Which Nutrient or Nutrients are Accurate Predictors of Total Energy in Food Items?

Elements of Data Processing - COMP20008

Assignment - 2

Group - 041

Group Members:

Sahil Khatri - 1280740 (Leader)

Daksh Agrawal - 1340113

Harshit Badam - 1332157

Ray Kye Tan - 1307508

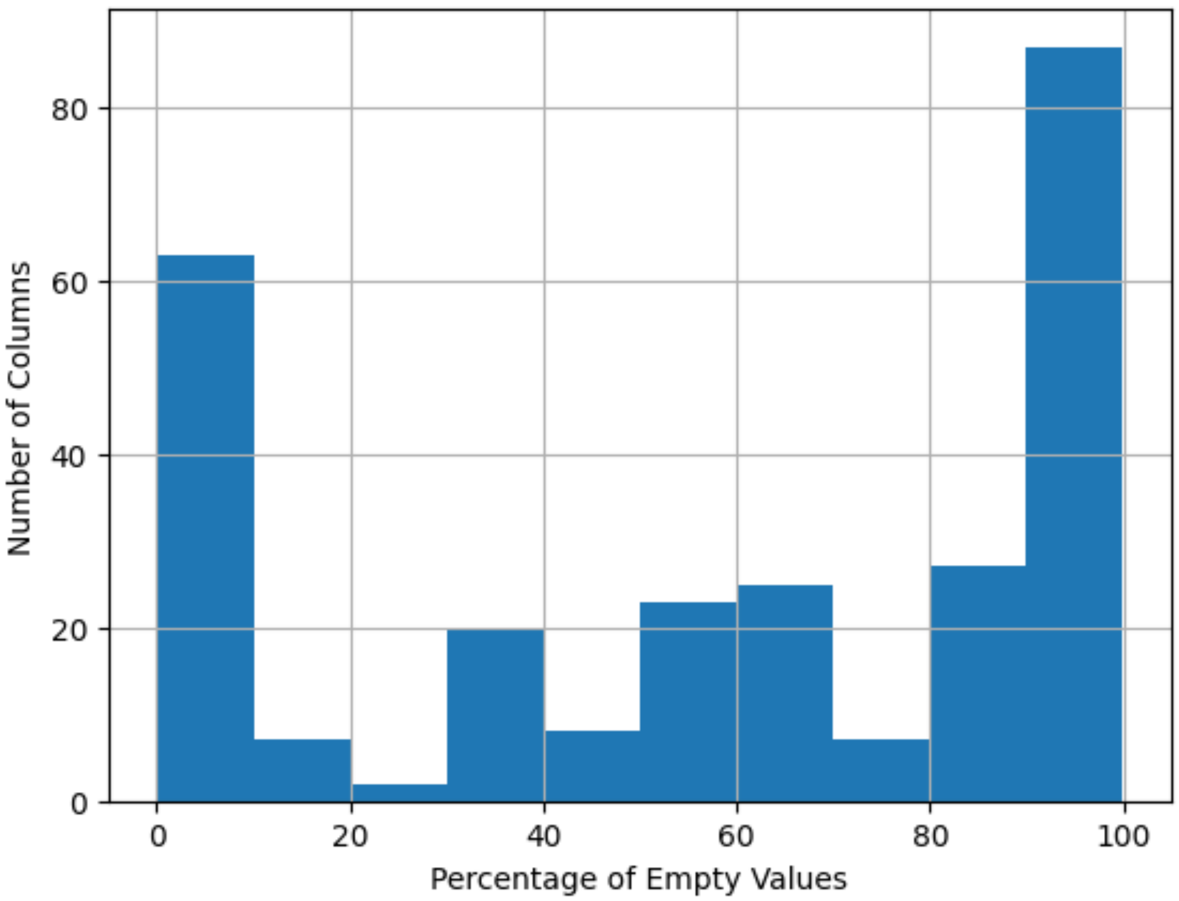
Table of Content:

Aim:

The aim of our project is to indentify and illustrate which nutirent or set of nutrients can help us in predicting the Total Energy content in various food items and illustrate how well they are co-related. The result of this project can help Food Scientists check and validate their analysis, Food Manufacturing Comapnies in finding the energy content in their products for meeting set criterias and saving costs on nutritional analysis for energy, and nutritionists in making nutrition plans and offering dietry recommendations.

Dataset:

We choose the dataset “Australian Food Composition Database data” which is an excel format file which we used for our research, model training and analysis. The dataset is derived from the “Nutrient File” which is originally found on the official website of Food Standards Australia New Zealand (Food Standards, 2022a). There are 1616 records - representing different food items - and 293 fields - representing the Public Food Key, Classification, Food Name and 290 nutrient fields. 57% of the cells in the dataset are empty. Out of the 293 fields, there are 169 fields which are empty for more than 50% of the records while there are 63 fields which are almost full (that is, they are full for more than 99% of the records).

