IBS Symptoms Simulation Study

COIS 4470H: Research Project

By: Punyaja Mishra

April 16, 2023

Table of Contents

Introduction

This report is a comparative simulation study in identifying the relationship and factor of affect between the Irritable Bowel Syndrome Symptoms and individual IBS patients. The Simulation is performed in programming language, Python, using the provided datasets which were collected based on patients' inputs. The dataset used and the objective for the study was presented by the Injoy company during the Hackathon PharmaHacks2023 as 'Injoy Challenge'. The details have been fetched from the challenge proposal document which has been attached with the project report. Injoy works on research and treating patients gut health issues and the dataset provided were collected from their patients diagnosed with IBS based on the symptoms they experienced. The anonymity of the patients was maintained in the dataset by replacing them into auto incremental numeric values as "user_number".

Purpose of Study

“Irritable Bowel Syndrome (IBS) is a common functional gastrointestinal disorder that affects a significant portion of the population. The primary symptoms of IBS include abdominal pain, bloating, constipation, and diarrhea. The precise cause of IBS is not yet fully understood, but it is believed to be related to a combination of genetic, environmental, and lifestyle factors.

One of the most significant contributors to IBS symptoms is food. For people with IBS, certain foods can trigger symptoms such as abdominal pain, bloating, and diarrhea. Some common trigger foods include dairy products, FODMAPs, gluten, alcohol, and artificial sweeteners. Conversely, some foods, such as fiber-rich fruits, vegetables, and whole grains, may help alleviate IBS symptoms.

The impact of food on IBS symptoms varies greatly from person to person, making it challenging to determine the most effective diet for managing IBS. As a result, many people with IBS rely on a trial-and-error approach to identify their trigger foods and develop a personalized diet to manage their symptoms.

Therefore, determining the effect of food on a person’s gastrointestinal symptoms is essential for people with IBS to effectively manage their condition and improve their quality of life.” (Injoy Challenge, PharmaHacks 2023, p. 02)

Objective

This Simulation study aims to identify the link between food and if or not they trigger any gastrointestinal IBS symptoms for each individual. To achieve this we need to accurately predict the symptom score and be able to explain the contribution of each food to the final prediction for each individual user. By doing this we try to narrow down the food types that should be avoided specific to the individual. Although each user records their diet and symptoms, the dataset is limited with only a few data points per user.

Simulation Language

Simulation Language used for this study is the Programming Language **Python**. Python provides multiple Machine Learning Algorithm libraries that serves that purpose of simulating the dataset to produce meaningful results. In addition to data analysis algorithms, Python also provides with multiple data visualization libraries, such as the most commonly used library – matplotlib that helps put the results into a visualized concept.

Understanding Provided Dataset

Datasets

For this study, the dataset provided is "*data.csv*". This data frame contains the data points for 50 users. The number of rows corresponding to each individual vary and have no correlation among each other whatsoever. There are 20 food categories labeled as F1, F2, F3 and so on till F20. The columns include the 20 Food Categories, Symptom value and User Number. User Number Column defines which user the particular row's data belongs to. Each User recorded the numeric values for each specific food category defining the amount of intake/serving for them – the unit of measurement is constant for each food category and user to maintain continuity in the study. Symptom value was calculated for each datapoint based on how much symptoms the user experienced. This value is defined as a float value and symptom value equal to 0 means that user experienced no symptoms.

Assumptions

There were a few assumptions taken into consideration based on which the algorithm was chosen for the simulation.

- For the *data.csv* dataset, there was no information provided on how the different values of intake defined for each food category were quantified, as in the unit of their measurement. As the method of receiving these values wasn't mentioned, it was assumed that the symptomatic value above zero meant a symptom was seen and it showed negative effects over IBS, which means it triggers IBS for that particular user.
- Similarly, food categories were labeled as "F1", "F2" and so on, which means, we do not care exactly what food is affecting how. There is a correlation between certain food categories - the food data can be combined into broader categories like FODMAP, gluten, and wheat, which can lead to multicollinearity.
    - However, For food hierarchy, as there was no documentation provided with the food types, we thought on not using these, as it meant replacing the occurrence of parent food type with child ones. That approach was avoided as there were some entries which showed complete contrasting results for some of the cases.
    - From some cases it was concluded F5 to be the food type leading to food causing trouble in most IBS cases, from the food hierarchy we can see that f3,f19,f10,f4 don't show no significant symptomatic value or cases where they can seem to effecting symptomatic value.
    - From these observations it was concluded to not using food hierarchy avoids any misinterpretation of data and lead to inconclusive results.

Approach

The problem was solved in 2 approaches. Both of the approaches have been discussed below along with description of any obstacles that were faced.

Multilinear Regression Model

In order to establish relationships between the columns of a given data frame, a multilinear regression model was utilized, with a 20-80 ratio employed for training and testing the data. Prior to analysis, the data was examined to account for instances where symptom values were recorded as zero, indicating a lack of health issues for the selected individual.

Subsequently, the correlation between various food types and the severity of Irritable Bowel Syndrome (IBS) symptoms was investigated to identify potential contributors to the development of IBS over time. The data was then segregated into X (food groups) and Y (symptom value) segments to train a multilinear regression model using *XGBoost*. To further understand the model's predictions, *SHAP* values were employed to determine how each food group contributed to the overall result, which was subsequently presented through a data table, bar chart and extremities chart.

This process was replicated for individual users and then extended to the collective group, to identify if any food category has a significant impact on the development of IBS, regardless of the individual. Such insights could aid in identifying a potential leading cause of IBS among the population.

Auto-sklearn Model

AutoML (Automated Machine Learning) is a growing field of research that aims to automate the machine learning pipeline, from data preprocessing to model selection and hyperparameter tuning. Auto-sklearn is one of the most popular open-source AutoML libraries that provides a simple and efficient interface to automate the machine learning pipeline.

Auto-sklearn is based on the Scikit-learn library and uses a meta-learning approach to select the best machine learning pipeline for a given dataset. It automatically searches for the best algorithm, preprocessing steps, and hyperparameters using Bayesian optimization and ensemble techniques.

