Build a model to predict the Vitamin B12 content based on other nutritional data in Animal-based product and dishes

September 20, 2023

Contents

1 Introduction

Vitamin B12, also known as Cobalamin, is an essential nutrient that plays an important role in many physiological processes, including red blood cell formation, and nervous system function (NIH, 2022). However, vitamin B12 deficiency is a widespread problem that not only affects vegetarians and vegans, but also those with poor diets or health conditions.

The aim of this project is to develop models to predict the vitamin B12 content in different foods based on other nutritional data, specifically in animal-based products and dishes. The results of this study can provide valuable insights for individuals experiencing vitamin B12 deficiency to enhance their dietary choices, as well as for health professionals seeking to promote healthy nutrition.

2 Background

2.1 Datasets

The data source for this project is from the Australian Food Composition Database. It includes the identification, classification, and the nutrient data available for each food, as of January 2022. In total, there are 1616 different foods and 290 nutrients (macronutrients, micro-nutrients,...) reported. The database file is in Excel format and has over 62% of missing data entries.

2.2 Target Audience

Firstly, the main target audience for predicting vitamin B12 are people with poor dietary habits and especially elderly, who are susceptible to vitamin B12 deficiency.

Secondly, the prediction models might prove useful for laboratory scientists in estimating the content of vitamin B12 in food. Since it is challenging to precisely measure the level of vitamin B12 in food, developing prediction models based on other nutritional data can provide a more cost-effective approach.

Noted that, vegan and vegetarians is suggested to take supplements of vitamin B12 to ensure sufficient intake since this model only investigates vitamin B12 content in animal-based products.

2.3 Features Description

The variables used in the regression model are described as follows.

Table 1. Features Description

Feature	Description	Measurement
Cobalamin (B12)	A water-soluble vitamin	ug
C22: 5w3	An omega-3 fatty acid, with a chain length of 22 carbon atoms	%T
Iron (Fe)	A mineral	mg
Iodine (I)	A trace mineral	ug
Copper (Cu)	A trace mineral	mg
C17	A saturated fatty acid, with a chain length of 17 carbon atoms	%T

Notes:

(1) All variables are continuous variables.

3 Data Pre-processing

This report section aims to collect clean and continuous data. Figure 1 illustrates key steps to prevent possible problems generated by non-informative data, high numbers of missing data and noises. This will guarantee a high-quality dataset for efficient uses of training algorithm models.

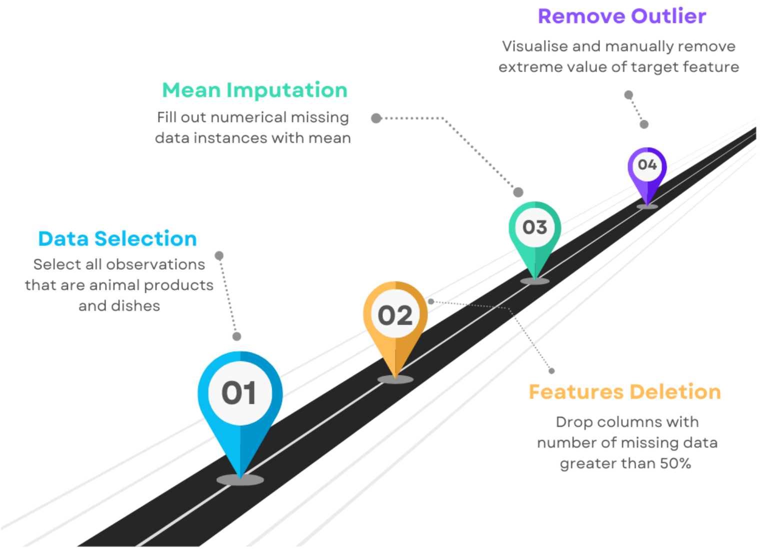

Remove Outlier
Visualise and manually remove extreme value of target feature
Mean Imputation
Fill out numerical missing data instances with mean
Data Selection
Select all observations that are animal products and dishes
01
02
03
04
Features Deletion
Drop columns with number of missing data greater than 50%

Figure 1: Pre-processing steps

3.1 Data Selection

The National Institute of Health (2022) reports that food and dishes of animal origin are known to naturally contain Vitamin B12, thus this report will pay attention to animal-based products. This implies that non-animal items (vegetables, seeds, . . .) will be removed. After performing data selection, 673 observations are left.

3.2 Feature Deletion

The dataset includes a significant number of missing fields, accounting for roughly 40% of the dataset. This poses a challenge when applying mean imputation, as it could introduce bias in columns with a low number of filled data fields. To address this issue, this study selects only columns that have a high number of filled data fields, exceeding a threshold of 50%. This results in the dataset size reduced by nearly half (from 193 to 120 columns).

3.3 Mean Imputation

There is a considerable number of columns that still contain missing data values, despite meeting the threshold of the number of filled data fields. To address this, this report will utilize mean imputation. This approach ensures that the missing data are filled while preserving the statistical properties of the dataset.

3.4 Remove Outlier

Finally, we examine the scatter plot of the target feature Cobalamin B12. Figure 2 highlights an extreme data point ”Lamb, liver, grilled, no added fat” with a vitamin B12 level of 76.5 ug, which is notably high compared to other animal-based foods. It is necessary to remove this outlier from the model, in order to prevent any extreme data point from exerting a disproportionate impact on the analysis.

