Seiveright Cargill Dujohn microcargill@hotmail.com

Introduction

After opening the Excel file containing the dataset, it was noted that there are exactly 1315 entries with 18 columns on the 'data' Sheet. Also, it was discovered that the data consists of the physical statistics (weight, height, age) of men and women, their respective nutritional requirements (energy in kcal, protein, fat, and carbohydrate in grams), the respective information for the meal they consumed (energy in kcal, number of dishes, protein, fat, carbohydrate, vegetable, and salt in grams), and meal type, whether breakfast, lunch, or dinner. Also, the dataset contains the assigned scores of the meals ranging from 1 to 4 where 1 is the worst, and 4 is the best. Moreover, it was highlighted that the salt requirement for men is anything less than 8 grams and less than 7 grams for women, as well as the vegetable requirement which is at least 350 grams per day.

With all this data, it can be gathered that the overarching theme is nutrition. There is potential to extract much information from the provided data. For instance, the BMI of each entry can be calculated since we have their weight and height. In additional, the registered nutritional requirement can be compared with their actual meal to check whether they are fulfilling those requirements. It is widely known by nutritionists that nutrition plays a vital role in health and well-being. Furthermore, nutritional requirements (amount of nutrition needed to maintain health and reduce the risks of diet-related diseases) vary between individuals and life stages[1].

Preprocessing

It is important to perform exploratory data analysis and preprocessing to see trends in the data and decide whether feature engineering is required. Seeing trends in data can give insights that are not obvious by just looking at numerical values, show critical information in an easy-to-understand format, and help decide which algorithm would work best for the data. Additionally, we can figure out if the data is clean enough before feeding it to an algorithm. Providing data with anomalies, and or null values can render an algorithm's predictions inaccurate.

First, the data was skimmed through to have a general understanding of its contents. Afterwards, the data was loaded using the Pandas library and the first 5 entries were outputted. The first column of indices was deemed unnecessary and was therefore removed. A quick statistic of the columns with numerical values was printed:

[1] https://www.nutrition.org.uk/healthy-sustainable-diets/nutrient-requirements/

microcargill@hotmail.com

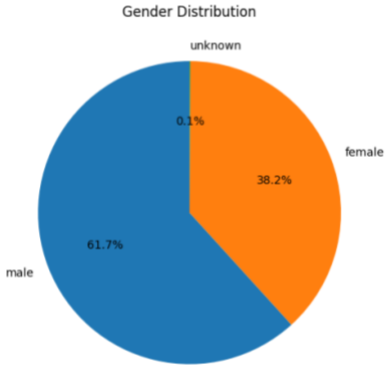
	age	height	weight	EER[kcal]	P target(15%) [g]	F target(25%) [g]	C target(60%) [g]
count	1314.000000	1314.000000	1314.000000	1315.000000	1315.000000	1315.000000	1315.000000
mean	39.703196	166.562405	58.714612	2176.253992	81.604021	60.447423	326.416085
std	9.369062	6.498895	8.368238	313.538696	11.757694	8.709403	47.030777
min	22.000000	152.000000	45.000000	1545.000000	57.949864	42.925825	231.799455
25%	35.000000	160.000000	51.250000	2020.000000	75.731250	56.097222	302.925000
50%	39.000000	167.000000	58.000000	2105.000000	78.918750	58.458333	315.675000
75%	44.000000	173.000000	63.000000	2376.000000	89.114456	66.010708	356.457825
max	62.000000	179.000000	91.000000	3380.000000	126.750000	93.888889	507.000000

