Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Predicting Plasma Vitamin C Using Machine Learning

Daniel Kirk, Cagatay Catal & Bedir Tekinerdogan

To cite this article: Daniel Kirk, Cagatay Catal & Bedir Tekinerdogan (2022) Predicting Plasma Vitamin C Using Machine Learning, Applied Artificial Intelligence, 36:1, 2042924, DOI: 10.1080/08839514.2022.2042924

To link to this article: https://doi.org/10.1080/08839514.2022.2042924

Published online: 24 Feb 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2042924 (2481 pages)
https://doi.org/10.1080/08839514.2022.2042924

Predicting Plasma Vitamin C Using Machine Learning

Daniel Kirk [a], Cagatay Catal [b], and Bedir Tekinerdogan [a]

[a]Information Technology Group, Wageningen University & Research, Wageningen, The Netherlands; [b]Department of Computer Science and Engineering, Qatar University, Doha, Qatar

ABSTRACT

Precision Nutrition makes use of personal information about individuals to produce nutritional recommendations that have more utility than general population level recommendations. In many cases, being able to predict current status is a necessary first step in offering tailored nutritional advice. The objective of this study is to predict plasma vitamin C using machine learning. The NHANES dataset was used to predict plasma vitamin C in a cohort of 2952 American adults using regression algorithms and clustering in a way that a hypothetical health application might. Variables were selected based on a known or hypothesized relationship with plasma vitamin C, and variables that are expensive or difficult to obtain were excluded in order to more closely replicate the situation of a real health application. The best performance was seen with the XGBoost regressor, with random forest performing almost identically. Clustering was also investigated as a means of improving regression accuracy by splitting the data up into smaller yet more homogeneous groups, however, this was not successful. The low R-squared scores obtained by the models are likely to be due to the low resolution of the NHANES data, particularly the dietary data. This emphasizes the need for high-quality data sets in Precision Nutrition research.

ARTICLE HISTORY
Received 26 November 2021
Revised 3 February 2022
Accepted 11 February 2022

Introduction

Precision Nutrition (PN) is centered around the idea that nutritional requirements are not the same throughout the population, and that personal factors govern requirements for dietary components such as vitamins, minerals, calories, etc. These personal factors can include biological components such as genetics, nutrigenomics, and the microbiome, but also lifestyle factors such as diet, activity, sleep, stress, and more (de Toro-martín et al. 2017; Kirk, Catal, and Tekinerdogan 2021; Ordovas et al. 2018). The collection of such data – and likely, the integration of this data into a model – will then ultimately lead to the generation of dietary recommendations that have more utility and relevance to the individual than recommendations on the whole population level. However, given the many areas of health and disease that nutrition

CONTACT Cagatay Catal ccatal@qu.edu.qa Department of Computer Science and Engineering, Qatar University, Doha, Qatar

This article has been republished with minor changes. These changes do not impact the academic content of the article.

relates to, it is uncertain whether something will be designed that can make nutritional recommendations to optimize health as a whole, especially given that nutritional recommendations may be conflicting based on different areas of health. Instead, research so far has focused on specific areas of health, such as postprandial blood-glucose response (Zeevi et al. 2015), postprandial lipid response (Berry et al. 2020), bodyweight (Ramyaa et al. 2019), chronic disease management (Baek et al. 2019; Kim and Chung 2020), and cancer (Shiao et al. 2018).

For many dietary components, being able to estimate the nutritional status of an individual with regards to a dietary component is a necessary first step in providing tailored dietary recommendations. That is, for some nutritional recommendations to have value, they must first take into account the current status of an individual. This represents a prediction task whereby personal information is used to predict the nutritional status of an individual with regard to a given dietary component. After predicting current nutritional status, nutritional recommendations can be made with more certainty. One such dietary component that is suitable for PN is vitamin C. Vitamin C is an essential vitamin and thus must be consumed through the diet or supplementation (NIH 2020). However, diet is not the only contributor to vitamin C status, since smoking, gender, age, genetics, and more all impact the vitamin C status of an individual (Carr and Rowe 2020). Thus, following a national guideline is arbitrary and cannot be expected to lead to adequate nutritional status in many individuals. Instead, combining personal information that considers multiple relevant factors could lead to accurate prediction of vitamin C status, which can then be used to make nutritional recommendations suitable for the needs of the individual. For example, if status was predicted as low in one individual, then recommendations would focus on suggesting vitamin C rich foods for consumption. Otherwise, the suggestion may be that the individual does not have to invest too much energy in obtaining vitamin C and can instead focus on other dietary components.

Machine learning (ML) is a sub-branch of artificial intelligence and has enabled processing of data to support smart decision making. By learning patterns in training data, an algorithm is able to predict an outcome in classification, regression, prediction, and clustering tasks (Rowe 2019). However, this can be done in a way that facilitates the integration of data in large amounts and to a high degree of complexity, facilitating analyses that would not otherwise be possible. ML can be applied in a supervised setting, where algorithms are trained on known labels, or in an unsupervised manner, where similarities between data points form the basis of the output of ML algorithms; both have utility in PN. Thus, developments in ML are making headway in previously inaccessible problems. In the realm of PN specifically, ML has great applicability. Data required for generating PN-related outcomes can be complex and large, however, using ML to generate predictions based on

patterns in the data means nutritional recommendations can be made with more accuracy than has ever been previously possible. Not all work in the field of PN has made use of ML, and whilst at times promising results have been found (Celis-Morales et al. 2017), approaches that cannot handle a large number of complex variables; data which contains variables that have had their dimensionality reduced; and datasets of a large volume, are naturally limited. Indeed, these advantages of ML also explain why it frequently outperforms traditional statistical methods such as basic linear regression or Bayesian inference on datasets of larger complexity (Bzdok, Altman, and Krzywinski 2018; Kirk, Catal, and Tekinerdogan 2021).

To investigate the applicability of ML in PN, the present study set out to use various models to predict plasma vitamin C based on dietary and lifestyle variables. Data was collected from the National Health and Nutrition Examination Survey (NHANES) of America because of its public availability and ease of access, the large and nationally representative sample size, and the fact that it contains a vast array of personal and dietary information, as well as actual laboratory-measured plasma values for a variety of compounds. The 2017–2018 NHANES cycle contains plasma vitamin C measurements for participants, and thus this cycle was selected. Various regression algorithms were used, with XGBoost proving to be most effective. Clustering – alone and in combination with principal component analysis (PCA) – was used in order to attempt to improve regression performance.

The following sections are organized as follows: The next section describes the methodology. Section 3 presents the results, Section 4 the discussion andSection 5 the conclusion.

Methods

The goal of the current work was to use personal information from a publicly available dataset in order to predict plasma vitamin C, as may be performed on a digital health application service. Hence, in order to be representative of the true working nature of such an application, only information that is likely to be readily available from the user is used. Information obtainable through expensive or impractical methods of assessment are excluded. All analyses were performed in Python 3.7.6.