It is important to note that despite there being many data points with very high values of vitamin B12 compared to the rest, we decide to keep these points since those are the foods that should be especially prioritised by those who are experiencing vitamin B12 deficiency. By doing so, we would appeal better to the target audience.

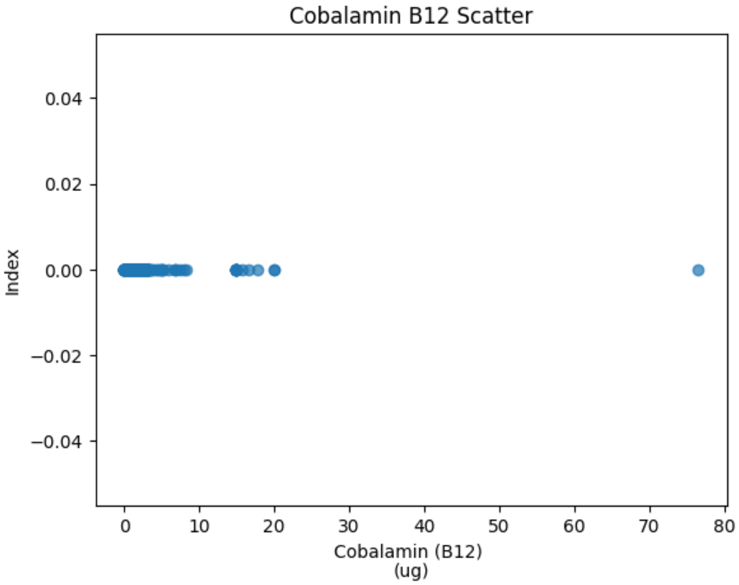


Figure 2: Outlier Scatter Plot

4 Feature Selection

Prior to building the model, it is necessary to perform feature selection to identify and extract the most relevant features. By removing redundant features, the process of analyzing can reduce overfitting, improve model performance, and enhance interpretability of the model.

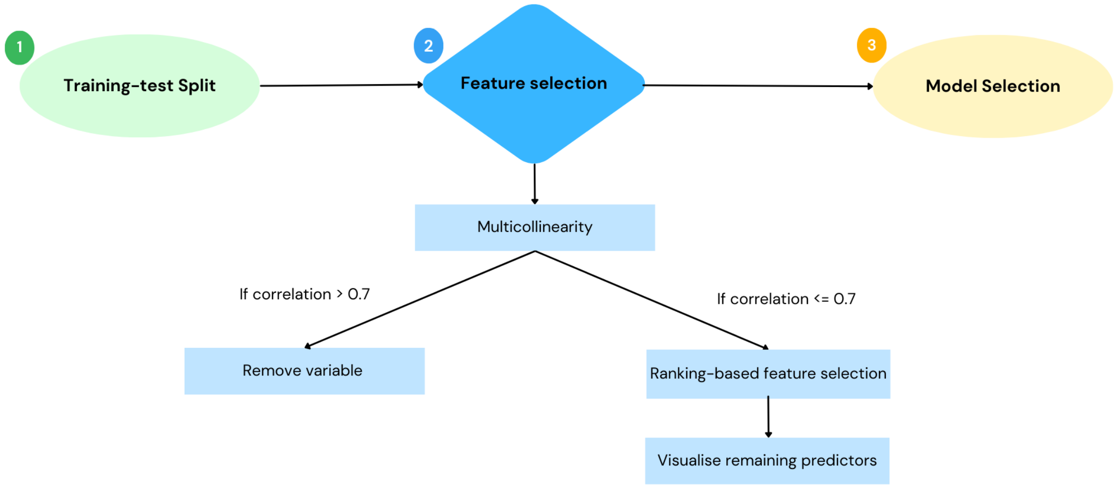


Figure 3: Feature Selection Flowchart

4.1 Training-Test Split

The remaining dataset is divided into a training-test set, using a common split of 70-30. The training set is used to train the machine learning model. Meanwhile, the test set is used for performance evaluation of models. This serves as an unbiased measure of how well the model can generalize to new, unseen data. Moreover, k-fold cross validation is also utilized to further assess models' goodness-of-fit.

4.2 Multicollinearity

Multicollinearity - high correlation among predictor variables - may result in model overfitting and reducing model performance. Therefore, it is necessary to remove multicollinearity. First, a correlation matrix of the predictors is calculated to determine pairs of predictors which are strongly correlated. Then, going through the matrix to remove highly correlated predictors. Two variables with correlation greater than the threshold 0.7 will be considered as generating high multicollinearity, and one will be removed from the dataset. After addressing this, there are 71 possible predictor variables left.

4.3 Ranking-based feature selection

Out of 71 remaining predictors, top 5 variables with the highest correlation with the target variable - Cobalamin (B12) - are chosen. This is because the highest correlated variables will have the highest explanatory power for vitamin B12. The correlations values of 5 remaining predictor variables for vitamin B12 are described below.