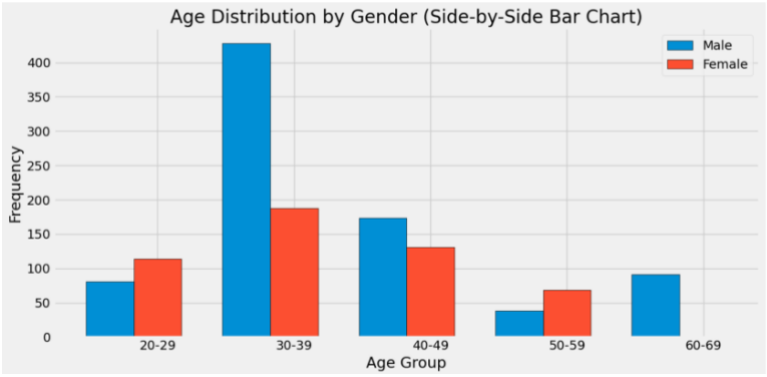
The previous image shows the table split vertically. The other section is below:

number of dishes	E[kcal]	P[g]	F[g]	C[g]	Salt[g]	Vegetables[g]	Score(1:worst 2:bad 3:good 4:best)
1315.000000	1315.000000	1315.000000	1315.000000	1315.000000	1315.000000	1315.000000	1315.000000
2.707224	594.993894	21.301460	23.551567	69.478464	3.002091	72.791148	1.965779
1.613367	309.082985	12.804031	18.162848	36.229788	1.995371	80.957033	0.712713
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
1.000000	371.690000	12.265000	9.945000	42.360000	1.490000	0.000000	2.000000
2.000000	564.170000	20.150000	19.440000	67.750000	2.720000	47.000000	2.000000
4.000000	780.430000	28.125000	33.085000	92.715000	4.200000	120.850000	2.000000
13.000000	2382.340000	94.010000	141.660000	239.760000	14.740000	621.500000	4.000000

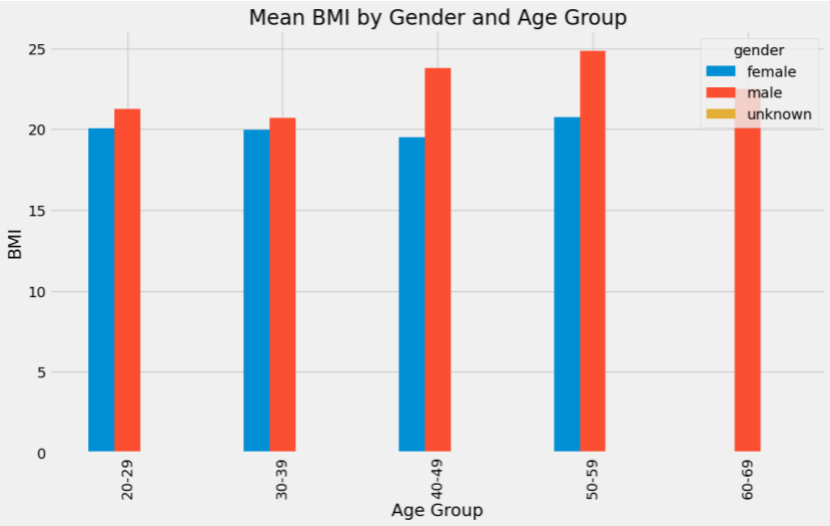
Histograms were plotted for age, height, vegetables, salt, carbohydrate, protein, fat, and number of dishes. Taking in the previous statistical description of the data, nothing seemed out of the ordinary and most of the data points were seen as normal. The only feature engineering done was to factor in the unknown gender and replacing the unknown age, weight, and height with the median. The histograms can be viewed in the Jupyter Notebook The data was then check for any empty values. Only one entry contained empty values for meal type, gender, age, height, and weight. It was decided that the entry would be kept until before training an algorithm.