One of the main advantages of Auto-sklearn is that it can save time and resources, especially for users who have limited knowledge or experience in machine learning. Auto-sklearn automates the time-consuming and complex tasks of selecting the best algorithm, tuning the hyperparameters, and optimizing the model's performance.

Additionally, Auto-sklearn provides a variety of performance metrics to evaluate the model's performance, including accuracy, precision, recall, F1-score, and others. It also provides detailed information about the selected model's architecture, hyperparameters, and feature importance, making it easier to interpret the model's results.

The conceptual model and specification model were developed to approach the problem and generate a model that could predict the correlation between food categories and IBS symptoms. Auto-sklearn Regression was chosen as the machine learning algorithm for its ability to automatically select and optimize available regression ML algorithms and hyperparameters, resulting in better performance and faster model creation. 20-80 ratio was implemented to train and test the data. Followed by the prediction of the R2 Score to check the efficiency of the algorithm.

Conceptual Model

Out of the datasets that were provided, data.csv is being used as not much context was provided about the data in the other dataset.

Input Data

- User Number whose data is being trained and relationship is being predicted
- Food Categories, labeled as 'F1' to 'F20', with numeric values for each specific food category. They are represented as the amount of serving consumed. For instance, F1 = 4 implied that 4 units of Food Category F1 were consumed for that data point
- Symptom Value is a measure of symptoms experienced by the user per data point. A symptom value of '0' indicated that no symptoms were observed

Output Data

- The symptom value predicted after training the model and calculating the accuracy of our models for both approaches separately

Specification Model

Two Programs were developed to tackle this problem - foodUser.py and foodCollective.py.

Total Number of Users = 50
User_Number = User Id whose correlation values are calculated in foodUser.py. This is an input to the program and belongs in range (0,50) as in the dataset
Number of Data Points per User = Varied per User (No certain Pattern as it was Patient Input Data)

Feature Selection = Spearman Correlation

foodUser.py

Input: User Number between 0 and 49 and generates a model that calculates the correlation between each food category and the symptom value for that individual user

Machine Learning Algorithm Used 1 = XGBoost

foodUser_auto.py

Input: User Number between 0 and 49 and generates a model that calculates the correlation between each food category and the symptom value for that individual user

Machine Learning Algorithm Used 2 = Auto Scikit Learn Regression

foodCollective.py

Input: All the correlation values generated by foodUser.py for all users and aggregates them to identify the food category with the highest and lowest correlation.

Machine Learning Algorithm Used = XGBoost

Spearman Correlation

- We implemented the Spearman Correlation method for feature selection to identify the Food Categories that has the highest effect on the Symptom Value Per User.
- Spearman correlation is a statistical measure of the strength and direction of the monotonic relationship between two variables. Unlike Pearson correlation, which measures the linear relationship between two variables, Spearman correlation is based on the ranked values of the variables, making it suitable for non-parametric data.
- In regression models, Spearman correlation can be used to identify and quantify the strength and direction of the relationship between the independent and dependent variables. This information can be used to select the appropriate regression model and to evaluate the goodness of fit of the model.
- Additionally, Spearman correlation can be used to detect outliers and influential observations that can affect the regression model's performance. By identifying and removing these observations, the regression model's accuracy can be improved. Overall, Spearman correlation is a useful tool for regression modeling, particularly in cases where the relationship between variables is non-linear or non-normal.
- Since our model needed to predict the relationship between each food and how it affects the symptom value, spearman correlation was used
- **Output** : list of food categories

XGBoost

- To find correlation values, we needed a regression model. XGBoost was our choice of ML Algorithm over other regression models because of its ability to consider multiple independent parameters.
- XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm used for supervised learning tasks, such as classification and regression. It is a type of ensemble learning method that combines multiple decision trees to create a more accurate and robust model.
- XGBoost is particularly useful because it is designed to handle large datasets and can train models quickly. It has a built-in feature that allows it to parallelize the training process, making it faster than other algorithms that use a single processing unit.

Autosklearn

Auto-sklearn is a powerful automated machine learning tool that can be used for a variety of tasks, including regression. Auto-sklearn regression is based on the Scikit-learn library and uses a meta-learning approach to select the best regression model and hyperparameters for a given dataset.

The process of using Auto-sklearn regression is relatively simple. The user provides the input dataset, selects the performance metric to optimize, and specifies the time and resource constraints for the search process. Then, Auto-sklearn runs a search algorithm that tries out different regression models, preprocessing steps, and hyperparameters combinations to find the best model that fits the given dataset.

Auto-sklearn regression supports various types of regression models, including linear regression, decision trees, random forests, support vector regression, and gradient boosting regression. It also supports different types of preprocessing steps, such as feature scaling, feature selection, and feature engineering, among others.

One of the advantages of using Auto-sklearn for regression is that it saves time and resources by automating the process of selecting the best model and hyperparameters. It can also handle large datasets and can be used to solve complex regression problems that would otherwise require a significant amount of time and expertise to solve.