Table 2. Top-5 highest correlation with Cobalamin (B12)

Features	Correlation
C22: 5w3	0.466
Iron (Fe)	0.417
Iodine (I)	0.404
Copper (Cu)	0.390
C17	0.357

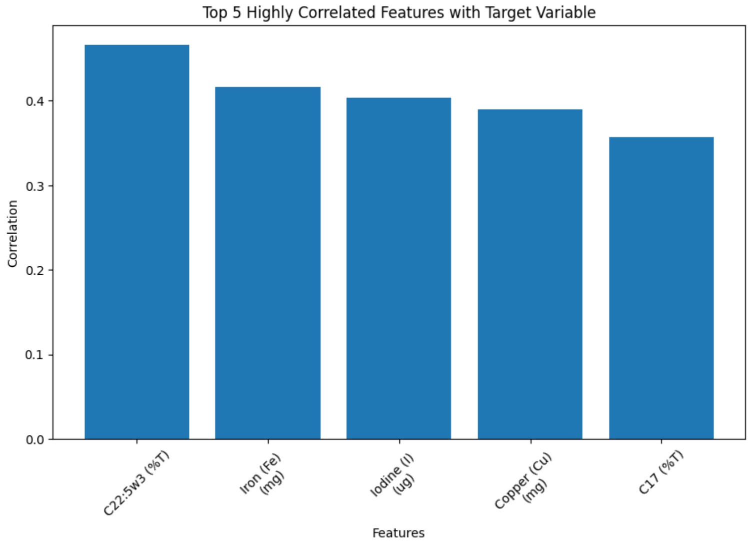


Figure 4: Correlation of Top-5 Most Correlated Variables with Cobalamin (B12)

4.4 Visualise Remaining Predictors

From the correlation matrix of the remaining predictors, high multicollinearity issues are under control, where all correlations are lower than 0.39.

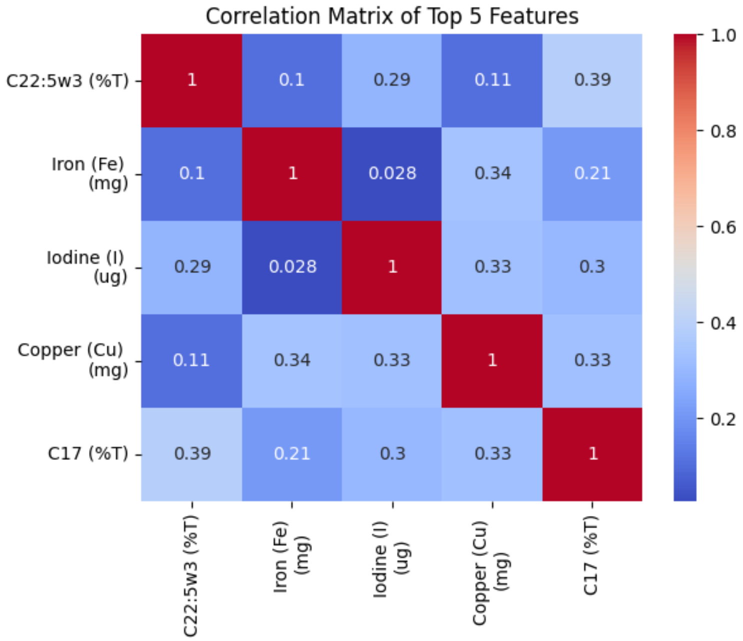


Figure 5: Correlation Matrix of Top-5 Most Correlated Variables with Cobalamin (B12)

Finally, in Figure 6, which displays a scatter plot matrix of the remaining predictors. The goal is to look at the predictors' diagonal distributions and see whether there is any substantial collinearity amongst them. Inspection reveals that the majority of predictors' distributions are skewed to the right. This finding could be explained by the mean imputation method used in the pre-processing phase. This skewness should be taken into account when interpreting the findings.