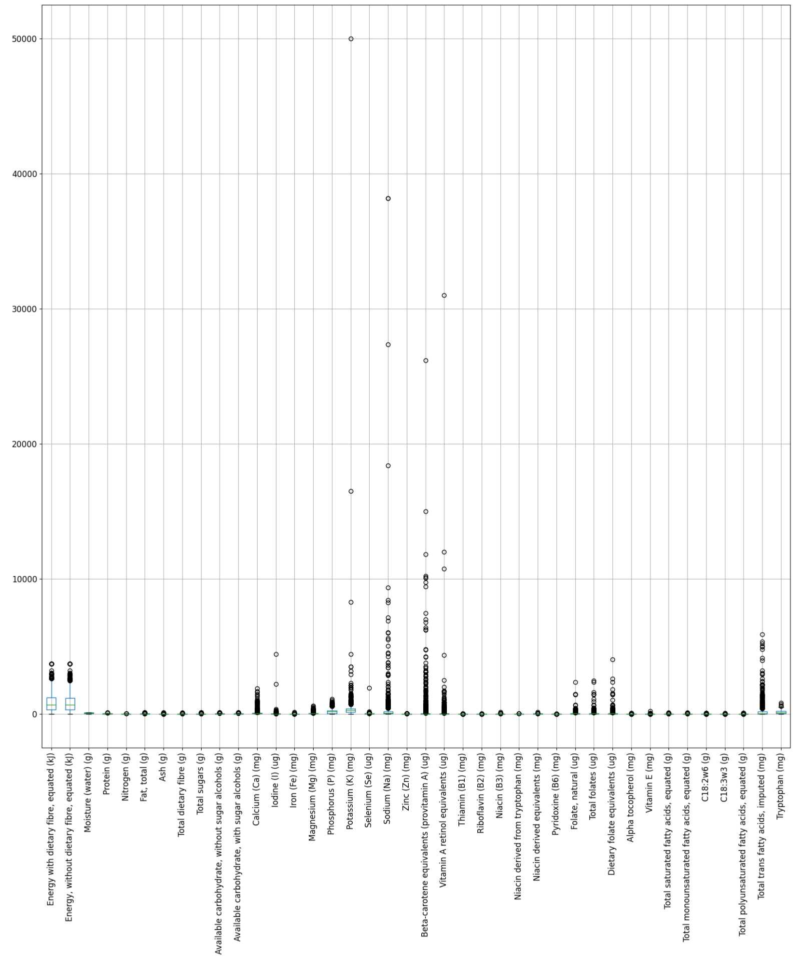


Our target variable in the dataset is Energy with dietary fibre, equated (kJ).

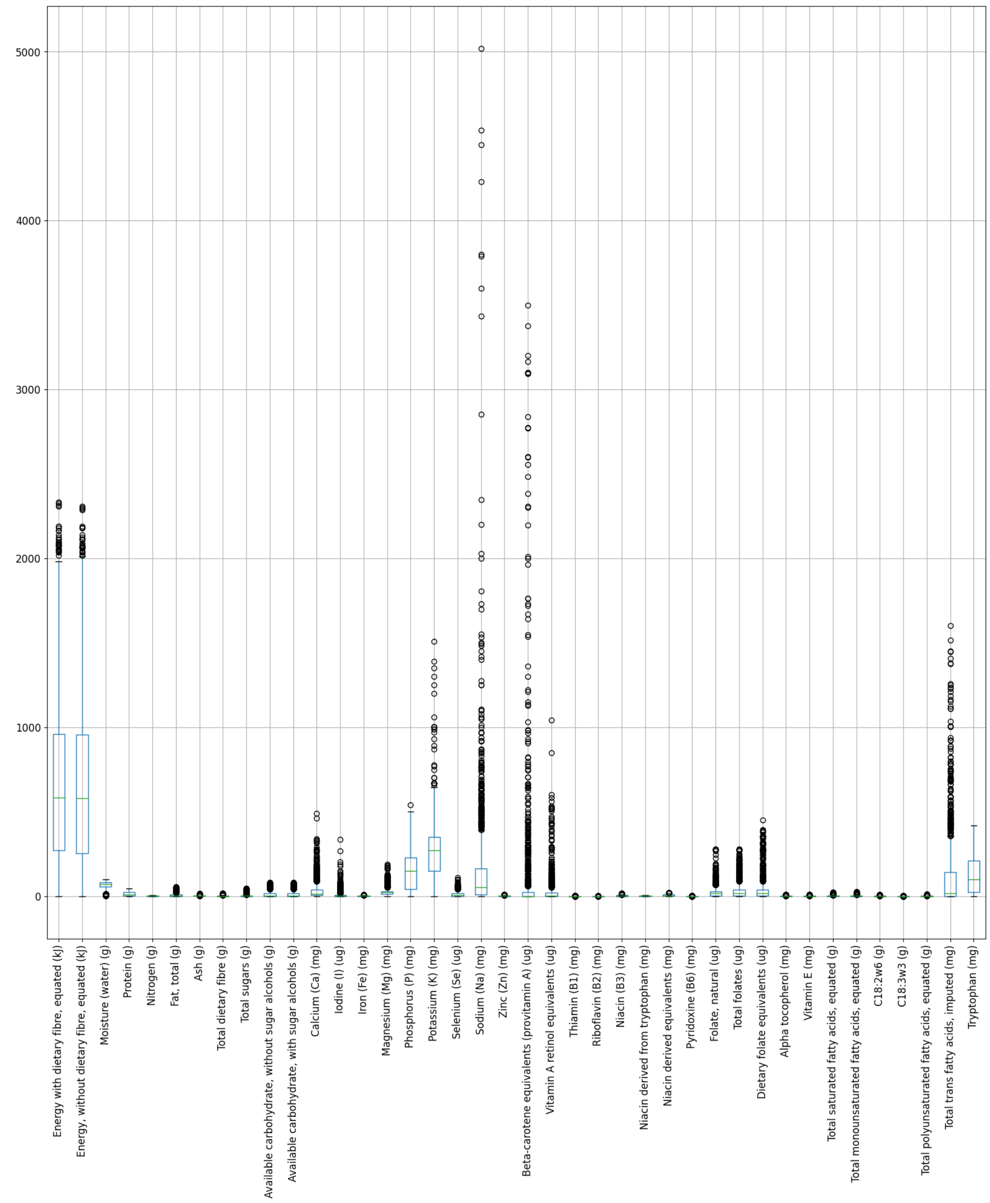
3) Pre-Processing and Wrangling:

Before using the dataset when the excel sheet was initially opened, it was filled with empty columns and unusable data which needed to be filtered out and cleaned. Firstly, the excel sheet format of the dataset was not in a python usable format so it needed to be manually exported to python by the process of first converting it into csv file and then importing it to python. Then, we split the data into training (applied to train or fit your model) and testing (is used for unbiased model evaluation) datasets as it is an unbiased evaluation of prediction performance.

Next, a notable portion of the columns had a significant persentage of the values missing these columns had to be filtered out as they couldnt provide good insights, imputing these missing values with the mean of the given values only works if the number of missing values is small, instead median or mode could have been used but it would have resulted into a bias towards the chosen value. Then, columns with more than half the values as zeros were dropped as these values did not provide solid grounds for usability.



(Before Outlier Removal)



(After Outlier Removal)

This was followed by droping the columns which were not numbers such as 'Classification', 'Public Food Key', and 'Food Name' as these provided no real use in our test case. Next, We created a table providing us a depper understanding of the data given, this table described the mean , median, varients, and persentiles which pointed out that significant outlies were present in the dataset, this would consequently bias the outcome of the model, therefore these outlires needed to be removed.

Finally, since linear regression is highly sensitive to the range of the features we normalize all the columns using sklearn minmaxscalar and apply the same normalization to the test dataset as well.

4) Analysis Methods:

4.1) Methods:

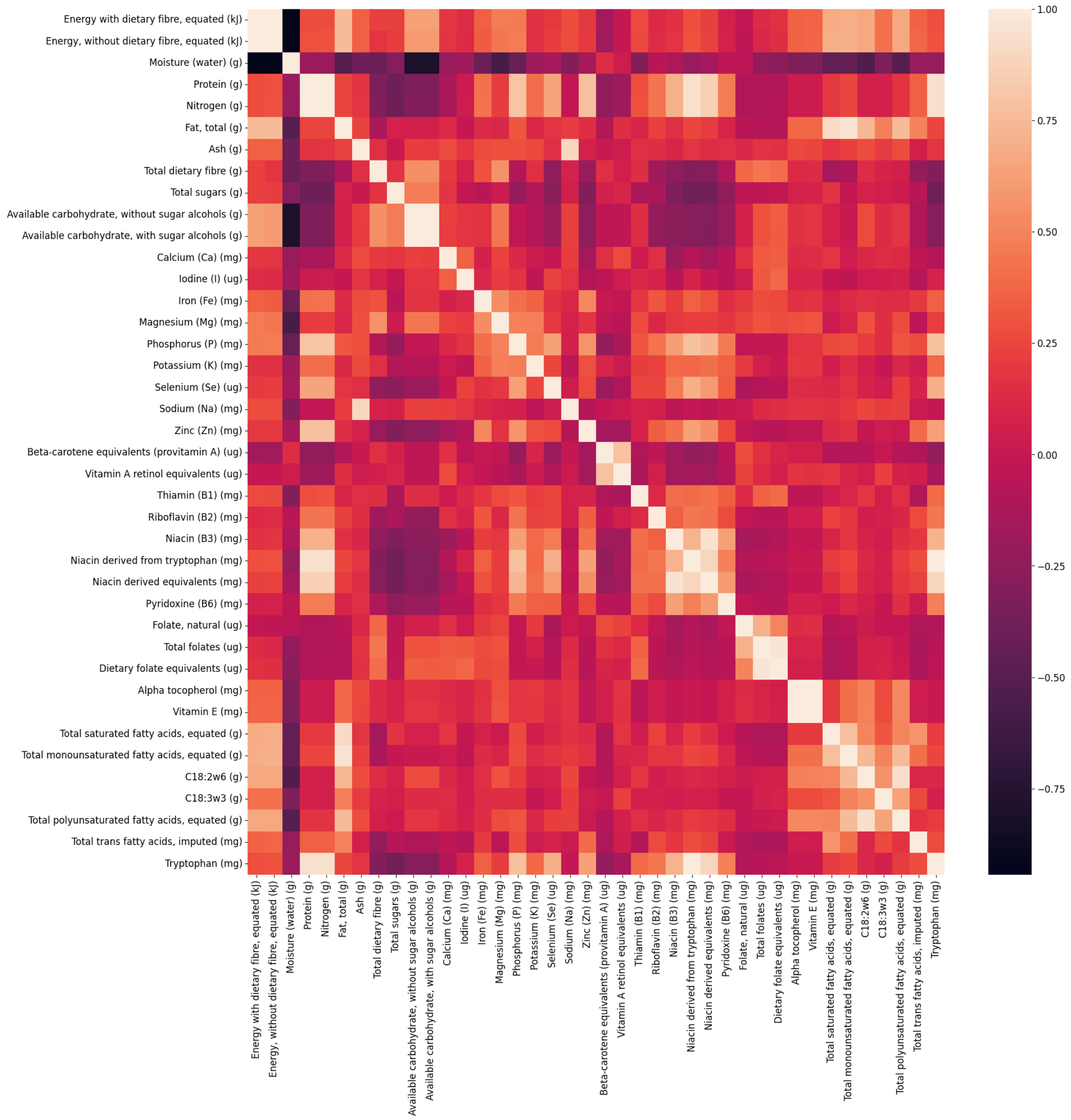
The nutrition dataset being used in this project contains mostly numeric data which represents the amount of various nutrients in different food items. Keeping this in mind along with our aim and approach helps us indentify from a range of available techniques and methods. We used Supervised learning techniques including linear regression - for finding the highest contirbuting nutrients to the total energy conent (with dietry fibres) - and Pearson correlation - to confirm the relationship between moisture, folates and energy. In this project we have not used any proper feature selection techniques, however we did remove certain fields which were too empty or contained to many zeroes. The validity of our results can we been seen from the plots given below

4.2) Train-Test Split:

After importing the dataset to python, we performed a 80-20 split to split the data into training and test sets. We used the training set for fitting the model and the preliminary analysis. After that we used the test set to analyse the fit and the accuracy of the model. We did not employee K-fold or Bootstrap testing techniques because the size of the dataset is large. This means that the test and training sets are large and thus eliminating the need for repeated creation of new training ans test sets for accuracy.

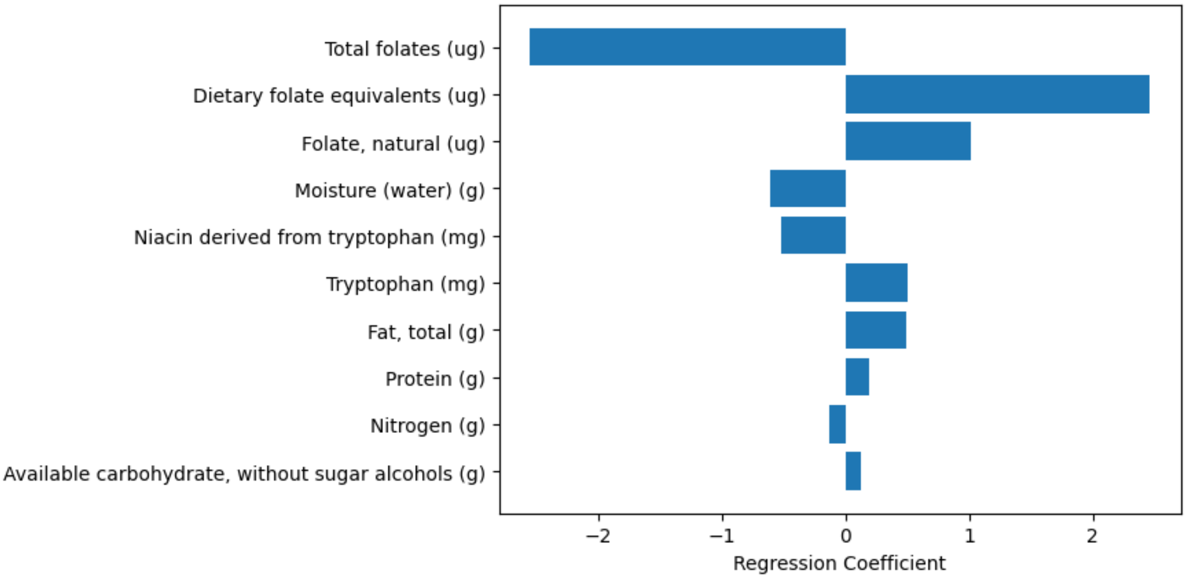
4.3) Preliminary Analysis:

To get a broader understanding of the data we plotted a heat-map of the Pearson correlation with the columns remaining in the dataset after cleaning . In the plot given below, we observe that the dark band corresponding to moisture shows a strong negative correlation with the other nutrients and that the strongest relation moisture shows is with total energy with dietry fibres (kj).