The Dataset

The NHANES 2017–2018 cycle was selected as the dataset for the current work (NHANES 2017). NHANES uses an anonymized nationally representative selection of the non-institutionalized American population to collect a wealth of data used to obtain information about the health status of the American population. The data is composed of information obtained from

physical examination and interviews. The physical examination is composed of both physical assessment and laboratory assessment. Information obtained from the interview includes demographic, socioeconomic and health-related questions. Within the interview section is an abundance of nutrition-related information such as information on food intake via two 24-hour dietary recall assessments and supplement use (NHANES 2017).

Although not a precision nutrition dataset, NHANES does contain personal information on every individual. Furthermore, corresponding laboratory data is available to provide ground truth labels for prediction, something which is not common in publicly available datasets. Hence, the NHANES data can be suitable for some work in the precision nutrition field by predicting health status or parameters related to health by using personal and lifestyle data. In this case, personal and dietary information is used to predict the level of plasma vitamin C, which is available in the 2017–2018 NHANES data cycle. The prediction of plasma vitamin C was chosen due to its established relationship with information available through NHANES such as dietary and lifestyle information and socioeconomic data (Carr and Rowe 2020).

Data Processing

The cohort of the NHANES 2017–2018 cycle that underwent plasma vitamin C investigation contains 6740 data points, which represents the maximum number of possible entries for the current study. As more variables are added the number of missing data also accumulates, causing this number to drop. Furthermore, in some of the questionnaire data, answers such as "Don't know" or "Refused" made up a very minor portion of the response; these participants, too, were removed. Finally, outliers in the target variable were removed based on a z-score of more than 3.25. This was decided not only from a statistical standpoint but also on plasma vitamin C ranges described in the literature (Hagel et al. 2018; Travica et al. 2019; NIH 2020; Kraemer, 2020). Of course, in such a large dataset, it is possible that these results do not represent anomalies and are possible plasma vitamin C readings, however by the same token such a large dataset increases the chances of reading or handling errors that could taint the results. For example, three individuals demonstrated plasma vitamin C levels of between 200–300 μmol/L. Indeed, this level appears extraordinarily high. Since model performance decreased if the outliers were included, the decision was made to exclude them. Future work could investigate if such results are to be expected, and, if so, how they can be incorporated into a model.

After all variables for the final models were added, a total of 2952 participants aged 18 years or older were used for analysis. The authors have no reason to believe that the missing data follows any pattern and that exclusion of these individuals would create a bias in the remaining

Table 1. The variables used in the model are shown, with their corresponding NHANES code.

Variable name	NHANES code	Categorical or Continuous	Notes
Gender	RIAGENDR	Categorical	
Race/Hispanic origin with non-Hispanic Asian	RIDRETH3	Categorical	
Annual Family Income	INDFMIN2	Categorical	
Vigorous work activity	PAQ605	Categorical	
Vigorous recreational activities	PAQ650	Categorical	
Smoked at least 100 cigarettes in life	SMQ020	Categorical	
Do you now smoke cigarettes?	SMQ040	Categorical	All missing values were assumed to be not currently smoking. This is because all those who answered "No" to the SMQ020 skipped question SMQ040.
Vitamin C (mg) (average daily supplementation)	DSQTVC	Continuous	To minimize data loss, all missing values were assumed to represent the absence of supplementation and thus were converted to 0 mg.
[Number] of people who live here smoke tobacco?	SMD460	Categorical	
Systolic: Blood pressure (first reading) mm Hg	BPXSY1	Continuous	
Diastolic: Blood pressure (first reading) mm Hg	BPXDI1	Continuous	
Which type of arthritis was it?	MCQ195	Categorical	As an inflammatory condition, information was obtained on participants rheumatoid arthritis. Other types of arthritis were excluded, and missing values were all assumed to be free of any arthritis (coded as 0).
Ever told had congestive heart failure	MCQ160B	Categorical	
Ever told you had coronary heart disease	MCQ160C	Categorical	
Ever told you had heart attack	MCQ160E	Categorical	
Ever told you had a stroke	MCQ160F	Categorical	
Ever told you had cancer or malignancy	MCQ220	Categorical	
Average daily dietary vitamin C intake	N/A (see Notes column)	Continuous	Average of two calculated dietary vitamin C intakes taken to provide average daily intake. This value is denoted "TVC" in the following figures.

data. The selection of variables was based on relationships to plasma vitamin C established in the literature or hypothesized relationships based on prior knowledge. These variables are presented in Table 1, where relevant processing notes are also mentioned. The only variable in Table 1 that does not exist in the NHANES data is the final variable, Average daily dietary vitamin C intake. The NHANES data offers total vitamin intake calculated from the 24 h dietary recall for each participant by corresponding the foods with vitamin levels in the USDA database. Because there are two 24 h recalls, there are two values for each vitamin. Hence, in order to attempt to provide a more reasonable estimate of dietary vitamin C intake, the average of the two values was taken and used as the total average vitamin C intake. In this way, the influence of one single day is halved.

Regression

A wide range of regression algorithms exists, each with different modes of operation and parameters that mean performance can vary across different conditions and datasets. Thus, it is common practice to use multiple algorithms and select the best performer. The data was a mix of continuous and categorical variables. Continuous variables can be used directly in the regression models, but categorical variables must first be processed. One-hot encoding was used to achieve this. One-hot encoding alters the data so that each entry of a categorical variable now becomes a new column entry, and ones and zeros represent the presence or absence of said categorical variables for each sample. R-squared (R-sq) was used as the principal criterion for regression model grading. The same input data and random state value of 7 were used across algorithms.

XGBoost

The first algorithm chosen for regression was XGBoost, whose name is short for "Extreme Gradient Boosting." As the name implies, XGBoost uses gradient boosting to provide more accurate predictions in classification and regression models. Gradient boosting is an ensemble method wherein the machine begins with weak learners that, with each iteration, are gradually improved upon. This process repeats until a given number of iterations are complete, at which point the model is finalized. In the case of XGBoost, the learners are decision trees, allowing competency in both classification and regression.

In addition to gradient boosting, XGBoost also has other features that improve its performance such as penalization of more complex models through regularization; internally dealing with sparse and missing data; hyperparameter tuning capabilities; built-in cross-fold validation; and tree pruning to prevent overfitting (Chen and Guestrin 2016). It is also extremely fast compared to other boosting and ensemble algorithms. Feature selection occurs internally to some degree but feature importance analysis is also readily available. Finally, it is easily deployed on many interfaces, and easy to use. For example, in comparison with other regression algorithms, fewer data processing steps are required. Collinearity need not be dealt with, since tree methods do not suffer from issues due to collinearity, and, besides, XGBoost also utilizes regularization which mitigates the issue of collinearity. Furthermore, other processing steps like normalization are not required, further reducing the workload for users of the algorithm. Given these advantages, it is unsurprising that XGBoost has boomed in popularity in recent years, demonstrated by its overwhelming use in online data science competitions and the fact it is generally regarded as a go-to algorithm in the data science world of today.

The model of the current work had hyperparameters tuned via grid search for performance optimization, resulting in 100 estimators, a learning rate of 0.1, a max depth of 2, an alpha regularization of 100, and a subsample of 0.8.

