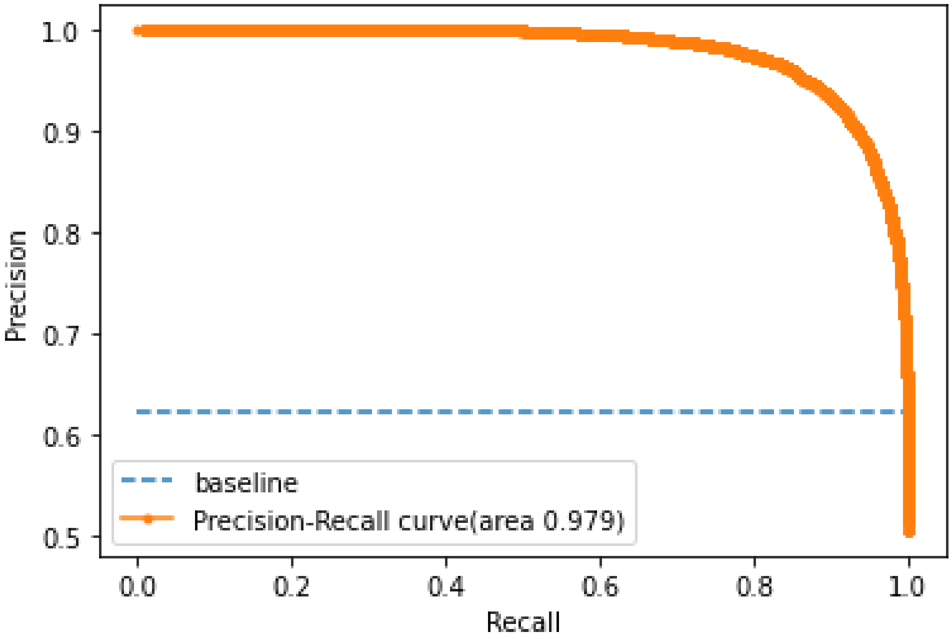

Figure 7. Precision-Recall curve for the best RF model

The computed feature importance with the best model, using permutation method, shows that Vitamin B12 and Zinc are the most important features followed by Selenium and Vitamin B6. The model also shows that there are several nutrients that contribute to the immune system's strength although with varying importance. This means these nutrients and other factors work synergistically to build and maintain a healthy immune system.