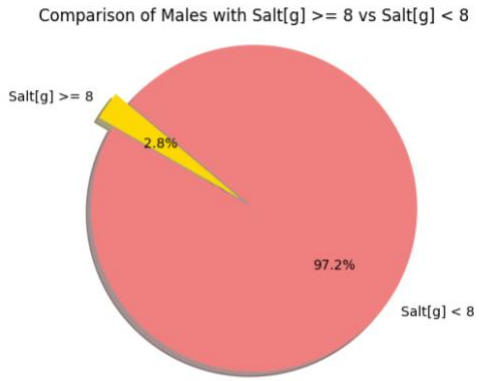
Below are some graphs and charts that were plotted. Reading from left to right: Gender Distribution Pie Chart, Side-by-Side Bar Chart showing Age Distribution by Gender, Bar Chart showing mean BMI by gender and age group, Pie Chart showing men who consumed >=8g of salt compared to men who consumed less, Pie Chart showing women who consumed >=7g of salt compared to women who consume less, Pie Chart showing men who consumed >=350g of vegetables compared to men who consumed less, Pie Chart showing women who consumed >=350g of vegetables compared to women who consumed less.

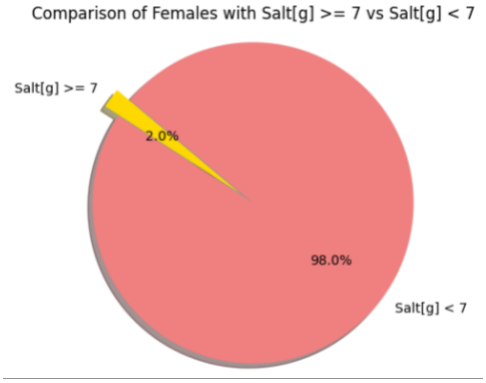


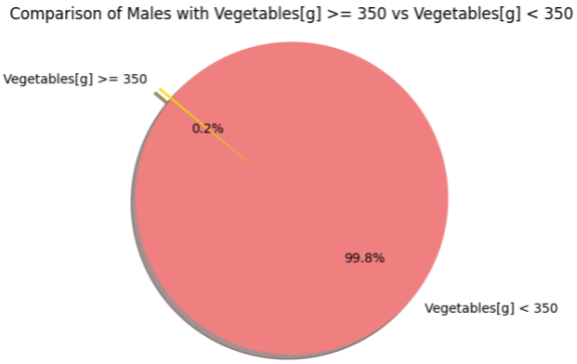


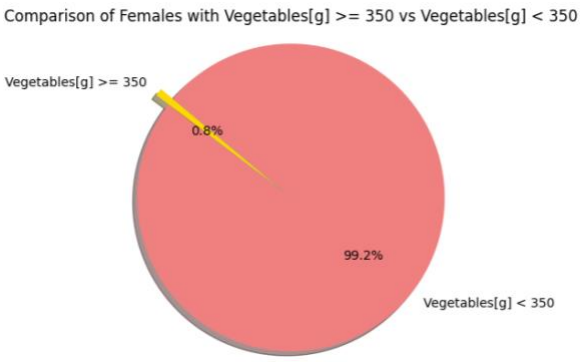