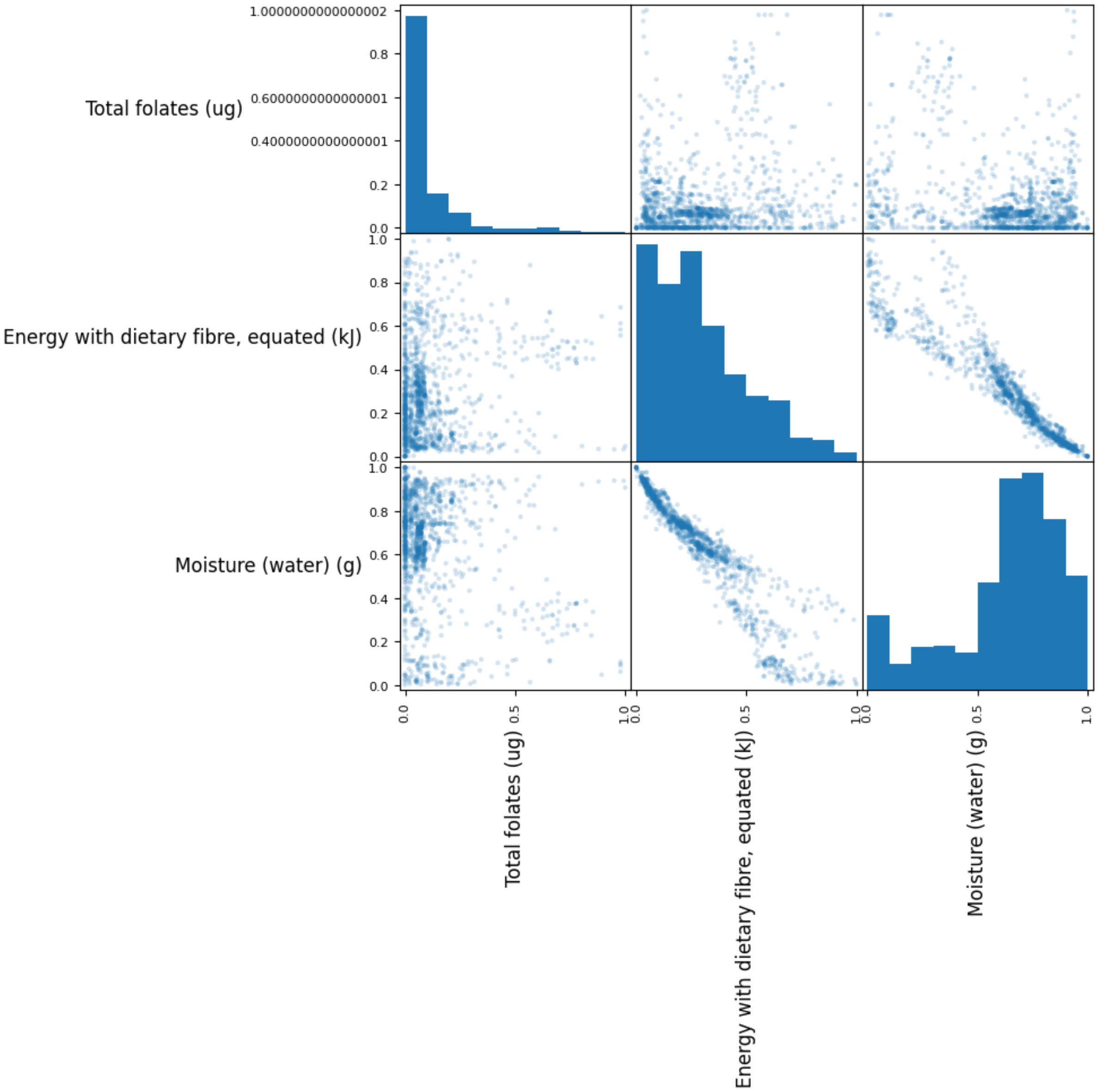


4.4) Modelling:

We fitted a linear regression model to better understand the dataset provided and the correlations between various nutrients using the line of best fit. To better comprehend the total prediction error of the model, we next compute the mean square error (value recieved for MSE for the test dataset is 0.0028846043713276768, and for train dataset is 0.0002202075936639337). The relationship between nutrients and energy is then discovered by coefficient analysis, which is then condensed to the top ten nutrients with the highest correlation value and plotted.