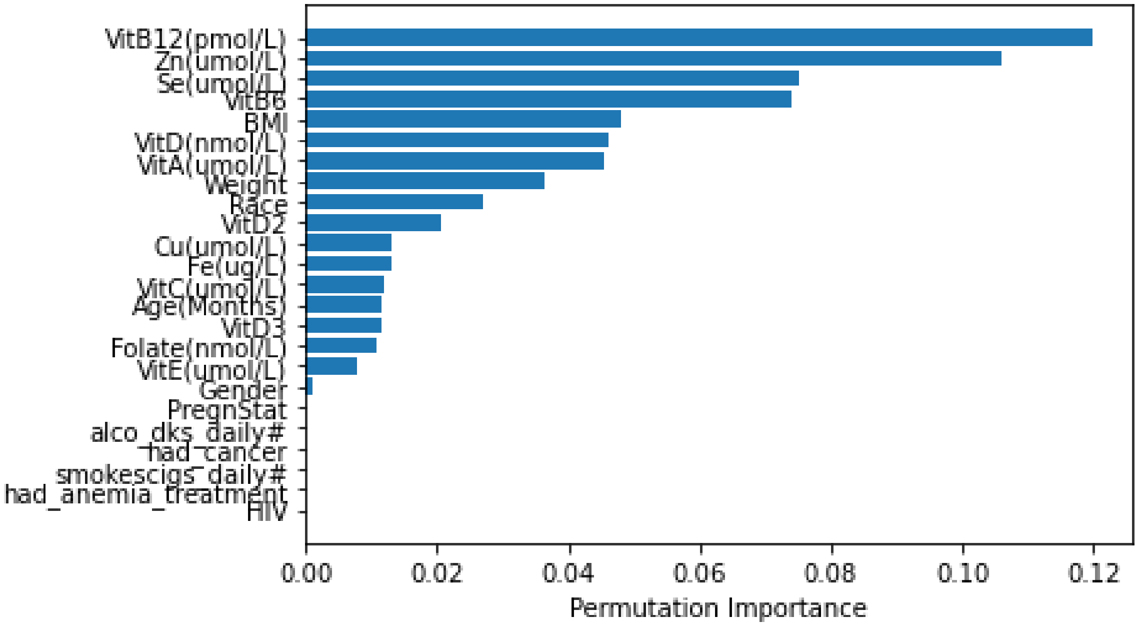

Figure 8. Permutation feature importance of the best Random Forest model

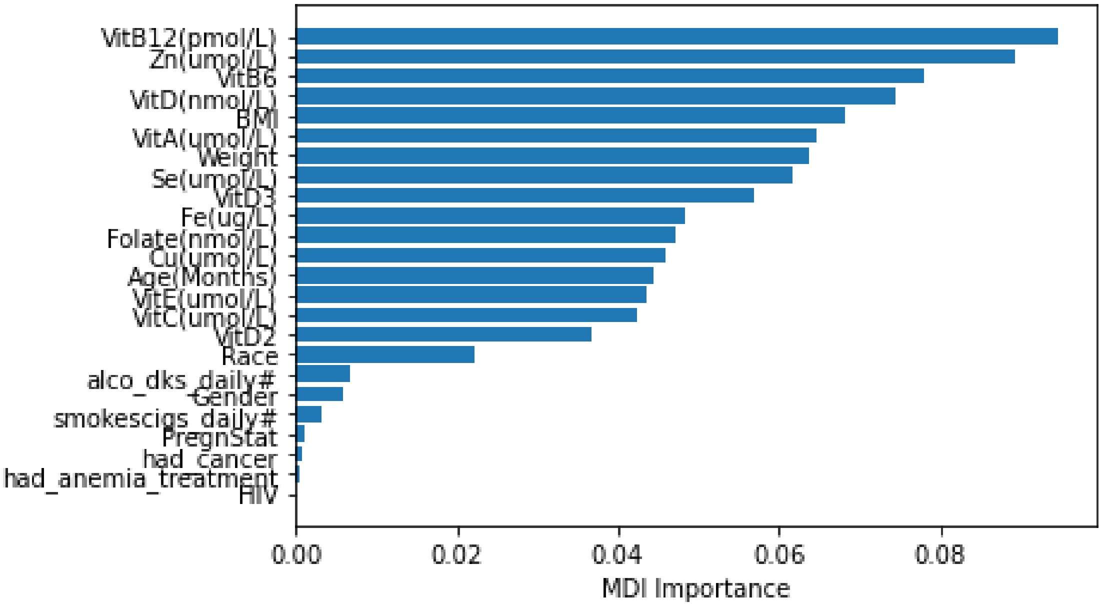

Figure 9. Random Forest built-in Mean Decrease in Impurity feature importance of the best Random Forest model

The feature importance computed by MDI method shows the same three features Vitamin B12, Zinc and Vitamin B6 in the top 4 list as observed with the permutation method. Although the there is some disparity in the relative importance, the general trend appears to be similar, with categorical variables at the bottom of the list. It is however important to highlight that impurity-based feature importance are prone to bias as they can be biased toward high cardinality features.

Interestingly, the decision tree model with 89% accuracy also showed the similar trend in feature importance with Vitamin B12 and Zinc as the most important features.