microcargill@hotmail.com







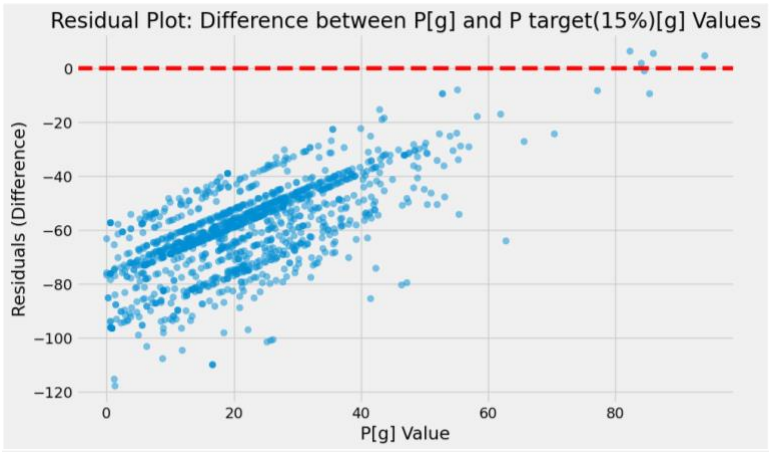


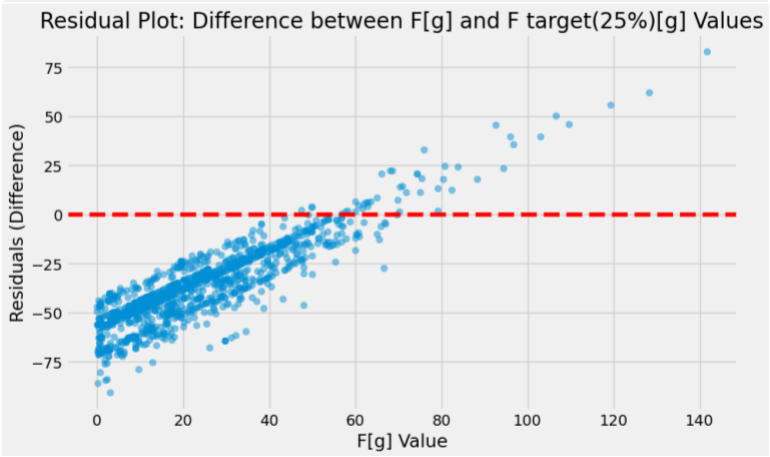


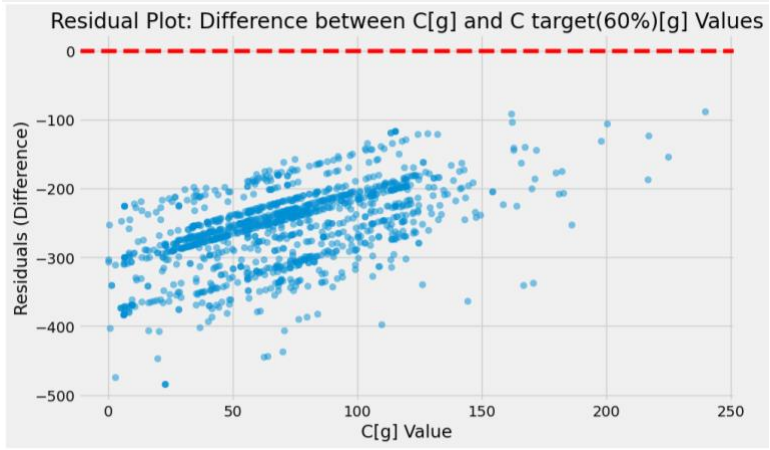
Below are residual plots (values on red dashed line means the values are the same, values above are over the target value, values below are under the target value where the y-axis shows the difference in magnitude) between:

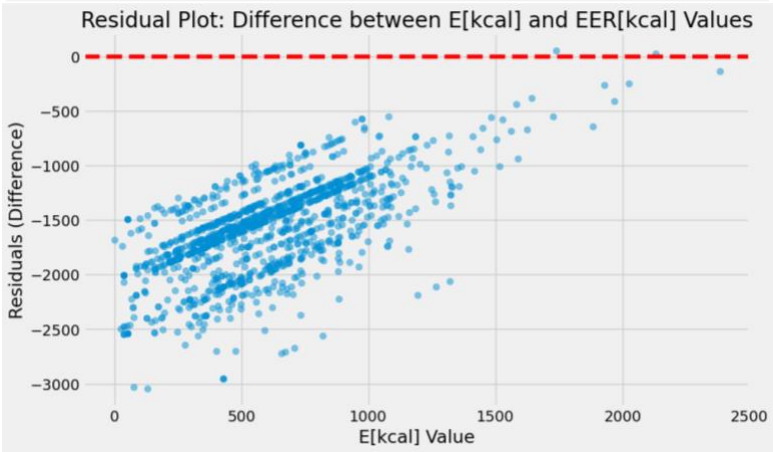
- Consumed protein (g) vs Target protein (g)
- Consumed fat (g) vs Target fat (g)
- Consumed carbohydrate (g) vs Target carbohydrate (g)
- Consumed energy (kcal) vs Target energy (kcal)

microcargill@hotmail.com









Seiveright Cargill Dujohn
Hacarus Test (Internship) 2023
microcargill@hotmail.com
Analysis Report