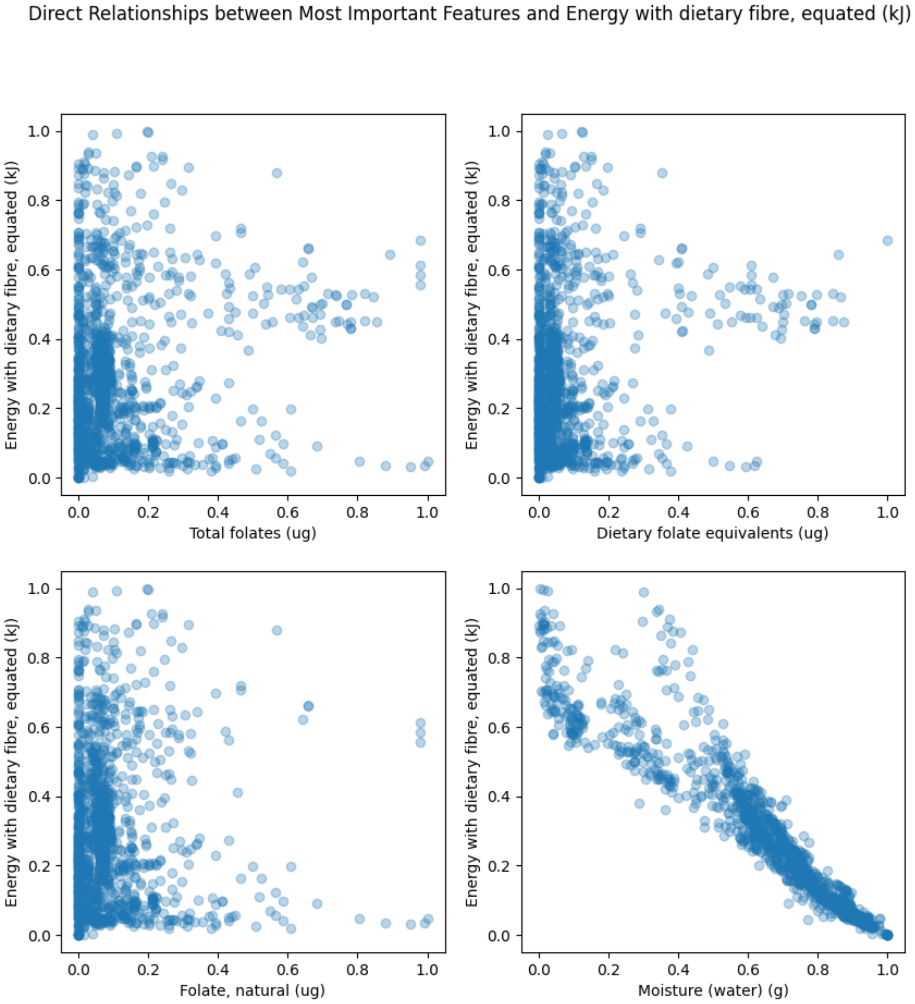


The next step is to use a scatter plot to plot the associations between the most crucial features and energy in order to look for any trends. To further illustrate how closely the features are connected with one another, we then compute the pearson correlation for the features from the previous step and produce a heatmap based on the results. In order to see the relationships, we generate a scatter matrix of the features.