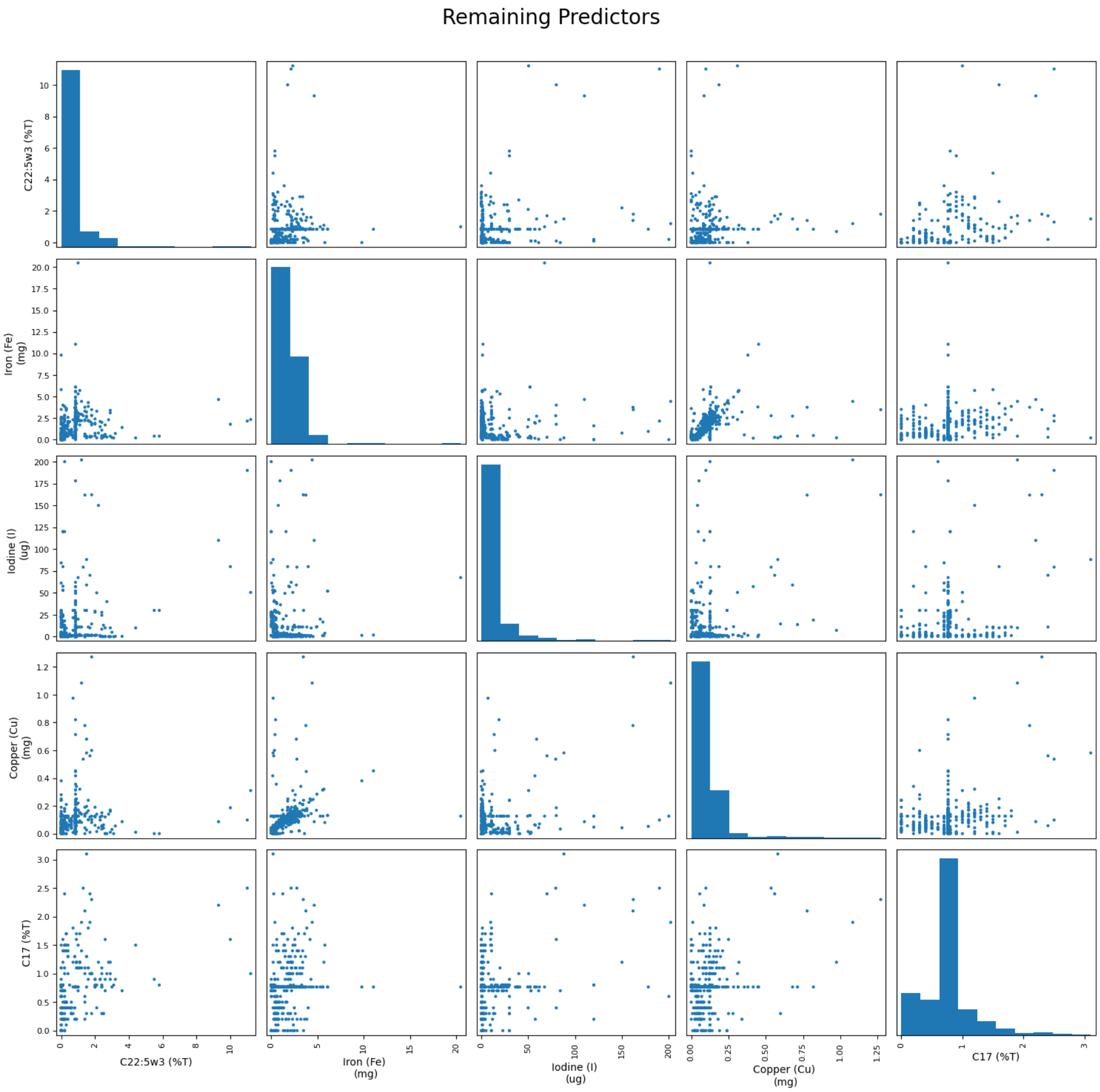


Figure 6: Remaining Predictor Scatter Plot Matrix

5 Analysis

5.1 Methods

In this project, as our target variable - the content of vitamin B12 - is a continuous variable, the appropriate supervised learning methods are Pearson correlation, linear regression, and random forest regression model.

Pearson correlation helps assess the linear relationship and determine the strength and direction of their association with vitamin B12 content. Linear regression can model the linear relationship between predictor variables and vitamin B12 content, while the random forest regression model can capture nonlinear relationships and interactions among variables, further enhancing the predictive accuracy. To evaluate the performance of our regression models, key metrics such as R-squared, RMSE, and 10-fold cross validation are examined.

5.2 Preliminary Analysis

Scatter plots between Vitamin B12 and independent variables are used to identify any pattern. Overall, there are not strong linear relationships between independent variables and the target variable. The high correlation is mainly driven by the food especially high in B12 content, and also those that are both high in B12 content and one or many of the independent variables. An irregularity in the scatter plot between C17 and vitamin B12 is the discretization of data. This is due to approximation in the original dataset, and has little effect on the analysis and training of the models.