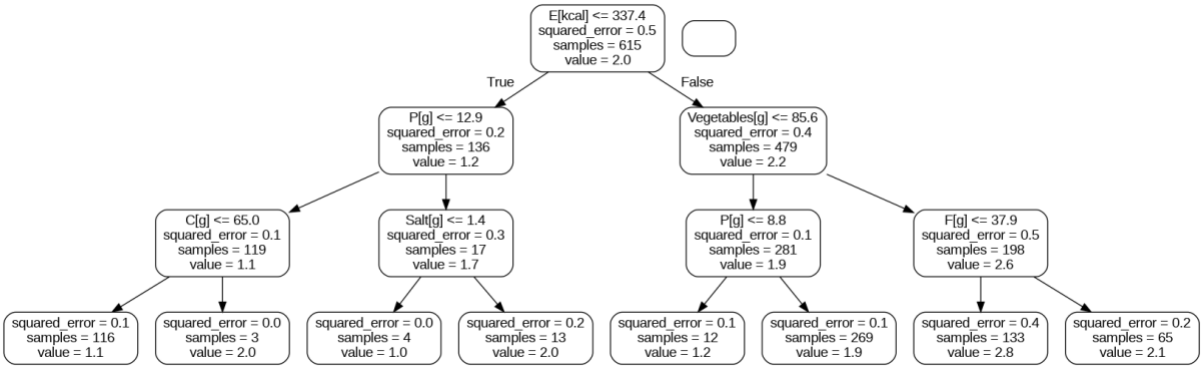
Algorithm Selection

Since the target values where provided, supervised learning algorithms are best suited. Therefore, Random Forest (RF), Support Vector Machines (SVM), and Linear Regression (LR) algorithms were considered. Random Forest and SVM were selected because there are multiple data points. Furthermore, the decision tree of the RF algorithm could be shown to get a brief overview of how the algorithm arrived at its predicted values. Despite the uncertainty of the potential performance of using LR, it was still tested.

Initial Results

RF achieved an accuracy of 87.55%, SVM achieved an accuracy of around 73%, while LR achieved an accuracy of 75.38%.

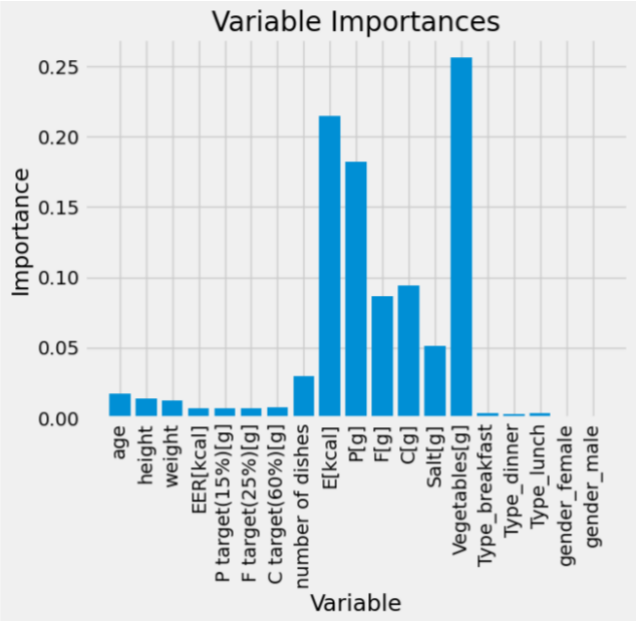
Below is a simplified decision tree of the RF algorithm. It shows for instance where the score 2 is assigned when the E[kcal] is less than or equal to 337.4, then diverges between P[g] being less than or equal to 12.9 and P[g] is greater than the 12.9. For the branch where P[g] is less than or equal to 12.9, the predicted value is 2 where C[g] is greater than 65.0 while for the P[g] is greater than 12.9 branch, the predicted value is 2 when Salt[g] is greater than 1.4.