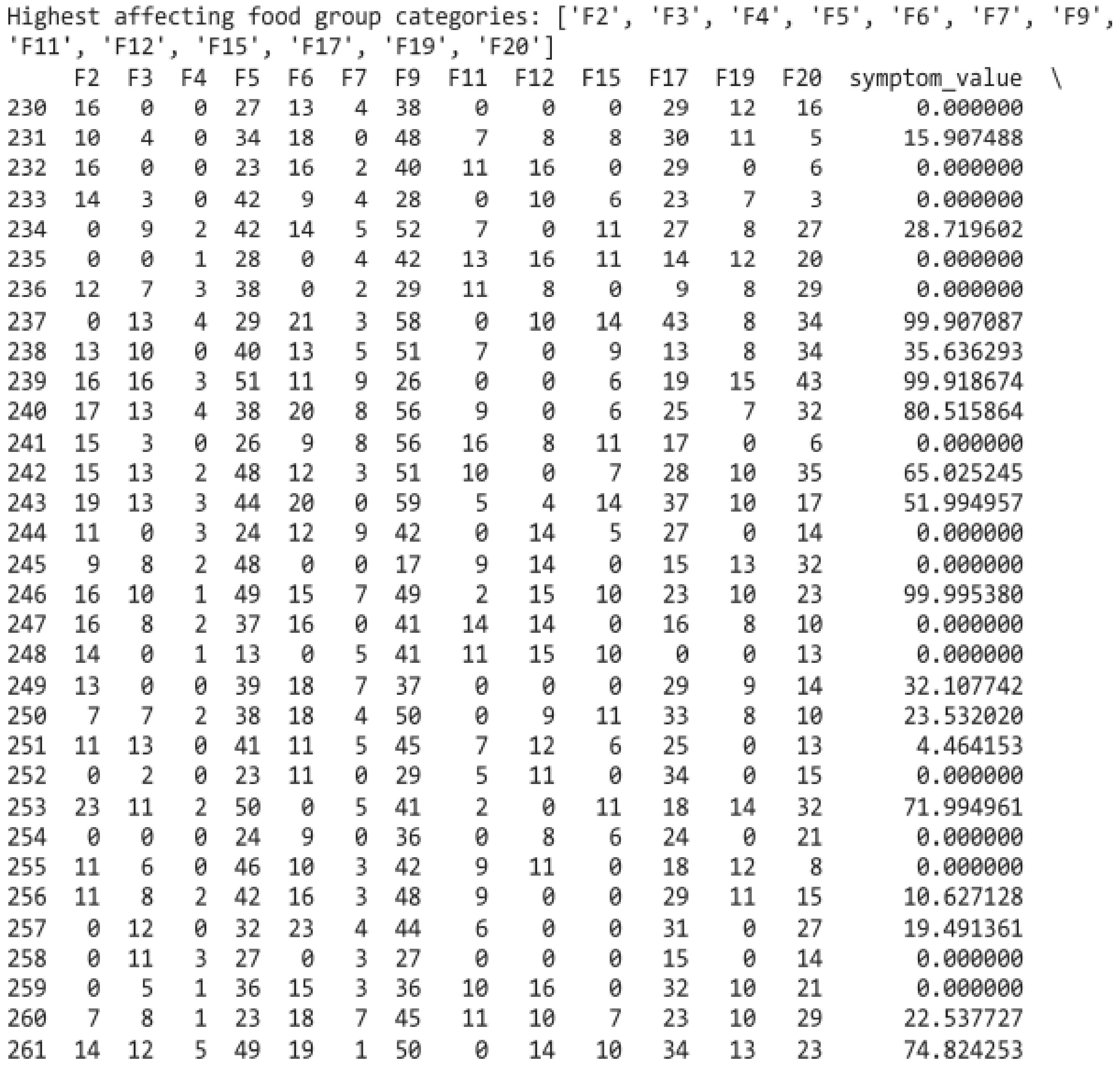
Simulation Results - foodUser.py

The first Model developed was using XGBoost, a regression Model, to perform on individual user data to identify the food categories that were affecting the IBS symptoms of that individual user and by how much, i.e., their correlation value per food category.

Simulation Run 1 – User 09

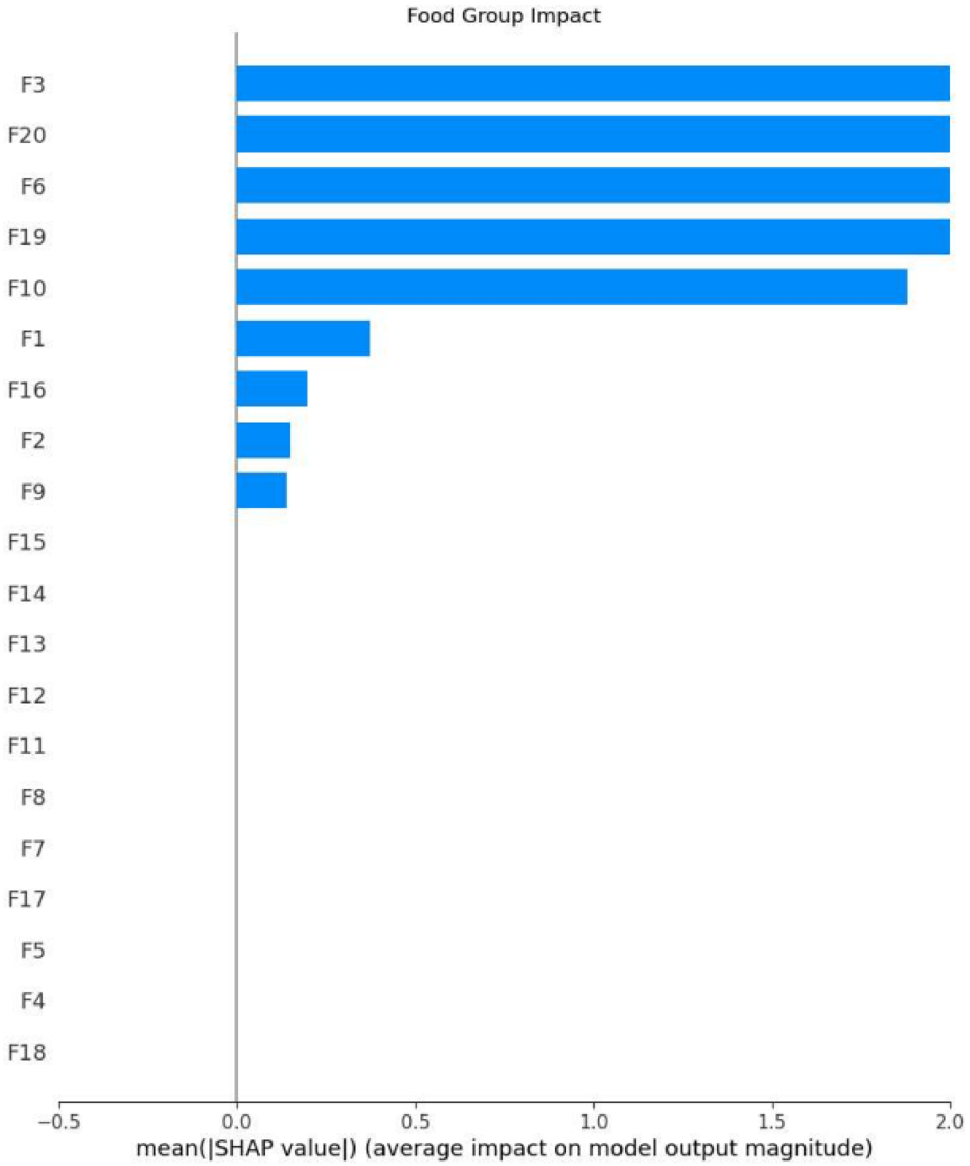
Highest Affecting Food Categories for User 09:

These values represent the food categories that have the highest affect on the individual user based on its symptom values for all data points. This is calculated using spearman correlation.