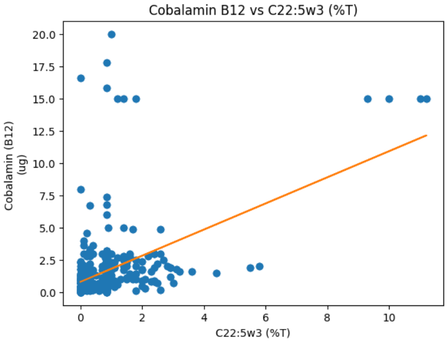


Figure 7: Cobalamin B12 vs C22:5w3

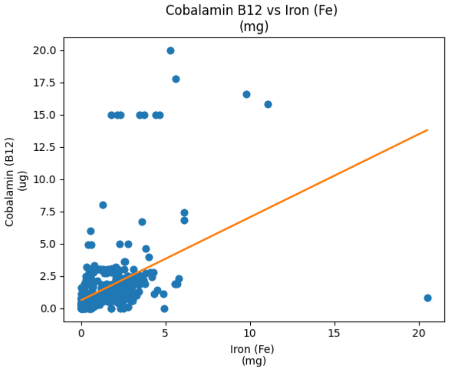


Figure 8: Cobalamin B12 vs Iron

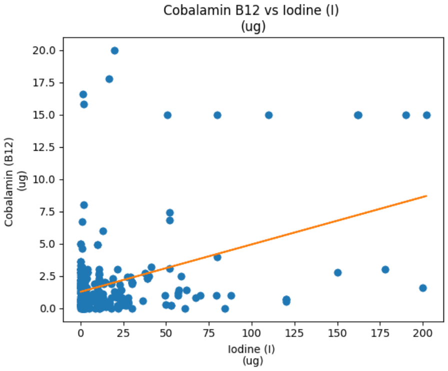


Figure 9: Cobalamin vs Iodine

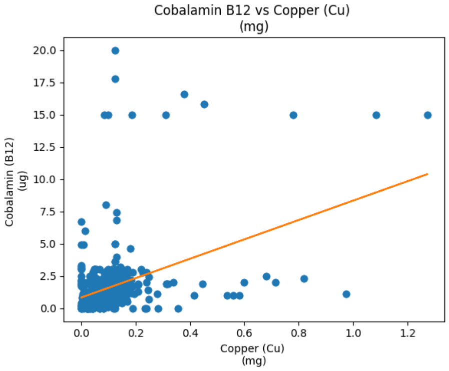


Figure 10: Cobalamin vs Copper

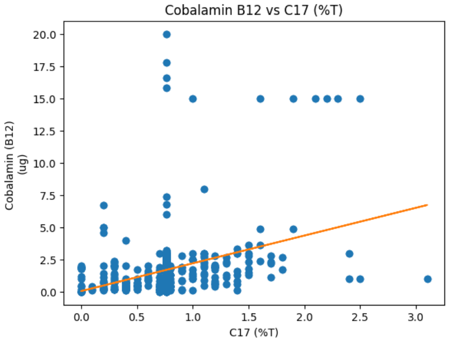


Figure 11: Cobalamin vs C17

5.3 Models

5.3.1 Model 1: Multivariate Linear Regression Model

To better understand whether the selected features had a meaningful relationship with Vitamin B12 content, a linear regression model is fitted:

$$Vitamin\_B12 = -0.5142 + 0.7245(C22:5w3) + 0.4837(Iron)$$
$$+0.0215(Iodine) + 3.0637(Copper) + 0.2408(C17)$$

The linear regression model returns an R^2 score of 0.497. The 10-fold CV mean RMSE value of the test set is 1.769, which is close to the RMSE in the original training-test split (1.551). Since this value is small compared to the range of values of vitamin B12 in the dataset, this model is relatively good at predicting vitamin B12 based on chosen features.

To evaluate the reliability of each model, this study looks at the linear regression scatter plot and residual plots:

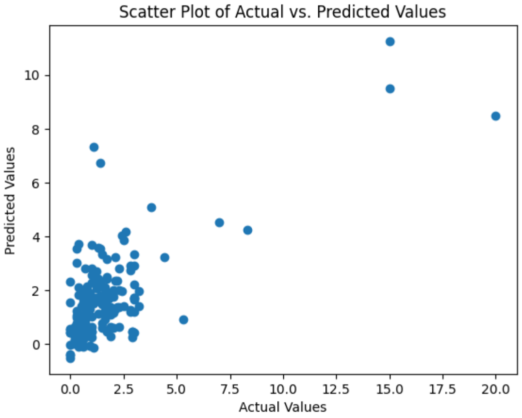


Figure 12: Actual vs Predicted Values Scatter

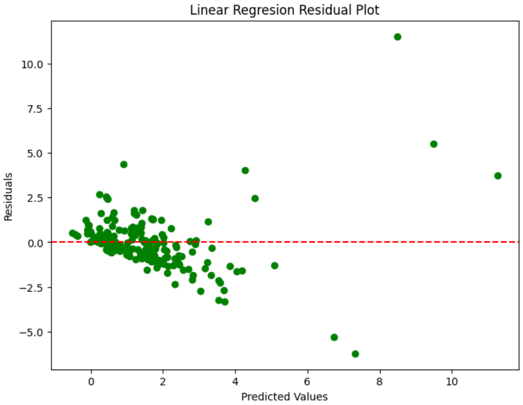


Figure 13: Linear Regression Residual Plot

The scatter plot of actual and predicted values follow the same findings. The data points in this plot cluster near the origin in a linear cone shape. The residual plot exhibits similar results, with a generally even scatter of points around the zero line for smaller values. However, at larger values, it becomes apparent that the variance of the error term is not constant, indicating the presence of heteroskedasticity.

5.3.2 Model 2: Random forest regression model

Since the relationship between vitamin B12 content and other nutrients might not necessarily be linear, it is appropriate to introduce another model that might perform better on variables with non-linear relationships. The scikit library offers a readily available and easy-to-implement solution in the form of a decision tree model. A decision tree model builds a tree-like flowchart to make predictions based on a series of decisions or rules. On each of its nodes, the tree splits into multiple branches based on some criteria on selected features, which splits the dataset into subsets accordingly. This study chose the random forest regressor model, which consists of an ensemble of multiple decision trees. This model has the benefit of having the same implementation in Python, while returning better results since more trees help to reduce the problem of bias.

This model returns an R^2 score of 0.45 and an RMSE of 1.62. The k-fold cross validation returns a similar RMSE score when taking the average from all folds (1.8). However, the difference lies in the fact that when taking the median, the score improves by a lot (1.18). When performing 10-fold cross validation, the model gets more training data (90% compared to 70% of the original dataset) hence it is expected to perform better. However, there is great variance among all the results, with the maximum RMSE going up to 7.2. This is partly due to the fact that the dataset has a bimodal distribution, with some foods being especially rich in vitamin B12. This could lead to the model being biased towards the technique of splitting training and test sets. A possible second reason is that decision trees and random forest models are notoriously bad at extrapolation, hence given a training dataset containing mostly foods low in vitamin B12, they would have a difficult time.

6 Discussion

6.1 Conclusion

This project developed models to predict vitamin B12 content in animal-based products based on nutritional data. The findings have implications for laboratory scientists to obtain an estimate of the vitamin B12 content based on the level of 5 different variables: C22:5w3, Iron, Iodine, Copper, and C17. To conclude, this contributes to nutritional science by enhancing the understanding of factors influencing vitamin B12 levels in food. These models can be improved and expanded upon in future study to increase their accuracy and adaptability to a wider range of dietary sources.