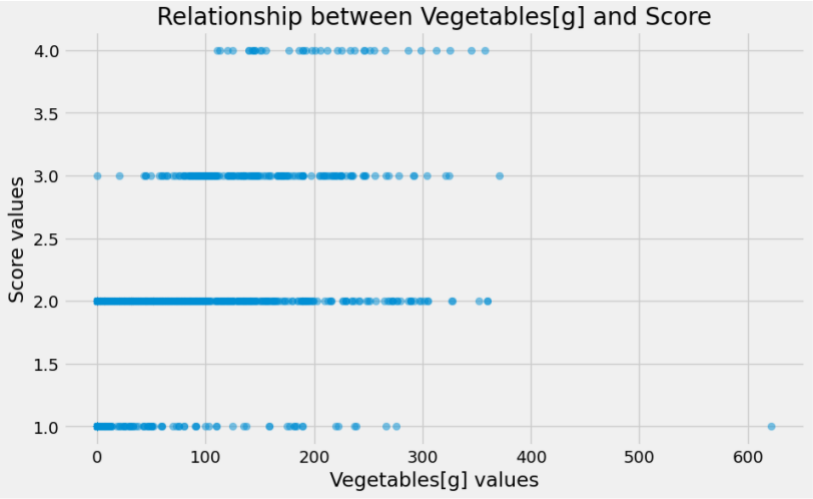
Below is a scatter plot showing the Test (actual scores) vs Predicted Values (predicted scores). Points closer to the red line are more accurate.

microcargill@hotmail.com

The importance of each variable is shown in the bar chart below. All the values add up to 1.00 (high indicates high importance)



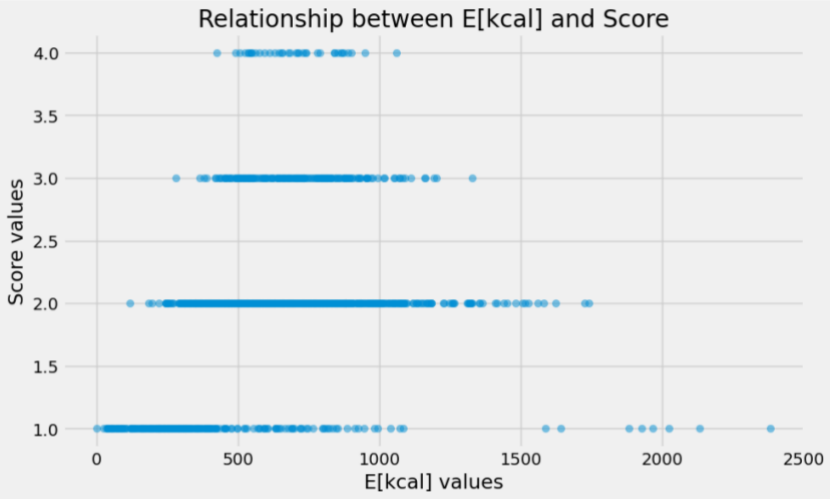
Since vegetables and consumed energy are the top two most important variables dictating the score, the relationships between those two respective variables and the score are shown below:



Higher scoring meals tend to have vegetables approximately between 90g and 300g.

Meanwhile the scatter plot below shows that higher scoring meals usually have energy approximately between 490kcal to 1200kcal.