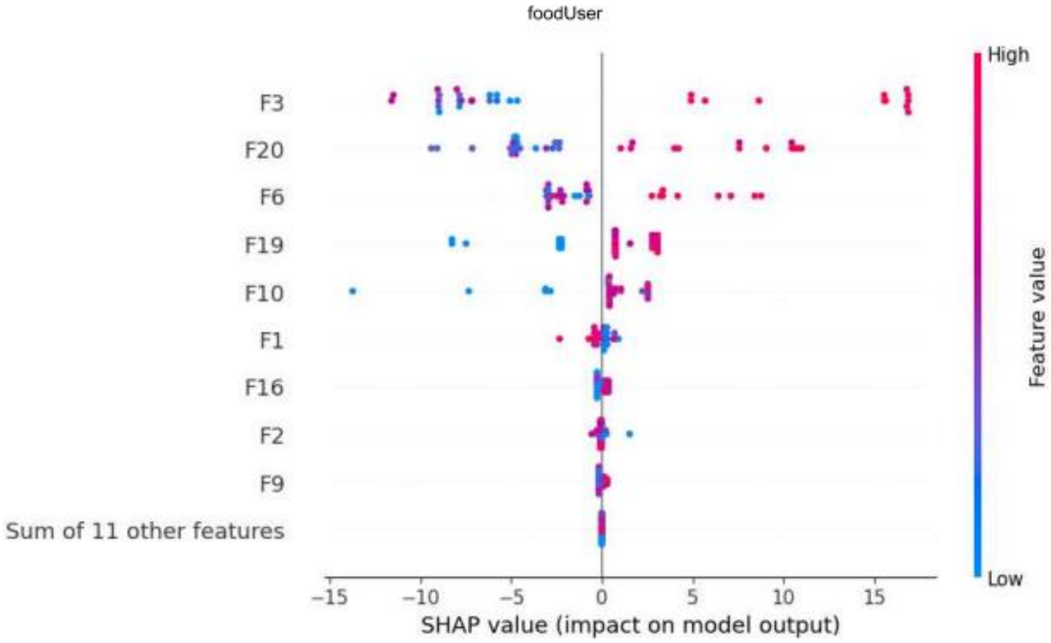
```
Highest affecting food group categories: ['F2', 'F3', 'F4', 'F5', 'F6', 'F7', 'F9',
'F11', 'F12', 'F15', 'F17', 'F19', 'F20']
     F2  F3  F4  F5  F6  F7  F9  F11  F12  F15  F17  F19  F20  symptom_value  \
230  16   0   0  27  13   4  38    0    0    0   29   12   16       0.000000
231  10   4   0  34  18   0  48    7    8    8   30   11    5      15.907488
232  16   0   0  23  16   2  40   11   16    0   29    0    6       0.000000
233  14   3   0  42   9   4  28    0   10    6   23    7    3       0.000000
234   0   9   2  42  14   5  52    7    0   11   27    8   27      28.719602
235   0   0   1  28   0   4  42   13   16   11   14   12   20       0.000000
236  12   7   3  38   0   2  29   11    8    0    9    8   29       0.000000
237   0  13   4  29  21   3  58    0   10   14   43    8   34      99.907087
238  13  10   0  40  13   5  51    7    0    9   13    8   34      35.636293
239  16  16   3  51  11   9  26    0    0    6   19   15   43      99.918674
240  17  13   4  38  20   8  56    9    0    6   25    7   32      80.515864
241  15   3   0  26   9   8  56   16    8   11   17    0    6       0.000000
242  15  13   2  48  12   3  51   10    0    7   28   10   35      65.025245
243  19  13   3  44  20   0  59    5    4   14   37   10   17      51.994957
244  11   0   3  24  12   9  42    0   14    5   27    0   14       0.000000
245   9   8   2  48   0   0  17    9   14    0   15   13   32       0.000000
246  16  10   1  49  15   7  49    2   15   10   23   10   23      99.995380
247  16   8   2  37  16   0  41   14   14    0   16    8   10       0.000000
248  14   0   1  13   0   5  41   11   15   10    0    0   13       0.000000
249  13   0   0  39  18   7  37    0    0    0   29    9   14      32.107742
250   7   7   2  38  18   4  50    0    9   11   33    8   10      23.532020
251  11  13   0  41  11   5  45    7   12    6   25    0   13       4.464153
252   0   2   0  23  11   0  29    5   11    0   34    0   15       0.000000
253  23  11   2  50   0   5  41    2    0   11   18   14   32      71.994961
254   0   0   0  24   9   0  36    0    8    6   24    0   21       0.000000
255  11   6   0  46  10   3  42    9   11    0   18   12    8       0.000000
256  11   8   2  42  16   3  48    9    0    0   29   11   15      10.627128
257   0  12   0  32  23   4  44    6    0    0   31    0   27      19.491361
258   0  11   3  27   0   3  27    0    0    0   15    0   14       0.000000
259   0   5   1  36  15   3  36   10   16    0   32   10   21       0.000000
260   7   8   1  23  18   7  45   11   10    7   23   10   29      22.537727
261  14  12   5  49  19   1  50    0   14   10   34   13   23      74.824253
```

Average Impact of each Food Categories on User 09:



The bar chart represents the average impact on the model impact per food category for the chosen individual user, User 09. As we see, the high correlation categories that were printed form the spearman correlation categories have the highest and most or all affect on the IBS symptoms.