Random Forest

RF is an ensemble of decision trees that use bagging and then taking the mode (in the case of classification) or the mean (in the case of regression) of the collection of trees (i.e., the forest). Bagging is an abbreviation of the two words bootstrap aggregating, which describes this procedure: bootstrapping consists of taking many random subsamples of the dataset (with replacement) and aggregation refers to combining the predictions of many learning algorithms to allow better prediction than any single algorithm alone. Bootstrapping occurs first, after which a subset of variables is randomly selected for splitting root node, and then again, each node after that. The number of variables selected in this case represents one of the hyperparameters that can be modified in an RF model. With each bootstrapped sample, a portion is omitted from being used for training in order to estimate the performance – this is known as the out-of-bag sample. Furthermore, variables of importance are identified by observing the drop in error when a variable is randomly selected. Note, this drop in error is also averaged out across all trees in the forest. These properties make RF a top-performing predictive algorithm and, unlike decision trees upon which it is based, it is robust to overfitting. However, this added complexity increases computational demands and therefore processing time substantially (Breiman 2001).

In the current study, sklearn's RF regressor was utilized. Hyperparameters were tuned to give the following values: a max depth of 6, minimum samples at each leaf of 27, 1000 estimators, and a minimum number of samples required to split a node of 3.

Linear Regression

Linear regression from sklearn represents the simplest model used in the current study. Unlike the other algorithms used here, there are no hyperparameters available for modification. Linear regression takes a single variable (in the case of simple linear regression) or multiple (in the case of multiple regression) variables to predict the numerical value of a dependent variable. Each of the variables has a coefficient value, which maps their relationship with the dependent variable. In addition, there is an intercept value that maps the point on the Y-axis that the regression line crosses. Thus, multiple linear regression is defined as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_i X_i$$

where Y represents the dependent variable, β_0 the value of the intercept on the Y-axis and $\beta_i X_i$ the beta coefficient of the i^{th} feature in the model. To enable such simplicity, some assumptions are made in linear regression. First, the relationship between the coefficients and the outcome variable is assumed to be linear. Second, errors between the residuals (the observed values) and the predicted values are normally distributed. Thirdly, there is no multicollinearity between the variables used in the model. Finally, homoscedasticity is assumed. That is, there is no pattern in the distribution of the residuals versus the predicted. Violation of some of these assumptions may mean processing steps such as normalization or log transformation are required to be applied to the data before being used. Alternatively, violations of these assumptions may suggest a different model that does not have such prerequisites would be more fruitful. Typically, the quality of the model is measured by various methods that check the distance of the residuals from their predicted value. Larger values indicate the model is predicting the values more poorly (Olive 2017).

Lasso Regression

Lasso stands for "least absolute shrinkage and selection operator" regression, where "shrinkage" refers to shrinking data points around a central value, such as the mean. Lasso is a regularization technique, which means it makes use of a penalty that limits the strength of coefficients on the variables. This is L1 regularization, and coefficients can even be shrunk to 0, meaning they are effectively removed from the model. The term used to signify the strength of penalization in the model is λ, and thus with higher λ the greater the magnitude of the features is dampened, whereas with lower λ values the opposite is true, to the point where $\lambda = 0$ is the same as standard linear regression. The regression penalty λ is then multiplied by the absolute of the coefficient of the variables and this is added to the sum of the squared residuals (Tibshirani 1996). Hence, Lasso has the advantage of reducing the effects of multicollinearity as correlated variables are dropped by having their coefficients shrunk to 0. Reducing coefficients to zero also acts as an internal feature selection mechanism for Lasso, which makes it attractive to apply on datasets with many features. These characteristics of Lasso means it simplifies models, prevents overfitting, and potentially improves model performance over ordinary linear regression. Hyperparameter tuning in the current study was performed by testing model performance across a range of λ (denoted as "alpha" in sklearn) values. An optimal alpha value of 0.262 was found.

Ridge Regression

Lasso and Ridge regression are very similar in that they both penalize model complexity, reduce multicollinearity, and reduce overfitting, whilst potentially improving model accuracy. They have a key difference,

however. The regularization approach employed by Ridge regression is L2 regularization. Here, the penalty term of Ridge regression is λ multiplied by the squared magnitude of the coefficient, rather than the absolute magnitude (as with Lasso). The result of this is that coefficients can be shrunk ever closer to 0, but never quite reach (unlike with Lasso) (McDonald 2009). The optimal λ value for sklearn's Ridge regression used in the present study was found at 100, but a broad range of alpha values produced the same outcome.

Support Vector Regression

Support Vector Regression (SVR) is a regression option based on the algorithm support vector machines (SVM). SVM is most often used as a classification algorithm. In this instance, the machine attempts to classify data into groups by aiming to find an optimal split. The algorithm has an infinite number of possibilities in which to separate the data via decision boundaries but should choose the optimal one of these many possibilities. To achieve this, the algorithm focuses on the points in the classes closest to the decision boundary, allowing an optimal boundary to be chosen based on the distance of the points to the decision boundary. These points are known as the support vectors. To find support vector classifiers, kernels are used which can transform the data to a higher dimensional, allowing for better classification accuracy and application on non-linear data. These kernels exist in the form of linear, polynomial, or radial basis functions. In the case of SVR, the concept is similar but slightly different. Since there are no points to be classified, the decision boundary instead contains an accepted level of error known as epsilon (ϵ), which will extend ϵ and -ϵ above and below the decision boundary, respectively, forming a hyperplane. Points outside these margins are penalized, whereas points inside are not, and as with SVM for classification, the points on the decision boundary are the support vectors and have the most influence on the shape of the hyperplane. The function is then optimized to find the narrowest hyperplane (as determined by the support vectors on each side) and – like with other regression techniques – the line that minimizes the errors (Awad et al. 2015). These characteristics make SVM and SVR attractive for multiple reasons. For example, the ability to map non-linear problems is convenient because linearity is often not guaranteed in real-world datasets. Furthermore, only the subset of the data closest to the decision boundary is relevant for constructing the model, which greatly decreases processing time. Finally, the algorithm is flexible and highly effective in prediction using unseen data.

Optimization for SVR is less complicated than other algorithms but still has various hyperparameters available for manipulation. In the current study, the optimal performance was achieved with the following settings: a radial basis function kernel, with a C value of 10 and a gamma of 0.01.

Clustering

In order to attempt to improve prediction accuracy, we set out to find clusters in the data. Unlike the other algorithms described thus far, clustering is an unsupervised approach that groups together individual data points based on similar characteristics. In this way, data is broken down into *k* number of clusters, each of which is more homogenous than the data as a whole. As such, if machine learning algorithms are trained and applied within each cluster, they learn their function better and prediction accuracy may be improved over using the data as a whole, as was seen by (Ramyaa et al. 2019). Two clustering algorithms were used – k-means and k-prototypes.

k-means

k-means begins with *k* number of randomly placed starting centroids in the data and each data point is assigned to the cluster nearest to it. The centroids then relocate based on the mean of the points in each of the *k* clusters, and this process of assignment and relocation repeats iteratively until a stopping point is reached (Kanungo et al. 2002; Mannor et al. 2011). Euclidean distance is used to measure the distance between points (Kanungo et al. 2002).

scikit-learn's k-means was used and the data input was the same as the data used in the regression algorithms, i.e., continuous variables were log-transformed and categorical variables one-hot encoded. Silhouette score was used to identify both optimal number of clusters and quantify cluster purity. For each sample, the silhouette score calculates the average distance from one point to all other points in the same cluster, and then calculates the average distance from one point to all other points in the next nearest cluster. The distance between these values is then divided by the largest of the two values, providing a number between 0 and 1, with high numbers signifying clearly demarcated clusters and lower values indicating poorly differentiated clusters.

k-prototypes

Unlike k-means, k-prototypes permits the use of categorical variables as well as continuous variables. Thus, one-hot encoding was not required following log transformation of continuous variables. k-prototypes uses a mixture of Euclidean distance for continuous variables (like k-means) and a matching dissimilarity measure for categorical variables. The latter is the concept of k-modes, a clustering algorithm for categorical only data which used modes instead of means. Thus, k-prototypes is somewhat of a hybrid of the two, adding these distances to provide a value for similarity between sample points (Huang 1998). As with k-means, silhouette score across a range of cluster numbers was obtained to identify the optimal *k* value.