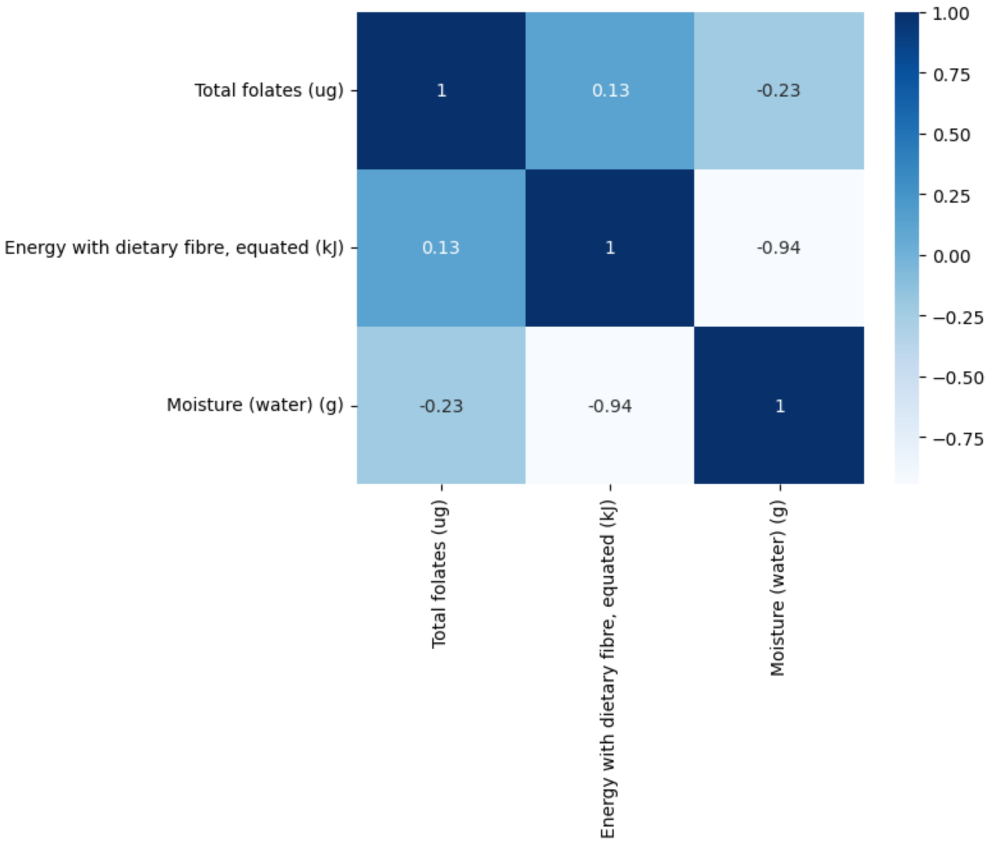


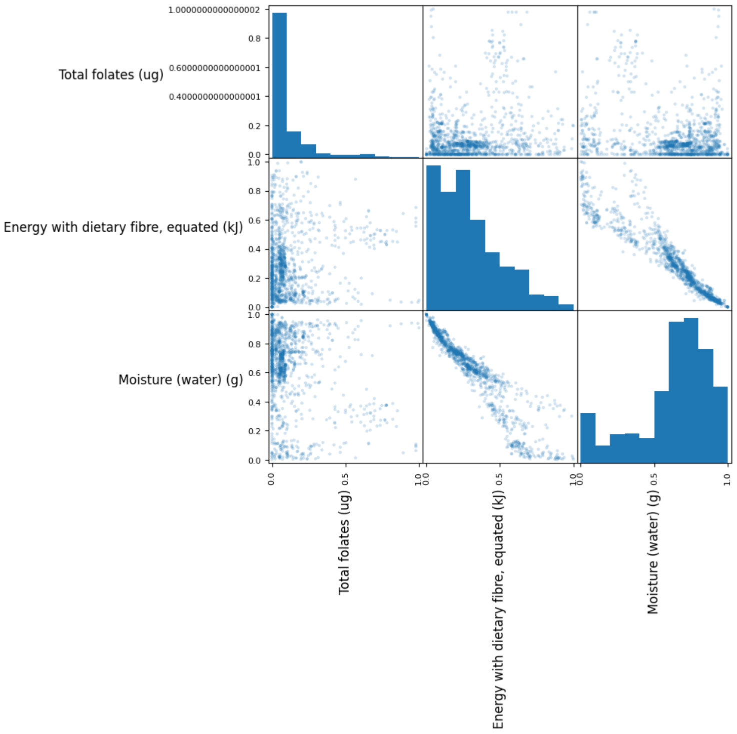
5) Discussion:

This section includes a discussion of the model's findings and their importance. First, we may infer from linear regression that folates and moisture are reliable indicators of energy. Second, our linear regression produced adequate results, as seen by the test dataset's mean square error of 0.2%. After calculating the mean square error, we note that it is incredibly low, which leads us to believe that the model's average prediction is quite accurate. The scatter figure shown below further supports the relationship between moisture and total energy, while the link between folates and energy is less clear.



Additionally, the scatter matrix and heatmap shown below confirm this.