6.2 Evaluation of Models

Table 3. Model Goodness-of-Fit

Model	R^2	RMSE	Mean 10-fold CV RMSE
Linear Regression	0.497	1.551	1.769
Random Tree Regressor	0.451	1.622	1.800

From the table, **the linear regression model performs slightly better than the random tree regression model**. This overperformance could be attributed to the way we selected features, as features with higher correlation with the target variable are selected, and correlation is linear. Overall, both models perform relatively well, given that the task is to predict a nutrient that is not so common in food, resulting in low RMSE relative to the range of values for vitamin B12.

6.3 Limitation and Improvement

6.3.1 Bias in mean imputation

Mean imputation is a common method for handling missing data. While mean imputation is a simple and straightforward approach, it has certain limitations, one of which is the introduction of bias into the imputed dataset. A way to improve this is to perform sensitivity analyses and comparing the results obtained from different imputation methods can provide insights into the robustness of the findings.

6.3.2 Linearity assumption

When using continuous data, the linear model assumes a linear relationship between the predictor variables and the target variable. However, if the relationship is non-linear, the linear regression model could be unable to fully represent the underlying pattern, which could result in inaccurate estimations or bad predictions. To address this, a non-linear regression model has been employed to estimate the relationship between vitamin B12 and the predictor variables.

6.3.3 Heteroskedasticity

Heteroskedasticity - when variance of the error term is not constant across all levels of independent variables - can affect the statistical significance of the estimated coefficients and undermine the reliability of the regression model. Methods to address heteroskedasticity include transforming variables, and using robust standard errors.

7 Reference

1) Australian Food Composition Database - Release 2. (2022, January).
https://www.foodstandards.gov.au/science/monitoringnutrients/afcd/Pages/downloadableexcelfiles.aspx

2) U.S. Department of Health and Human Services. (2022, December 22). Office of dietary supplements - vitamin B12. NIH Office of Dietary Supplements. https://ods.od.nih.gov/factsheets/VitaminB12-HealthProfessional/

8 Appendix

Appendix 1: Outlier Visual Plot

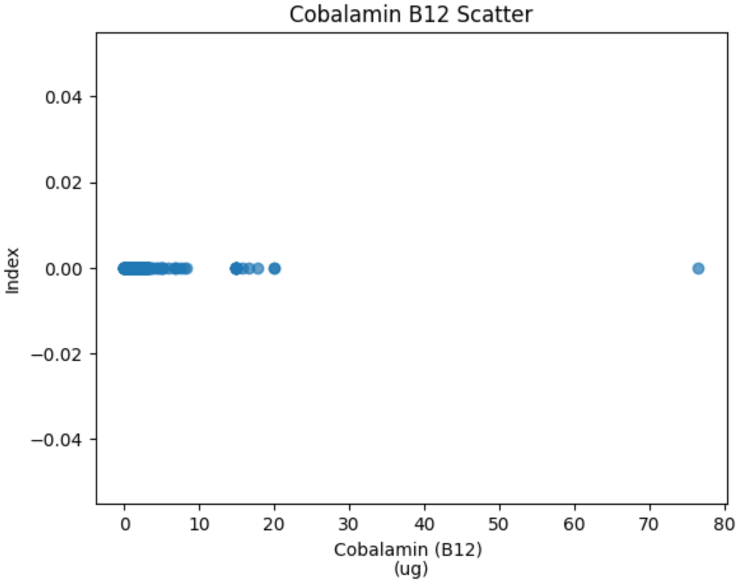


Figure 14: Outlier Scatter Plot

Appendix 2: Bar chart of top-5 correlated variables with Cobalamin B12

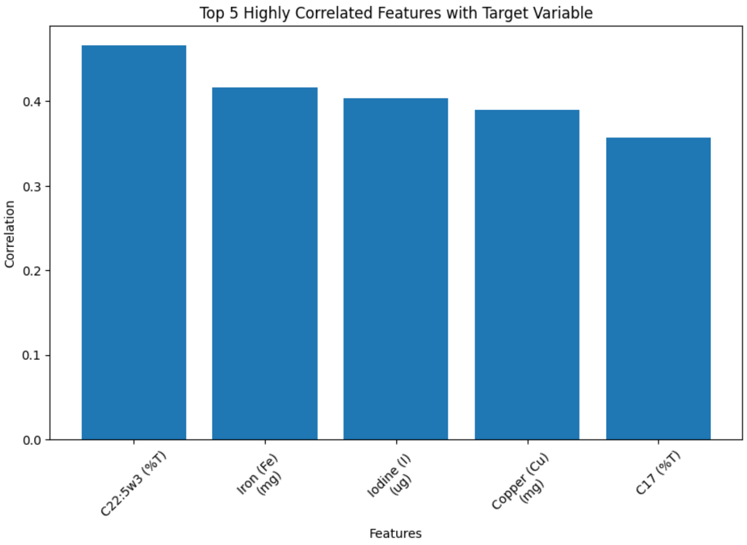
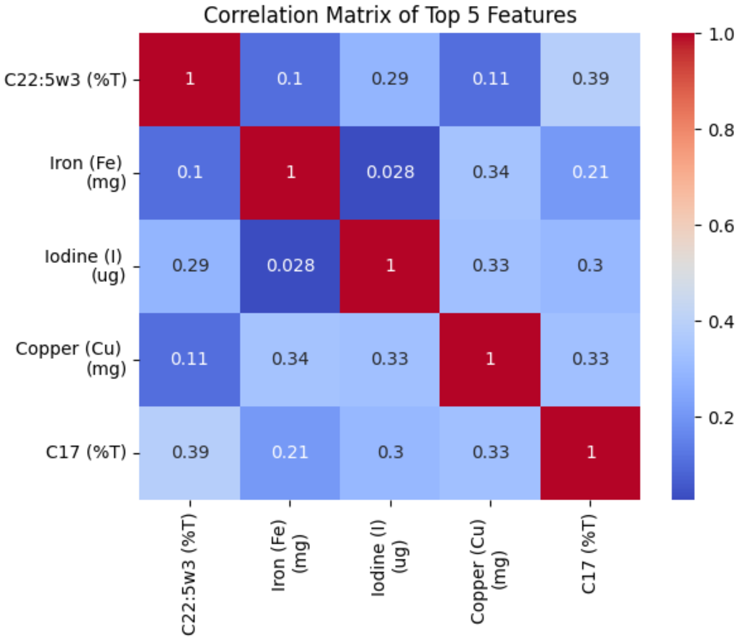