Principal Component Analysis

PCA was used experimentally with clustering in an attempt to improve results. PCA is a dimensionality reduction technique. In a dataset with n number of features, a maximum of n number of principal components are created and the variance captured by each principal component is known. First, the data is standardized, and a covariance matrix is made that provides information on how the variables differ from the mean to see if there are relationships between the variables. From this covariance matrix, eigenvectors, which provide directions of the axes containing the most information, and eigenvalues, which are the corresponding values on this new diagonal line, are calculated. In this way, new variables made in the previous steps are linear combinations of old variables. Ranking the eigenvectors in descending order of magnitude means the first eigenvector captures the largest possible variance; the second, the second most variance; and so on. These eigenvectors correspond to the principal components. In this way, PCA is designed so that the first principal component contains the most variance, and every principal component after that increasingly less until 100% of the variance is captured at the n^{th} principal component (Jollife and Cadima 2016). This process allows a large amount of variance to be captured with less total information by selecting a small number of principal components that capture most of the variance.

There are a few advantages to treating the data in this way. First, because the amount of information is reduced (sometimes substantially), processing time on proceeding operations using the new PCA dataset is reduced. Next, issues relating to collinearity are reduced. Since variables correlated together are captured in each principal component (because they convey information in a similar direction), principal components are uncorrelated. Third, redundant information is removed from the dataset. PCA is commonly used to capture the largest portion of the data for the smallest number of principal components. In some cases, this is much smaller than the number of features. Removing such redundant information could potentially enhance algorithm performance by reducing noise in the data. However, even if this is not the case, PCA may still be desirable as large enhancements in efficiency may be worth small decreases in accuracy. Finally, using only the first two components can allow the visualization of data with many features in a way that would otherwise not be possible.

In the present study, PCA was applied to the one-hot encoded data before clustering. Typically, clustering is performed alone for initialization, followed by PCA to reduce the dimensionality, and finally, clustering repeated again on the reduced dataset. This was the procedure followed in the present study. Notably, the data containing the new variables that are present after PCA are entirely numerical, which makes it incompatible with k-prototypes as k-prototypes requires at least one categorical variable. However, since continuous variables are treated in a k-means fashion in k-prototypes anyway, only k-means was performed after PCA transformation of the dataset.

Results

XGBoost

The full dataset contained 2952 samples and 16 features with a known or hypothesized potential relationship with vitamin C status. Following hyper-parameter tuning, the XGBoost regressor managed a cross-validated R-sq score of 0.303. XGBoost comes with an internal mechanism for allowing feature importance assessment. Visual analysis of this suggested that Annual Family Income (INDFMIN2) and Ever told you had cancer or malignancy (MCQ220) were not contributing to the model. Removal of Annual Family Income caused a small increase of R-Sq to 0.304 and removal of Ever told you had cancer or malignancy had no effect. It was investigated whether manual removal of other parameters also had the same effect on R-Sq, however, removal of other parameters caused model performance to worsen. Hence, a maximum cross-validated R-Sq of 0.304 was achieved with XGBoost regressor, as shown in Figure 1.

Figure 2 shows the importance (as weight)of each of the features in making the prediction according to the XGBoost regressor in the final model (that is, without INDFMIN2 and MCQ220). Continuous variables are represented unchanged, whereas categorical variables are represented by their name followed by an underscore and then the numerical value used in the original dataset which represents their categorical value, as occurs in the process of one-hot encoding. This is shown in the key below the figure. Unsurprisingly, average daily vitamin C intake (TVC) and average daily vitamin C supplementation (DSQTVC) compose the two most important features in predicting vitamin C. After this, both systolic and diastolic blood pressure are third and fifth, respectively. Fourth was gender as male. Smoking-related

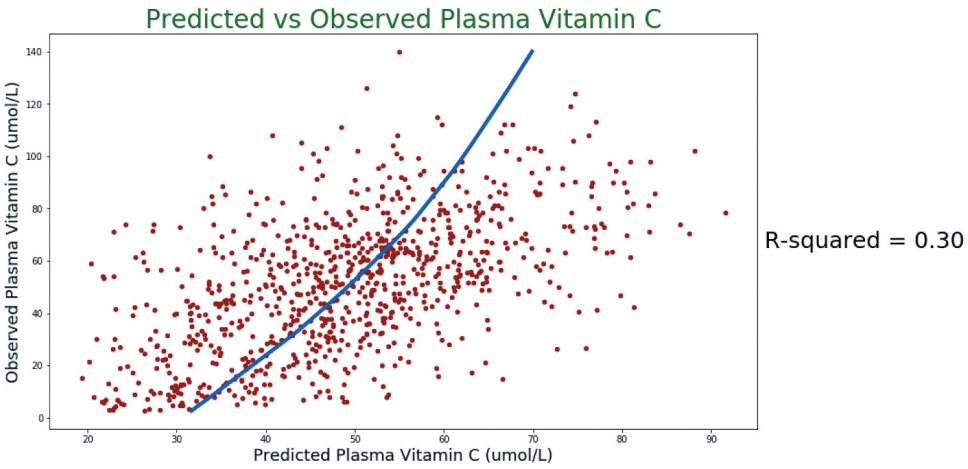
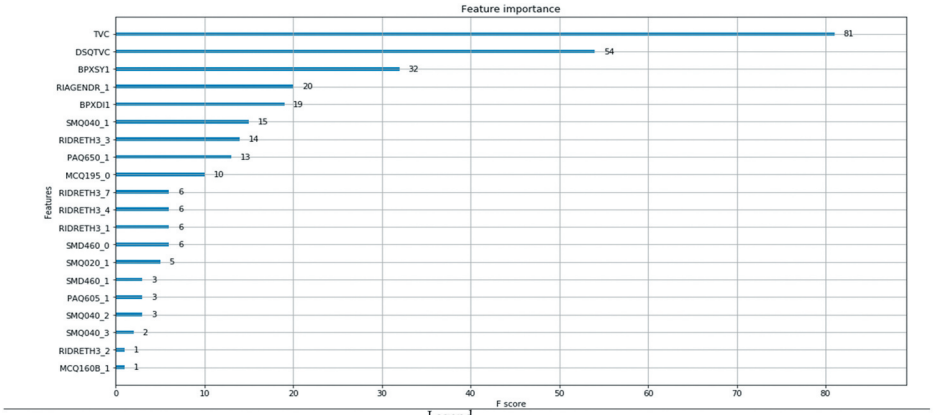


Figure 1. Actual plasma vitamin C is shown in comparison to the vitamin C predicted for the test split of the XGBoost model.