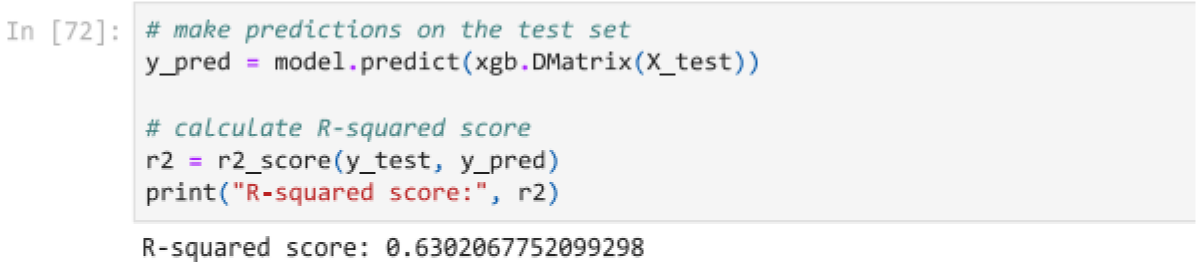
Extremities Chart for User 09:



The extremities chart represent the SHAP values of each food category for the user. These vary from negative to positive values representing that the combination of certain food categories can affect the affect on a user's symptom. Which proves our reason to believe that IBS has not been able to identify a root cause of its cause and therefore finding a generic solution or cure for IBS has been difficult.

Based on the analysis of the high correlation food categories table, bar chart, and extremities chart, we have identified the food categories that appear to be the most influential in causing IBS symptoms for User 09. The visual aids provided by these charts make it simple to understand which food groups have the highest correlations to symptom values and the severity level of a particular food group for a specific user. This approach can be used for any individual to determine which food groups may be provoking their IBS symptoms.

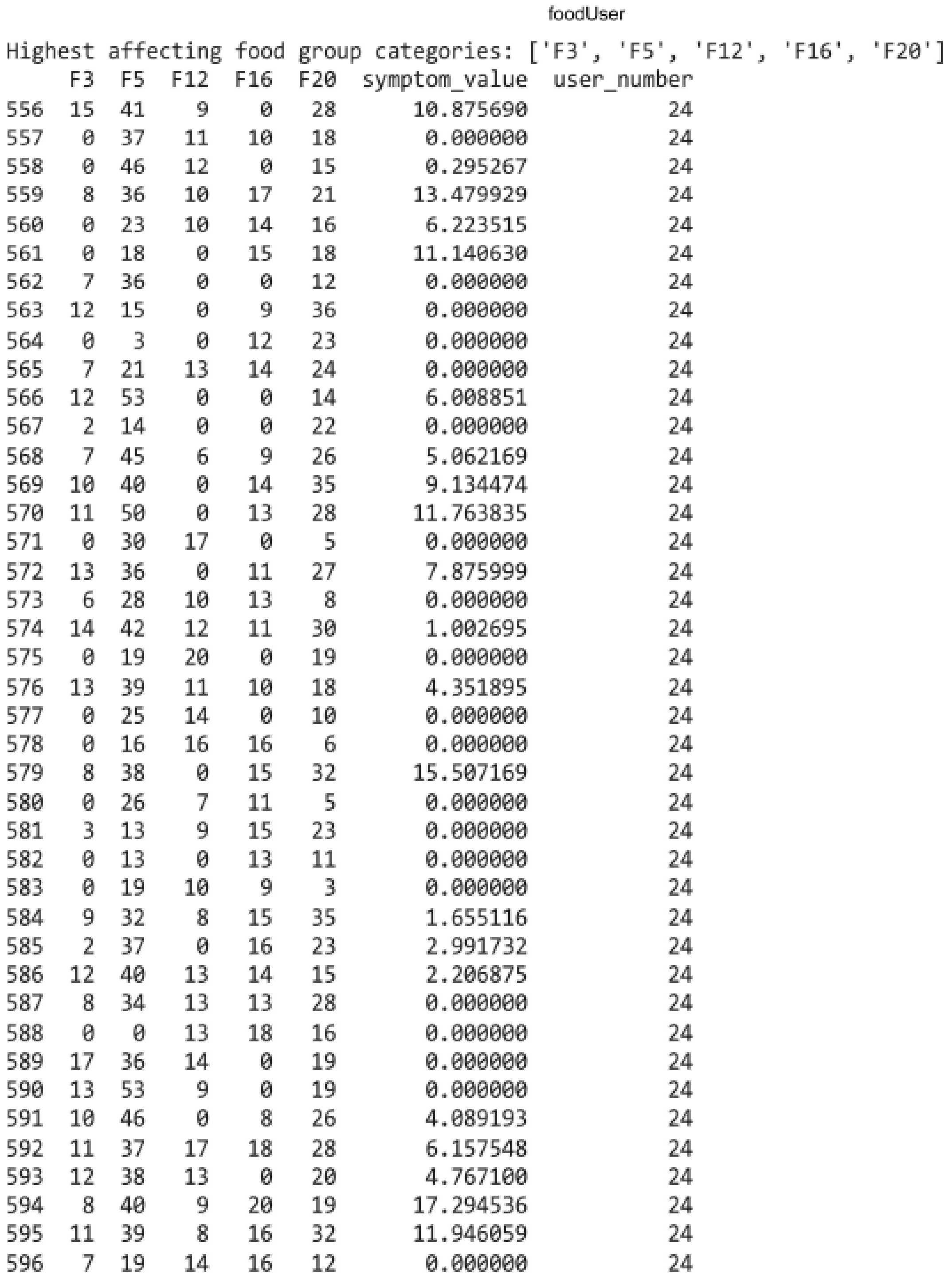
R2 Score is used to calculate the performance of a Regression Model. The predicted values for our XGBoost Regression Model for User 09 have a R2_score of 63%, which tells us that the model worked well for the user data.



```
In [72]: # make predictions on the test set
         y_pred = model.predict(xgb.DMatrix(X_test))

         # calculate R-squared score
         r2 = r2_score(y_test, y_pred)
         print("R-squared score:", r2)

         R-squared score: 0.6302067752099298
```