Figure 15: Correlation of Top-5 Most Correlated Variables with Cobalamin (B12)

Appendix 3: Correlation Matrix of top 5 most correlated variables



	C22:5w3 (%T)	Iron (Fe) (mg)	Iodine (I) (ug)	Copper (Cu) (mg)	C17 (%T)
C22:5w3 (%T)	1	0.1	0.29	0.11	0.39
Iron (Fe) (mg)	0.1	1	0.028	0.34	0.21
Iodine (I) (ug)	0.29	0.028	1	0.33	0.3
Copper (Cu) (mg)	0.11	0.34	0.33	1	0.33
C17 (%T)	0.39	0.21	0.3	0.33	1

Figure 16: Correlation Matrix of Top-5 Most Correlated Variables with Cobalamin (B12)

Appendix 4: Plot of remaining predictors after selection

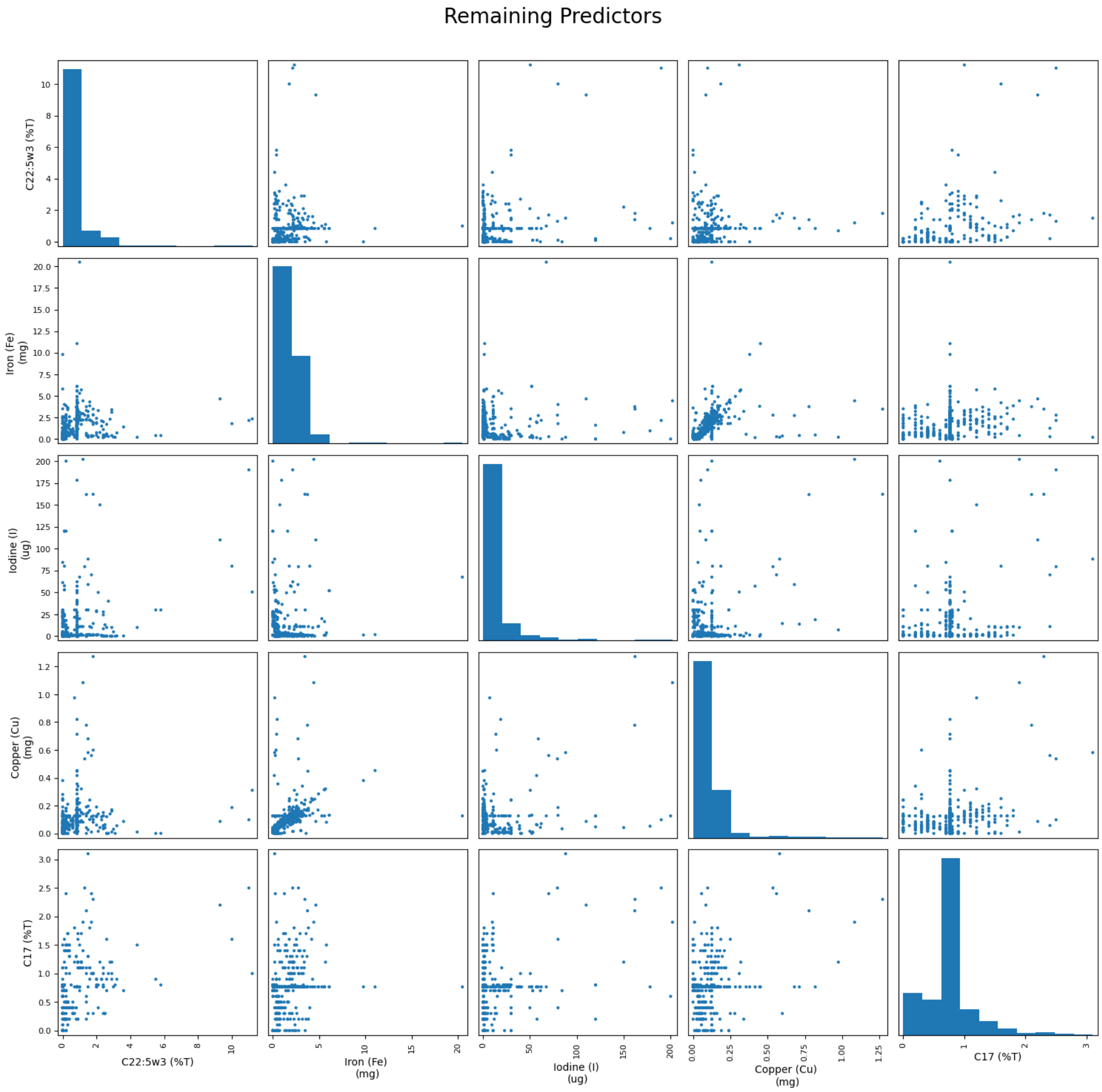


Figure 17: Remaining Predictor Scatter Plot Matrix