Legend

Label	Variable full name	Possible response
TVC	Average daily dietary vitamin C intake	Continuous
DSQTVC	Vitamin C (mg) (average daily supplementation)	Continuous
BPXSYS1	Systolic: Blood pressure (first reading) mm Hg	Continuous
RIAGENDR	Gender	1 - Male
		2 - Female
SMQ040	Do you now smoke cigarettes?	1 – Every day
		2 – Some days
		3 – Not at all
RIDRETH3	Race/Hispanic origin with non-Hispanic Asian	1 – Mexican American
		2 – Other Hispanic
		3 – Non-Hispanic White
		4 – Non-Hispanic Black
		6 - Non-Hispanic Asian
		7 – Other race – Including Multi-Racial
PAQ650	Vigorous recreational activities	1 – Yes
		2 – No

Label	Variable full name	Possible response
MCQ195	Which type of arthritis was it?	0 – No Arthritis
		1 - Osteoarthritis or degenerative arthritis
		2 - Rheumatoid arthritis
		3 - Psoriatic arthritis
		4 - Other
SMD460	# of people who live here smoke tobacco?	0 – No one in household is smoker
		1 – 1 household member is smoker
		2 – 2 household members are smokers
		3 – 3 or more household members are smokers
SMQ020	Smoked at least 100 cigarettes in life	1 - Yes
		2 – No
PAQ605	Vigorous work activity	1 - Yes
		2 - No
MCQ160B	Ever told had congestive heart failure	1 – Yes
		2 – No

Figure 2. Above is the importance of each of the one-hot encoded features in the XGBoost prediction of plasma vitamin C. Below is the key that explains the naming of the variables on the Y-axis.

features also demonstrate some importance, such as when the answer to SMQ040 ("Do you now smoke cigarettes?") was 1 ("Every day"). SMD460 ("# of people who live here smoke tobacco?") was the next highest smoking-related question and was concerned with the number of smokers in the home. SMQ020 ("Smoked at least 100 cigarettes in life") was after this and contributed to the model when the answer was 1 ("No"). Next was the feature race, followed by physical activity variables. Variables coded beginning with MCQ were medical conditions questions. These generally added little to the model.

Other Regression Algorithms

Other regression algorithms were also investigated for comparison with XGBoost. The data used was the same as the final input into the XGBoost model, and thus the pre-processing steps were the same. Unsurprisingly, RF also performed very well, with a score of 0.302, almost identical to XGBoost. For the non-ensemble methods, sklearn's linear regression model was used as a starting point, providing a cross-validated regression score of 0.258.

Subsequent analysis suggested that there was a high degree of covariance between the four continuous variables (one VIF score of 291.99, and two others exceeding 10). Stepwise removal of the variable with the largest VIF score was performed, however this decreased model performance. An alternative approach was to try linear regression algorithms that have capabilities to deal with covariance, namely Lasso regression. After optimizing the alpha value (the value responsible for modifying the strength of feature penalization), the results were basically unchanged (R-sq improvement of 0.04 to 0.262). Ridge regression was also performed but R-Sq did not exceed 0.260 after alpha optimization. Other hyperparameters on these models did not affect the outcome. The final algorithm investigated was SVM. After optimization, a maximum score of 0.294 was obtained using a radial basis function kernel. The results of all the regression algorithms can be seen in Figure 3.

Clustering

Clustering was performed in an attempt to improve regression results. The basis for this is that regression models trained on the individual clusters may be superior to models trained on the data as a whole. To achieve this, two clustering algorithms were used, individually and in combination with the dimensionality reduction technique PCA, which is often performed in tandem with clustering.

k-means

Silhouette score analysis of k-means clustering on the independent variables revealed 2 clusters to be the optimal number of clusters, evidenced clearly by the highest silhouette at $k = 2$ in Figure 4(a). The score was 0.787, indicating a high degree of quality of clustering. Thus, the labels of these two clusters were assigned to each of the participants in the data.

k-prototypes

As with k-means, silhouette score analysis was used to find the optimal k value, shown in Figure 4(b). Again, as with k-means, the optimal number of clusters was identified as 2, with silhouette score being highly similar (0.761) Regression was also performed on the two clusters identified with k-prototypes.

PCA

Following k-means clustering, it was also investigated whether PCA could improve clustering results further. Again, the one-hot encoded data was used for PCA analysis. First, for visualization purposes, only two components were used, the results of which are demonstrated in Figure 5. Next, all of the principal components and their respective and cumulative variance were

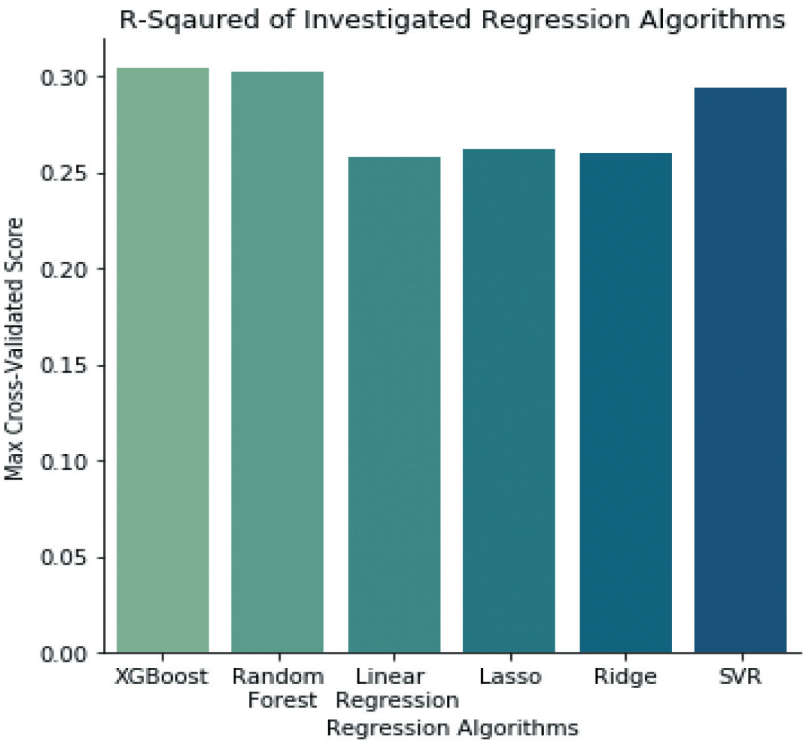


Figure 3. The cross-validated R-squared scores of all of the regression algorithms investigated in the present study.