Simulation Run 2 – User 24

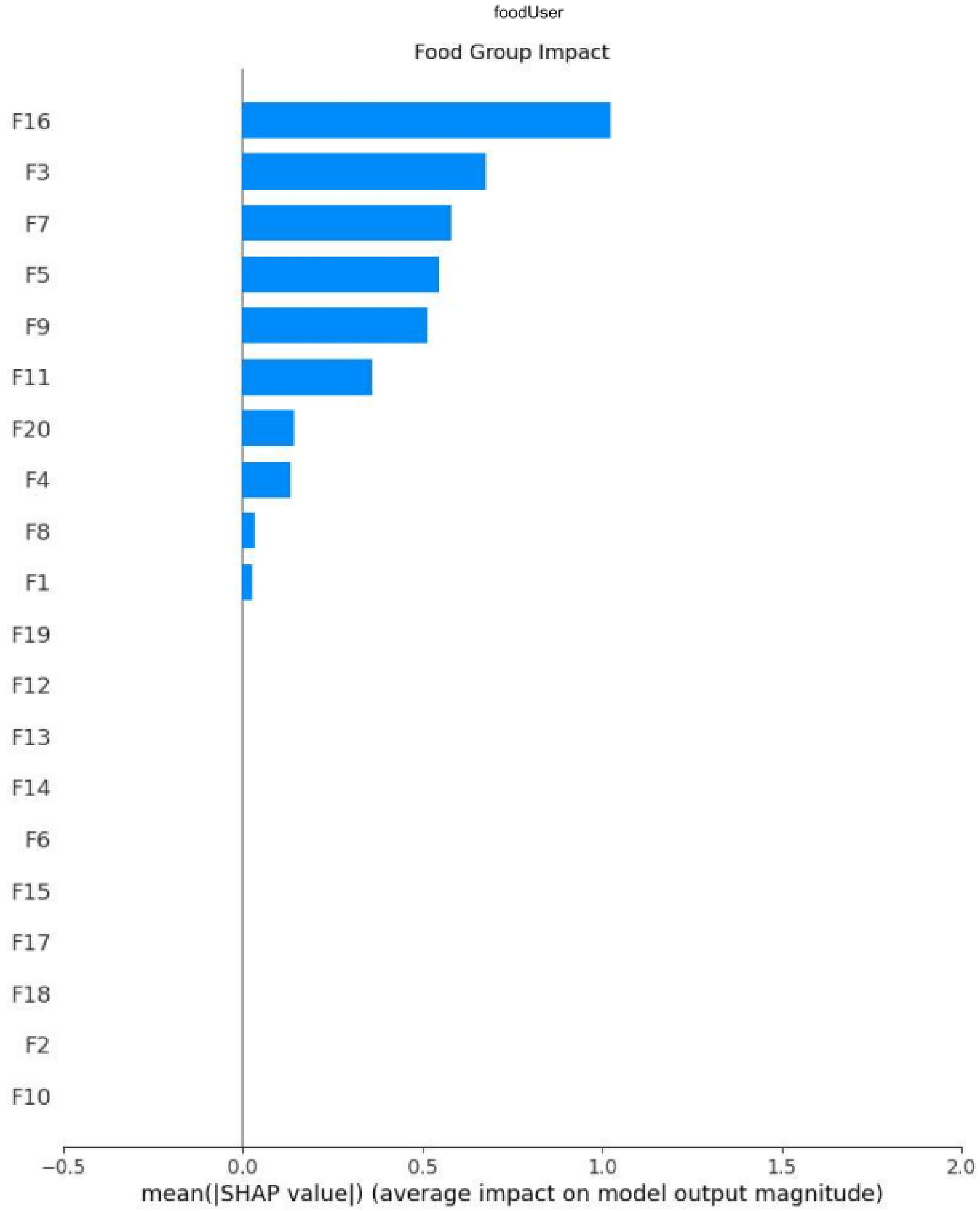
Highest Affecting Food Categories for User 24:



foodUser

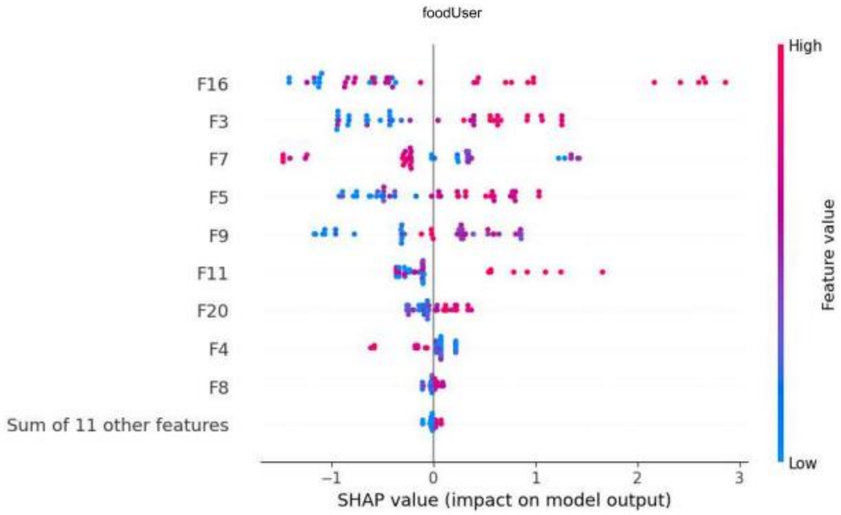
```
Highest affecting food group categories: ['F3', 'F5', 'F12', 'F16', 'F20']
     F3  F5  F12  F16  F20  symptom_value  user_number
556  15  41    9    0   28      10.875690           24
557   0  37   11   10   18       0.000000           24
558   0  46   12    0   15       0.295267           24
559   8  36   10   17   21      13.479929           24
560   0  23   10   14   16       6.223515           24
561   0  18    0   15   18      11.140630           24
562   7  36    0    0   12       0.000000           24
563  12  15    0    9   36       0.000000           24
564   0   3    0   12   23       0.000000           24
565   7  21   13   14   24       0.000000           24
566  12  53    0    0   14       6.008851           24
567   2  14    0    0   22       0.000000           24
568   7  45    6    9   26       5.062169           24
569  10  40    0   14   35       9.134474           24
570  11  50    0   13   28      11.763835           24
571   0  30   17    0    5       0.000000           24
572  13  36    0   11   27       7.875999           24
573   6  28   10   13    8       0.000000           24
574  14  42   12   11   30       1.002695           24
575   0  19   20    0   19       0.000000           24
576  13  39   11   10   18       4.351895           24
577   0  25   14    0   10       0.000000           24
578   0  16   16   16    6       0.000000           24
579   8  38    0   15   32      15.507169           24
580   0  26    7   11    5       0.000000           24
581   3  13    9   15   23       0.000000           24
582   0  13    0   13   11       0.000000           24
583   0  19   10    9    3       0.000000           24
584   9  32    8   15   35       1.655116           24
585   2  37    0   16   23       2.991732           24
586  12  40   13   14   15       2.206875           24
587   8  34   13   13   28       0.000000           24
588   0   0   13   18   16       0.000000           24
589  17  36   14    0   19       0.000000           24
590  13  53    9    0   19       0.000000           24
591  10  46    0    8   26       4.089193           24
592  11  37   17   18   28       6.157548           24
593  12  38   13    0   20       4.767100           24
594   8  40    9   20   19      17.294536           24
595  11  39    8   16   32      11.946059           24
596   7  19   14   16   12       0.000000           24
```