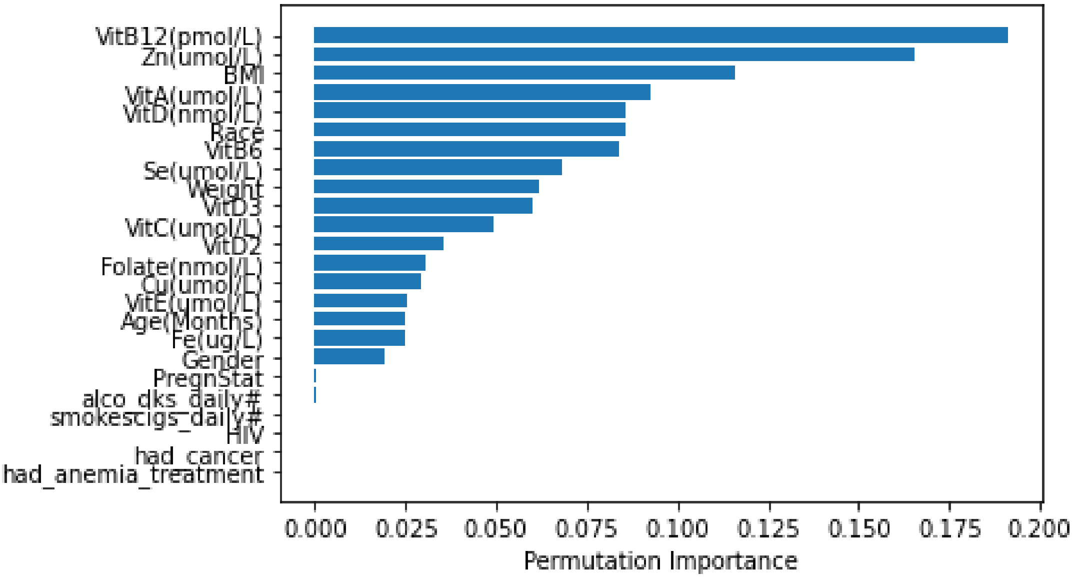

Figure 10. Permutation feature importance of the Decision Tree model

All the models and computed features importance consistently show that Vitamin B12 and Zn are the most important nutrients. Vitamin B6 is also a high ranking important nutrient as it was consistently in the top 4 features in all the models. Selenium appears to be an important feature although the computed feature importance using MDI method shows some discrepancy.
For the purposes of this project I used the ranking of the top 4 nutrients from the permutation method to score food items in the USDA Nutrition dataset.

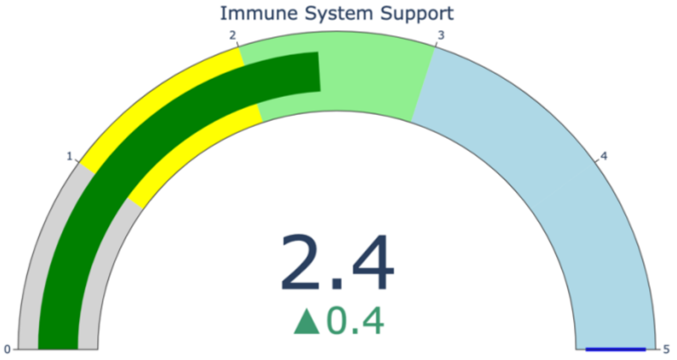

Figure 10. Score gauge showing the score of the immune-boosting score of a food

Conclusion

The Random Forest Classifier was the best model with high accuracy and precision. The model shows that most of the categorical features except for race were not important features for this model. All the micronutrients examined contribute to the immune system’s strength although with varying importance. This synergistic effect implies that optimizing one factor would not necessarily result in a healthy immune system. Thus, I created a scoring function based on the weighted contributions of the 4 most essential nutrients: Vitamin B12, Zinc, Selenium and vitamin B6 which allows the user to input the USDA code for the food of interest and visualize the score on a gauge chart .

Future work

The metrics of final model were good, however, there were few weaknesses in the model. Firstly, the dataset had significant amount of missing data that had to be imputed. Datawig worked well to impute the data, however, there are several methods to deal with the missing data that were not explored. It would be important to test the model on carefully selected data or imputed data to validate this model.

It would have been beneficial to explore other models that are not tree based. Tree based methods are good at handling categorical features compared to other classification models. Analysis of the feature importance shows that there are several features that could be dropped. In the future I would like to explore other classification models that normally perform poorly with categorical data.

The food scoring function could be improved by running further experiment to determine the actual proportional contributions of the nutrients to the immune system function.