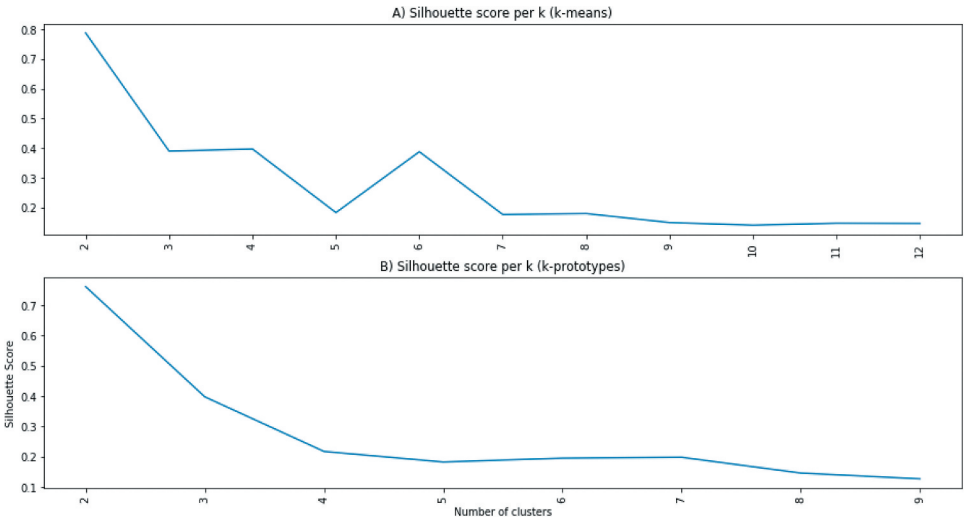


Figure 4. The silhouette scores at various values of k clusters for A) k-means and B) k-prototypes.

calculated. From this, a small number of components can be used to capture a large portion of the variance. This is presented graphically in Figure 6. It can

be seen that one component captures the vast majority (89.3%) of the variance, after which there is a small continual decline before a plateau. Since four principal components captured 95.42% of the variance, this was considered an appropriate number of components to use going forwards. Thus, PCA was repeated with only four principal components, and this PCA-transformed data containing four features was used again for k-means clustering. Again, the silhouette score was used to identify the optimal number of clusters with the PCA-transformed dataset, as seen in Figure 7. However, the maximum silhouette score is much lower here (0.500) than before PCA transformation, and thus the PCA-transformed data was not used for future analysis.

Regression on Clusters

Two clusters were identified in the data by the k-means and k-prototypes algorithm. Increasing homogeneity in the data may improve the predictive ability of the algorithm versus using the entire dataset as a whole. Thus, to investigate this possibility, the data was split according to these two clusters and the current best performing algorithm – XGBoost – was trained on the two clusters separately. In addition, because PCA analysis with 2 components for visualization (Figure 5) suggested four clusters may be present, the same analysis was also performed with $k = 4$. Cluster 0 contained 1034 participants and cluster 1 had 1918 of the participants. After splitting the data into the two clusters, XGBoost was trained and deployed separately, and hyperparameters were reoptimized with the following settings: a max depth of 2, a learning rate

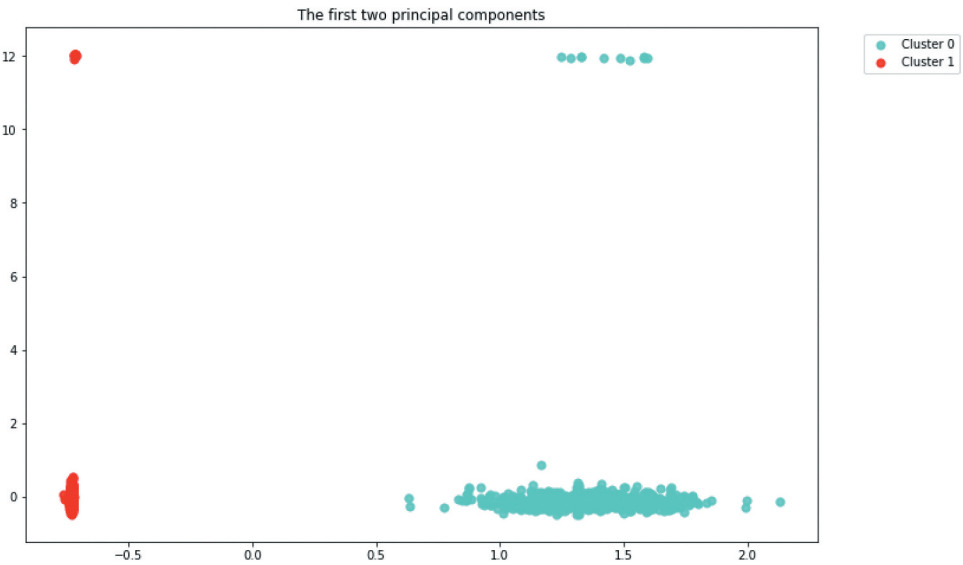


Figure 5. The first two principal components displayed following k-means clustering with 2 clusters.

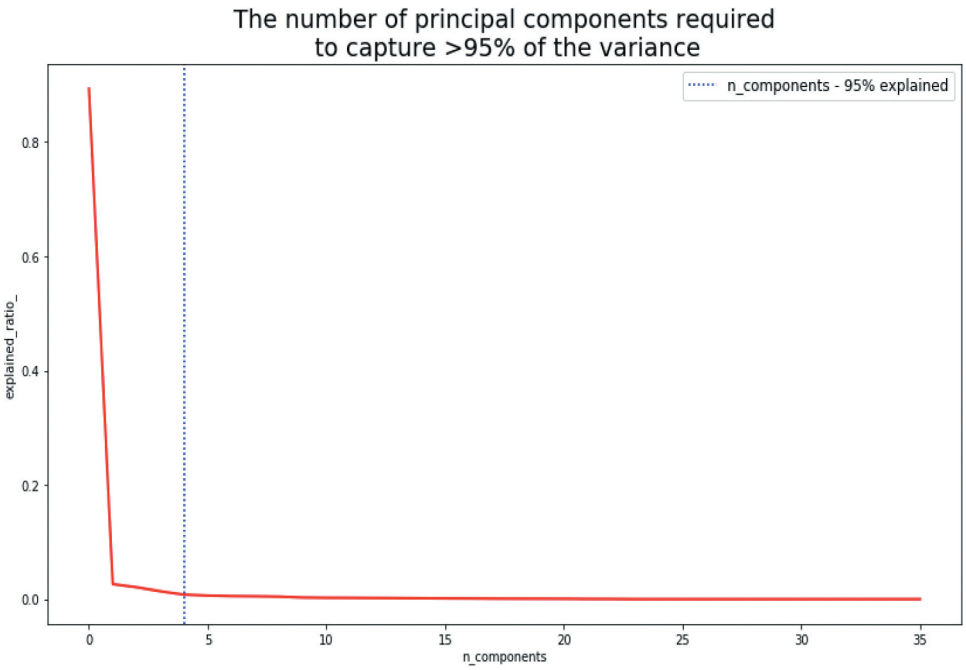


Figure 6. The number of components and the cumulative explained ratio is shown. The blue line represents the number of components required to capture >95% of the variance, which in this case is four.