Average Impact of each Food Categories on User 24:



The bar chart represents the average impact on the model impact per food category for the chosen individual user, User 24. As we see, the high correlation categories that were printed form the spearman correlation categories have the highest and most or all affect on the IBS symptoms.

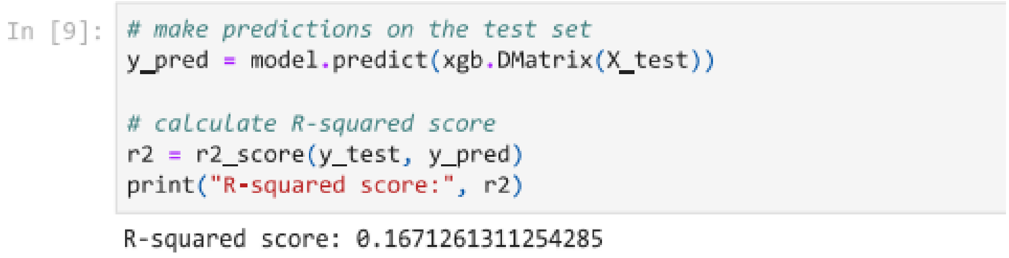
Extremities Chart for User 24:



The extremities chart represent the SHAP values of each food category for the user. These vary from negative to positive values representing that the combination of certain food categories can affect the affect on a user's symptom. Which proves our reason to believe that IBS has not been able to identify a root cause of its cause and therefore finding a generic solution or cure for IBS has been difficult.

Based on the analysis of the high correlation food categories table, bar chart, and extremities chart, we have identified the food categories that appear to be the most influential in causing IBS symptoms for User 09. The visual aids provided by these charts make it simple to understand which food groups have the highest correlations to symptom values and the severity level of a particular food group for a specific user. This approach can be used for any individual to determine which food groups may be provoking their IBS symptoms.

R2 Score is used to calculate the performance of a Regression Model. The predicted values for our XGBoost Regression Model for User 09 have a R2_score of 63%, which tells us that the model worked well for the user data.



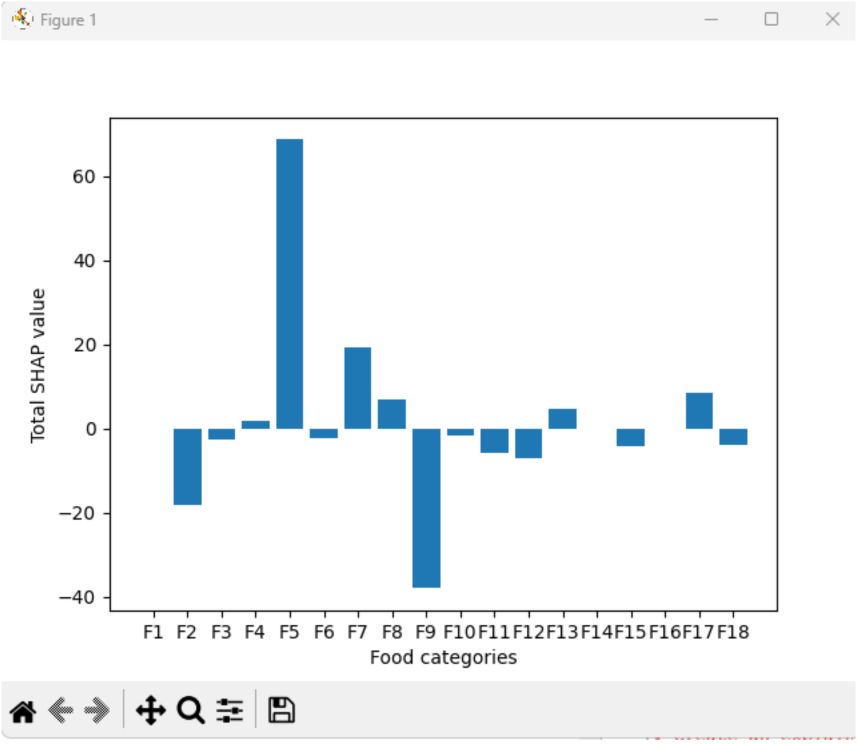
```
In [9]: # make predictions on the test set
        y_pred = model.predict(xgb.DMatrix(X_test))

        # calculate R-squared score
        r2 = r2_score(y_test, y_pred)
        print("R-squared score:", r2)

        R-squared score: 0.1671261311254285
```

While we got a 63% R2 Score for User 09, we had a dropping 16% for User 24. This led us to believe that there was a high inconsistency in the machine learning model. Even though XGBoost is a good choice of algorithm , there are several other regression models such as the Gaussian Process, KNN, decision trees that would perform well for the given data. Therefore, it was decided to use the Auto Scikit-Learn Regression Model and see how it behaves.