Appendix 5: Plot of Cobalamin vs all other selected features

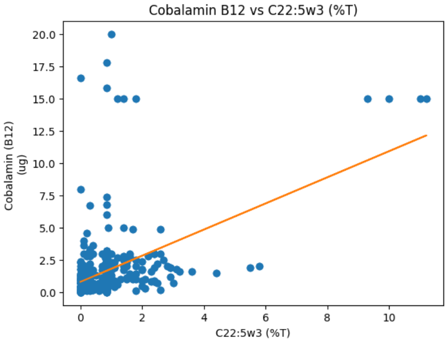


Figure 18: Cobalamin B12 vs C22:5w3

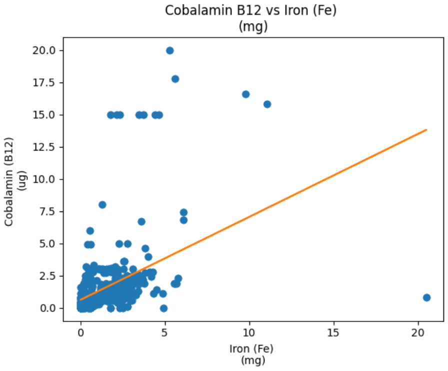


Figure 19: Cobalamin B12 vs Iron

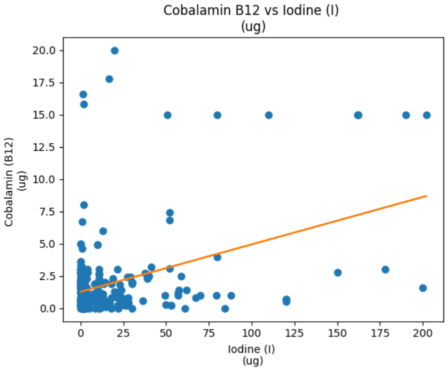


Figure 20: Cobalamin vs Iodine

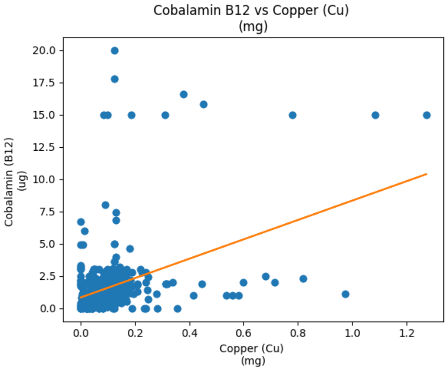


Figure 21: Cobalamin vs Copper

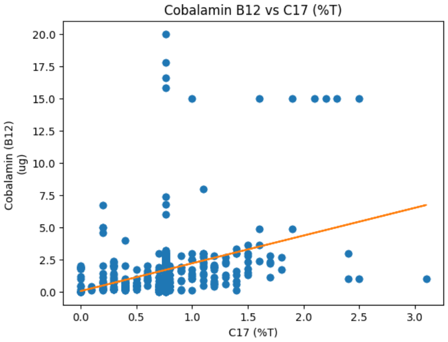


Figure 22: Cobalamin vs C17

Appendix 6: Plot of Actual vs Predicted and Linear Regression Residual Plot

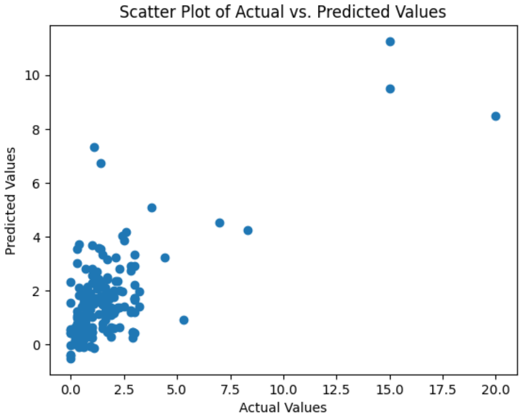


Figure 23: Actual vs Predicted Values Scatter

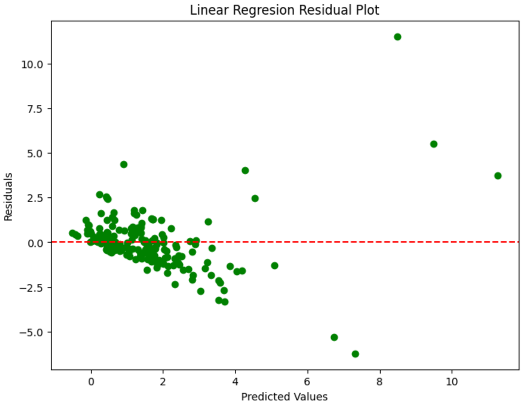


Figure 24: Linear Regression Residual Plot