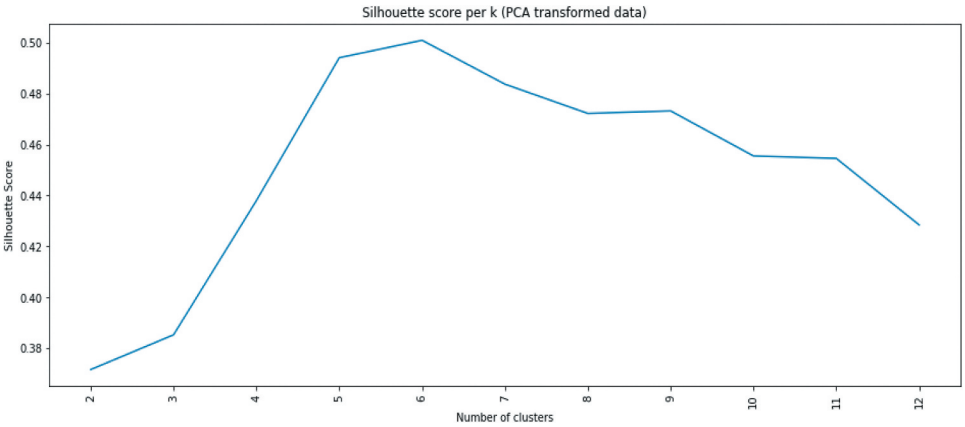


Figure 7. The PCA transformed dataset is used for clustering. The optimal number of clusters is seen as 6.

of 0.1, a subsample of 0.2, column sample by tree of 0.7, 80 estimators, and a regression alpha of 100 for cluster 0 and 200 for cluster 1. This led to R-sq values of 0.156 for cluster 0 and 0.206 for cluster 1, neither of which exceed the maximum R-sq score on the whole dataset. The same procedure was also applied following k-prototypes clustering however the clusters were the same length and R-sq almost identical to those following k-means. This suggests

that using k-means with one-hot encoding or using k-prototypes will give comparable results. Finally, repeating the analysis using four clusters instead of two caused a decrease in XGBoost performance across the clusters. In conclusion, NHANES data could not be used to predict plasma vitamin C with meaningful accuracy.

Discussion

The present study showed that NHANES data cannot be used to predict plasma levels of vitamin C in American adults, evidenced by a maximum R-sq value of only 0.304 using XGBoost regressor and variables likely to be readily available to users of a hypothetical health application. The growing field of PN aims to provide nutritional advice personalized to the individual instead of on a general level. One aspect of this area involves using information about the individual to approximate their current status of food substances such as vitamins and minerals, in order to make more precise recommendations and avoid malnutrition. Although vitamin C status cannot be entirely represented through plasma vitamin C due to peripheral placement, plasma levels can offer an approximation and thus can suggest whether an individual is in danger of insufficiency (NIH 2020). Besides this, plasma vitamin C is also more convenient to obtain and more widely measured than other metrics, meaning this data is more available. Finally, the relationship between plasma vitamin C and food intake may be more short term than in comparison to peripheral vitamin C, making it more suitable for measurement in the present study given the form of the dietary data available in NHANES (discussed below) (NIH 2020). These points formed the theoretical basis for the selection of plasma vitamin C over other food substances that data was also available on in NHANES. NHANES, however, is an epidemiological rather than a PN dataset. Therefore, the resolution of the data is low and instead the focus is on collecting data across a large number of people. This does not make it a good choice for the study of PN; however, it is attractive because it is free and publicly available. This is in contrast to other datasets that contain data more suitable for PN-orientated research, which are difficult to obtain. In most cases, contacting authors or formal application is required, which limits the accessibility of datasets to other researchers. These were barriers that could not be overcome in the timeframe of the current study, and therefore NHANES was selected, despite its shortcomings. Although data in PN research is witheld to respect the personal information that is collected, it also presents a barrier to the development of the field of PN. Accessible datasets such as NHANES suitable for the PN research would develop the field.

The field of PN is gaining traction because of the potential it has to revolutionize the health status of society by focusing on nutritional requirements on a personal level rather than on a population level, where

recommendations are based on large groups of people and thus relevance to the individual is diminished. This neglect of the personal factors that affect nutritional requirements partially explains the failure of existing approaches used to attempt to improve nutritional status (Cecil and Barton 2020). By using personal data to predict the status or requirements of an individual (or group of individuals, if on the stratified level) with regard to a nutritional component, targeted approaches can be deployed to prevent or amend deficiencies before health complications ensue. In this way, public health will be improved whilst treatment costs due to malnutrition-related complications are reduced. This scenario is very achievable considering that data is increasing in both abundance and accessibility and the tools for the processing and analysis of this data are ever-improving (Kirk, Catal, and Tekinerdogan 2021).

PN needs not only to work in theory but also in practice; in this regard, cost-effectiveness is relevant. In many cases, to generate results of importance for PN, one requires basic equipment that is considered widely available in the modern world, such as a computer, internet access, and an environment in which to execute ML experiments. This enables the generation of meaningful results using only publicly available data, or at least data that can be obtained easily and non-invasively. Indeed, the importance of easily accessible personal variables in PN approaches utilizing ML is also described elsewhere (Berry et al. 2020). Besides, even when high-resolution, expensive approaches are required in certain cases, this cost will likely be front-loaded rather constant; that is, once PN results are obtained, recommendations can be administered that can have a long-lasting impact. Whether or not the benefits – both financially and in terms of health gains – outweigh the costs should be investigated, but it seems unlikely that financial constraints will prevent PN from realizing its potential.

Whereas previous data analysis approaches have relied on traditional statistical methods to generate outcomes, they increasingly fall short in comparison to ML and other artificial intelligence approaches. This was also observed in the current study, in that, although no algorithm had sufficiently adequate performance, ML algorithms were consistently better than traditional approaches. ML and its subdivision deep learning can learn patterns in the data that would otherwise be inappreciable to their human operators and can do so on complex and large datasets. Using unsupervised learning, ML can also be helpful in the generation hypotheses, and reinforcement learning (another ML type) allows constant self-regulation in response to a dynamic environment, which will be relevant to PN in the context of the constant generation of data in real-time from devices such as wearables (Kirk, Catal, and Tekinerdogan 2021). These are aspects of PN where traditional statistical approaches are unable to compete with ML approaches. For a more detailed discussion, see the review of Kirk, Catel, and Tekinerdogan (Kirk, Catal, and Tekinerdogan 2021).

Although this study aimed to predict plasma vitamin C status, it is not surprising that this was not achieved given the dataset used. Firstly, diet is a major predictor of vitamin C status since vitamin C cannot be manufactured by the body itself (Carr and Rowe 2020), and hence demonstrates a moderate relationship with plasma status when assessed via food frequency questionnaire or dietary recall (Dehghan et al. 2007). NHANES collects data for the dietary portion of the survey using two 24-hour dietary recalls spaced between 3 and 10 days apart (NHANES 2015). This means that the information collected that led to the calculated average daily vitamin C intake (denoted "TVC" in the current study) was derived from only two days of food data. Even though plasma vitamin C correlates with short-term vitamin C intake (NIH 2020), this time frame may still be too short to get an accurate reflection of real vitamin C intake. This makes it likely that the assumed vitamin C dietary intake was not reflective of true daily intake in many of the participants, which would naturally reduce the correlation between intake and plasma levels, and therefore the predictive ability of the model. This could explain why the correlation of vitamin C intake and plasma vitamin C alone was lower in the present study (r = 0.29 on log-transformed TVC) than elsewhere (r = 0.46) (Dehghan et al. 2007). Besides these points, memory-based dietary self-assessment (as used by NHANES) has been criticized for being unreflective of true dietary intake (Archer, Pavela, and Lavie 2015), and this sentiment is echoed elsewhere, specifically with dietary intake and vitamin C (Dehghan et al. 2007). The resolution of data required for many aspects of PN is higher than tools like dietary recalls and food frequency questionnaires can provide. Thankfully, modern techniques such as dietary trackers on smart devices, image-based recognition of foods and barcode scanners are making higher resolution tracking possible (Kirk, Catal, and Tekinerdogan 2021). Thus, these developments may facilitate more accurate tracking of consumption and allow for stronger predictive models in the coming years.