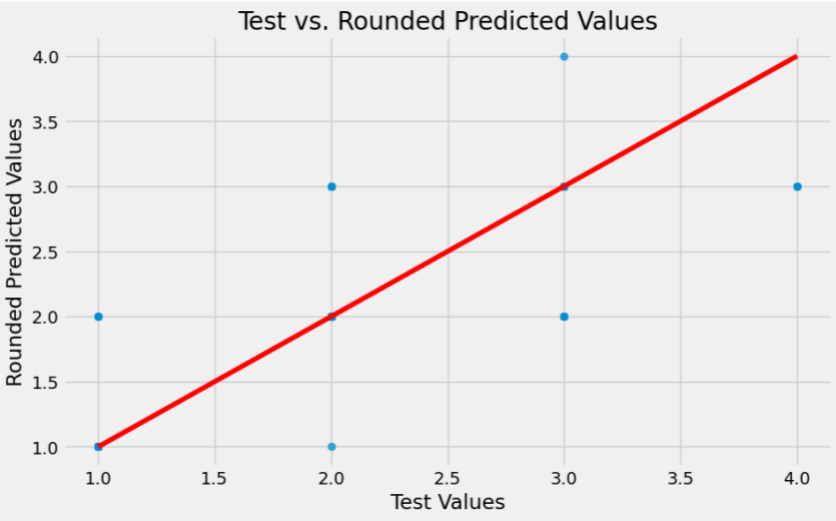
microcargill@hotmail.com



Hyperparameter Tuning

It was noticed that the RF algorithm predicted floating point scores even though the actual scores are only whole numbers between 1 and 4. Therefore, the predicted scores were rounded off and the accuracy increased to 90.63%. Moreover, RF with the 6 most important variables (Vegetables[g], E[kcal], P[g], F[g], C[g], and Salt[g]) was tested and achieved the score 87.27%. Grid search was also implemented to find the best possible parameters for RF. However, the search was time consuming and yielded parameters that did not outperform the previous parameters. Other hyperparameter tuning methods could be investigated and implemented.

Below is a scatter plot of the predicted scores rounded off vs the actual scores. Points closer to the red line are more accurate.



Conclusion

First, it should be noted that the total meal consumed for each person could not be calculated since there is no connecting variable for the meal types with the respective individual and trying to do so algorithmically would significantly increase the complexity and introduce uncertainty of the matched entries. Also, a significant portion of the entries, 61.7%, are males [Gender Distribution Pie Chart]. Most of these males are ages 30-49 [Side-by-Side Bar Chart showing Age Distribution by Gender]. Men and women have different nutritional requirements and are susceptible to diet-related diseases at different rates. Additionally, men between 40 and 59 are close to being overweight[2] according to the bar chart showing mean BMI by gender and age group. Only 2% and 2.8% of women and men respectively consumed higher than the daily requirement of salt in one serving. However, an overwhelming majority 99.8% and 99.2% of men and women respectively consumed less than 350g of vegetable in one meal serving. This increases their chances of diet-related diseases.

Although the combined meals eaten per day was not calculated for each individual, almost all of the entries did not surpass the required energy and carbohydrate for one meal (according to the respective consumed vs target residual plots). Nonetheless, the required fat for the day was exceeded in one meal for some individuals (according to the consumed vs target fat residual plot). Only a few individuals exceeded their protein requirements for one meal for the day (according to the consumed vs target protein residual plot).

Vegetables, total meal energy, protein, carbohydrates, followed by fats tend to be the most important factors for predicting the meal's score. Meals having vegetables around 90g-300g and total energy around 490kcal-1200kcal tend to score higher.

In summary, majority of the dataset's demographics include men between the ages of 30 to 49 where men between 40 and 59 are close to being overweight. In addition, majority of the entries eat meals with less than the daily amount of vegetables required, but it should be noted that the total vegetable consumed for the day was not calculated. Moreover, some of the entries consumed fat exceeding their daily requirements in one meal alone. Last, Random Forest was used to predict the scores and it was noted that some features of the dataset are significantly correlated to the score of the meal.

[2] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html