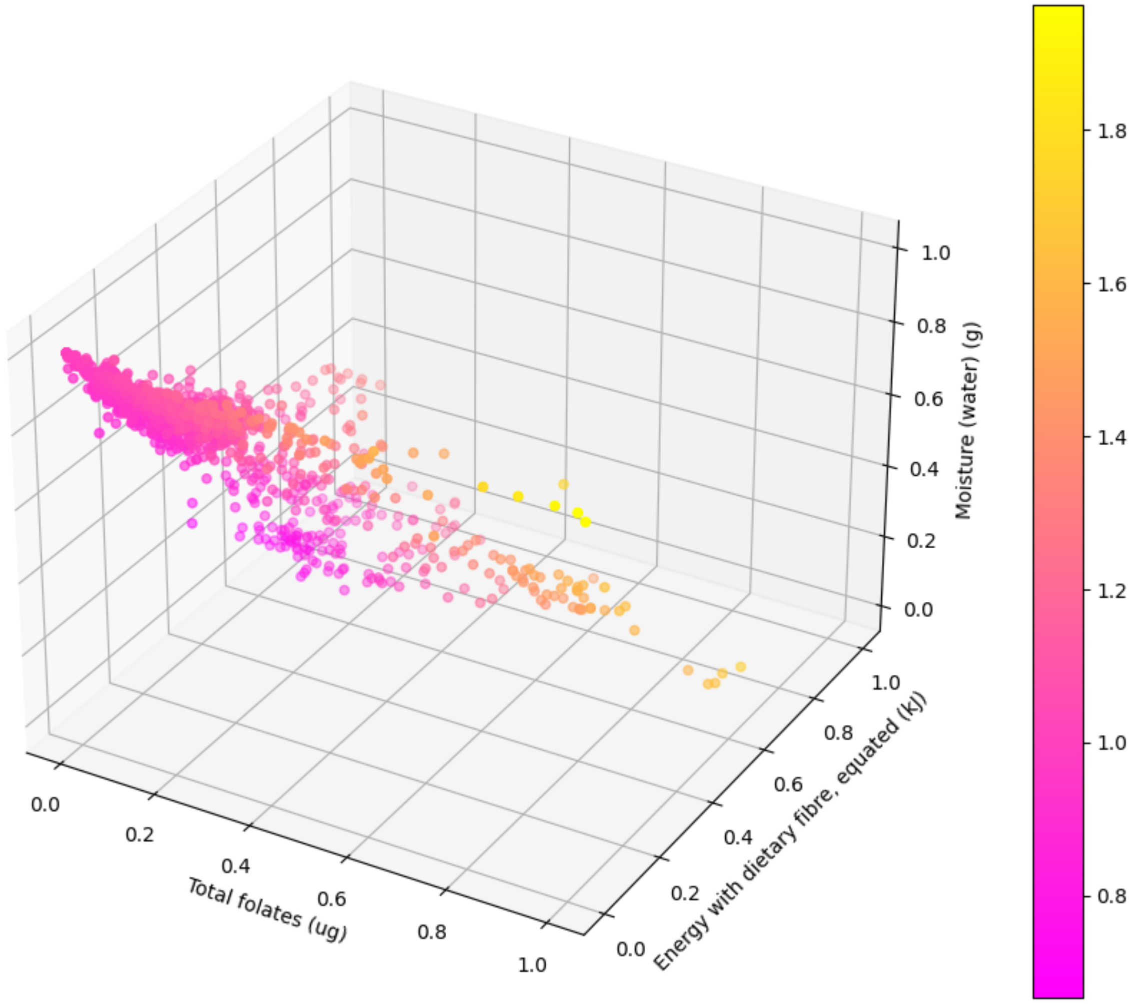




From the heatmap, we can see that the pearson correlation between energy and moisture is highly negative, indicating a significant inverse correlation. We also calculate the correlation coefficient, which leads us to the conclusion that moisture and folates have the strongest link with energy.

Then, we calculate pearson correlation which furthermore builds upon the stated conclusion of strong relationship between total energy and moisture.

Finally, we create a three-dimensional graph between the three features that shows how strongly they are related to one another.



These findings suggest a close relationship between energy, moisture, and folates. Additionally, we can conclude that even though folate by itself is not a particularly reliable indicator of energy, it significantly aids the prediction when combined with moisture.

6) Evaluation:

In this section we examine the limitations of our results and propose possible improvements:

1) Dataset Limitations and improvements :

- The data in the dataset used is collected over a period of 40 years from the 1980s. This long data collection timeframe in combination to changes in the analysis techniques may result in some inaccurate data (Food Standards, 2022b). Apart from this, the dataset is quite empty as a significant proportions of the columns present had mostly null values rendering them unusable. Another issue is the presence of significant number of zero values present in the columns as the trace values (less than 0.2g/100g) of nutrients in foods were reported as 0 (Food Standards, 2022b). This results in slight variation from the true results and inaccurate predict.

- These limitation can we overcome by including data from other external and preferably latest sources to fill the empty cells and to cross check the existing values. This can also helps in dealing with the absence of trace amounts on nutrients in the original dataset.

2) Analysis Methods Limitations and improvements :

- Pearson correlation and linear regression cannot capture and illustrate complex relations.Most nutrients are distributed in a negative exponential fashion. which may be better represented by logarithmic regression. There might also be some clustering in the dataset which cannot be modelled using linear regression thus alternative approaches might be more useful. Finally, linear regression assumes that the effect of each variable is additive which may not be the most appropriate method to pursue, other models may be more suitable for this approach.
- The complex relations between the nutrients can be better illustrated using alternate techniques such as logarithmic regressions. K-NN clustering can be used to deal with clustering in the dataset which may get overlooked by linear regression and Pearson correlation. Feature engineering, though very complex, can we used to capture and represent non-linear and non-logarithmic relations.

7) Reference List:

Food Standards. (2022, January). *Download Excel files (Australian food composition database - Release 2)*. Home. https://www.foodstandards.gov.au/science/monitoringnutrients/afcd/Pages/downloadableexcelfiles.aspx

Food Standards. (2022, January). *Foods and nutrients in the Australian food composition database*. Home. https://www.foodstandards.gov.au/science/monitoringnutrients/afcd/Pages/foods-in-release-1.aspx