Simulation Results - foodCollective.py

The second Model developed was also using XGBoost, a regression Model – this time to perform on the data as a collective to identify the food categories that were affecting the IBS symptoms of users in general – basically identifying if there are any food categories, that affect all or most users with the highest probability.

Correlation Values pe food Category for all users combined:



The purpose of this simulation is to identify the food groups that have a high probability of triggering IBS symptoms among the majority of users. To achieve this, we utilized the entire dataset of user-entered data and created a model. Our findings indicate that food group F5 is more likely to cause IBS symptoms, whereas food group F9 has the lowest likelihood of triggering such symptoms across the dataset.

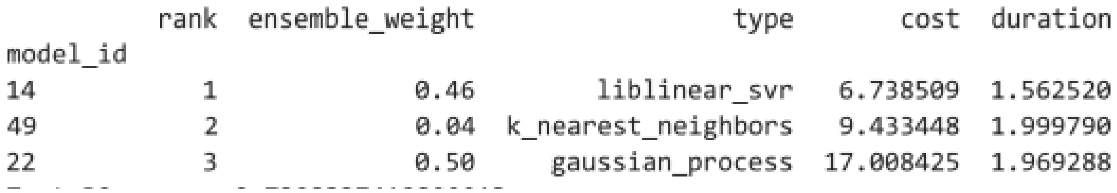
Furthermore, this analysis sheds light on the food hierarchy and the parent-child relationships between food groups. Interestingly, the assumption that a parent food causing IBS symptoms would lead to its child foods causing symptoms is proven false by the data. This underscores the fact that there may be deeper underlying factors beyond just the foods themselves that contribute to IBS.

It would be interesting to look more into this by implementing certain food hierarchy into the data and identifying how these affect our results.

Simulation Results – foodUser_auto.py

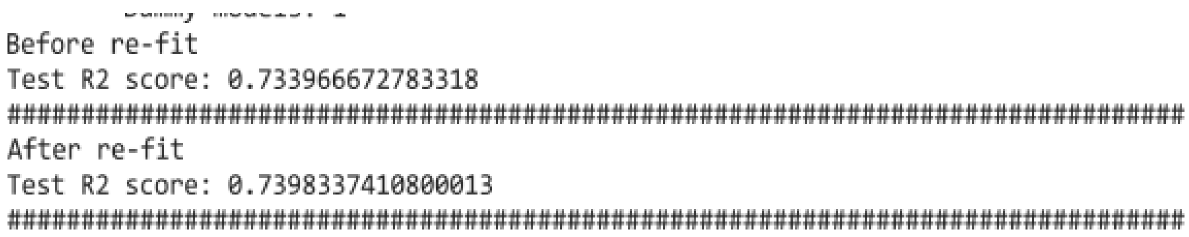
This model was built using Auto Sklearn Regression Library. It was run on Ubuntu Instance using Jupyter Lab IDE.

Auto Sklearn prints out the type of models it found that would fit best for the given data and expectations and then performs its algorithm choosing the most cost and time efficient one. According to the leaderboard, there were 3 processes that AutoRegressor Found – liblinear_svr, k_nearest_neighbour(KNN) and gaussian_process. Based on the most cost efficient and ensemble weight, the process chosen was liblinear_svr.



```
          rank  ensemble_weight                type       cost  duration
model_id
14           1             0.46       liblinear_svr   6.738509  1.562520
49           2             0.04  k_nearest_neighbors   9.433448  1.999790
22           3             0.50    gaussian_process  17.008425  1.969288
```

The model was tested on User 0, and we got adequate results. Upon Fitting and Re-fitting the model, we achieved an R2 Score of ~74% !



```
    Dummy models: 1
Before re-fit
Test R2 score: 0.733966672783318
################################################################################
After re-fit
Test R2 score: 0.7398337410800013
################################################################################
```

This means that the predicted values were almost 74% match to the test dataset.

Conclusions

Upon conducting our simulation and analyzing the results through the *foodUser.py* program that uses the XGBoost Model, we were able to cross-verify our findings with the *collective model* output of food group (F5) being the most significant contributor towards causing IBS symptoms, and food group (F9) having the least impact on the IBS symptom value. With this cross-verification, we can confidently conclude that our outcomes from the two models are validated and can be utilized to draw conclusions when building a treatment plan for IBS patients.

The scope of this model can be expanded to incorporate all food groups and their subcategories to identify which food categories are the common cause of provoking IBS symptoms in the human body at a larger scale. Additionally, we can explore the implementation of the Food Hierarchy, which categorizes food as Parent and its children, to investigate how they impact the correlation value based on their relationship. This analysis can also aid in the research of decomposing the microbiological structure of food compounds and identifying which sugars or acids can potentially cause IBS symptoms.

The implementation of the Auto Regressor in *foodUser_auto.py* led to better results in prediction values of the symptom score for a food category for individual users. Furthermore, the Spearman Correlation helps us identify the food categories affecting an individual user the most, and thus the output can be used to build a diet plan for the user. By integrating the two models and ideas, we can even predict the values of the symptom score for a particular food category intake pattern to implement in the diet plan.

Machine learning models, like the ones mentioned above, provide us with an opportunity to delve deeper into IBS symptoms and identify links between food categories and individuals. There is still much research that can be done in this field, and it would be fascinating to gather more datasets, information, and conduct further analysis.

References

Auto-sklearn Regression. Regression - AutoSklearn 0.15.0 documentation. (n.d.). Retrieved April 13, 2023, from https://automl.github.io/auto-sklearn/master/examples/20_basic/example_regression.html

XGBoost documentation. XGBoost Documentation - xgboost 1.7.5 documentation. (n.d.). Retrieved April 13, 2023, from https://xgboost.readthedocs.io/en/stable/

PharmaHacks 2023 Challenge. (n.d.). Retrieved April 13, 2023, from https://pharmahacks-2023.devpost.com/

Google. (n.d.). *Pharmahacks_resources*. Google Docs. Retrieved April 13, 2023, from https://docs.google.com/document/d/1npsnEwplm6xE-mlHjcpUq4r5dROXyC-cbc48Vx-I5do/edit