One of the variables least likely to suffer from quality is that of blood pressure measurements since this is an objective measure. Although this data may not be immediately available to a hypothetical user of a health app, it is generally inexpensive and easy to obtain, and thus remained in the final model. Both systolic and diastolic blood pressure were shown to be moderate predictors of vitamin C. Vitamin C and blood pressure have previously been shown to be associated (Myint et al. 2011; Ness, Chee, and Elliott 1997), with randomized controlled trials showing a causal role for vitamin C influencing blood pressure rather than simply association due to a concurrent increase in fruit and vegetable consumption (Juraschek et al. 2012). Thus, it is logical that those with higher blood pressures could be expected to have lower levels of plasma vitamin C. However, it is likely in our case that such associations with other lifestyle habits (such as fruit and vegetable consumption) are masked by not only the blood pressure but also physical activity variables, which showed

a relationship with plasma vitamin C. Future work could look to tease these contributions apart, which could increase understanding and prediction accuracy. One approach to this could be investigating the consumption of individual foods, rather than the calculated vitamin C intake. This is discussed in more detail below in the strengths and limitations.

Race was also shown to be a moderate predictor of vitamin C in certain circumstances. As with blood pressure, race is a variable that encompasses other information. Clearly, race conveys genetic information about the individual, but ethnic groups also differ in other ways, such as socioeconomic class and dietary and lifestyle habits. Such information may be falsely augmenting the importance of race in plasma vitamin C prediction. Thus, it would be beneficial for future work to investigate further the true contribution of race, with the aforementioned factors also accounted for. Moreover, whilst there is a genetic influence on vitamin C status – and there is a known relationship between race and these genetic variants (Carr and Rowe 2020) – race would serve as a poor proxy as a genetic variable. With the development of genetic testing and increasing public access to such services, it is likely that health applications will be able to incorporate both to allow the information contribution from either. Some genetic information is available from NHANES, although this is restricted and was not available for use in the time frame of the project. Other variables such as gender are known to have a relationship with vitamin C status. Whilst this has been described to be related to body composition differences between men and women (Jungert and Neuhäuser-Berthold 2015), such information was not included in the present study as this information is less accessible. Health-related questions were included as a variety of disease states are associated with vitamin C deficiency, even after recovery (Carr and Rowe 2020); however, health variables generally had poor predictive value.

Annual family income (INDFMIN2) was used to approximate socioeconomic status, which has an established relationship with vitamin C intake (Carr and Rowe 2020; NIH 2020). However, this was dropped from the model since it worsened the predictive quality of the model. Even if annual family income cannot adequately approximate socioeconomic status, it is strange that this variable would negatively impact the model. An alternative variable offered by NHAENS, Annual household income (INDHHIN2), was also investigated but offered no contribution to the model. These artifacts may again reflect the quality of the NHANES data. Finally, it should also be considered that the addition of other variables present in the NHANES data could have improved the prediction accuracy. However, NHANES has a large number of variables and each requires time to process and integrate. Furthermore, each additional variable comes with the cost of more missing data. Thus, executive decisions were made on which variables to incorporate based on prior knowledge and assumed relationship with vitamin C.

The prediction of a continuous value (as with plasma vitamin C) is a regression problem. The problem could have also been viewed as a classification problem with different levels of adequacy being the label classes, but this was ultimately decided against and instead a value was predicted. This is advantageous because more information is conveyed (i.e., an exact number rather than a broad label), and because the absolute value may be relevant to know for general health, such as was described above with the relationship between plasma vitamin C and blood pressure (Juraschek et al. 2012). Thus, users may wish to know exact amounts. Besides, assigning class labels to a numeric predicted value is straightforward, meaning this information can be additionally conveyed to a hypothetical user. The models selected tend to perform well on regression. The popularity of XGBoost is clearly demonstrated by its domination in Kaggle data science contests in recent times. For reasons discussed in the methods section, XGBoost has various advantages that enhance its performance above other regression algorithms. Hence, it is unsurprising that XGBoost was also the best performer in the current study, although only very marginally so compared to RF. RF and XGBoost are different algorithms and perform differently on different datasets, however as a boosting algorithm XGBoost has the ability to improve on the mistakes that previous weak learners made earlier in the process to provide a strong final estimate (boosting). The trees in XGBoost are very short, sometimes so short they are known as decision stumps. RF instead uses bagging to improve estimation over using a single estimator (i.e., a decision tree) and does this in parallel as opposed to XGBoost which does this iteratively. The advantage of boosting and the range of hyperparameters available for tuning mean XGBoost can be expected to have a slight edge over RF. SVR – another algorithm with strong capabilities in regression problems – achieved a maximum R-sq value comparable with those of XGBoost and RF, and thus this R-sq value may represent an upper limit of predictive ability based on the quality of the data. This reemphasises the fact that NHANES has low applicability in such types of PN-orientated research as this. As a baseline algorithm, multiple linear regression was chosen, which had the lowest performance. In an attempt to account for some of the covariance between continuous variables, Lasso and Ridge regression were also employed, which led to minor improvements. This could be owed to the fact that the majority of the data were categorical in nature.

Following regression, clustering was applied to the data. As an unsupervised method, clustering has the potential to group the data in a way that may not be intuitive to the human eye but ultimately makes groups that are more homogenous than the data as a whole. Doing so can have certain advantages. First, this is an example of stratification, which is a level of personalization in PN. Identifying the presence of groups in work such as this may lead to the identification of groups in the population that share characteristics that may

have an impact on health and disease. Thus, such groups – if relevant to vitamin C – may also be relevant to other areas of health and may have other nutritional requirements. Secondly, making groups more homogenous could improve the prediction accuracy of the model by reducing noise and allowing for better training, as was observed by Ramyaa et al. when predicting body-weight from macronutrient and exercise variables (Ramyaa et al. 2019). Thus, in ML, techniques such as clustering can provide the opportunity to improve on results or identify patterns that would otherwise go unnoticed. Silhouette score analysis (Figure 4) revealed two clusters ($k = 2$) with k-means provided the best clustering results. However, when these clusters were used to train the XGBoost model, results showed a significant decrease. The most logical explanation for this is that, although two clusters could be clearly defined, they had no relevance to the ultimate status of plasma vitamin C. Thus, the large decrease in n from 2952 (with the whole dataset) to 1034 and 1918 for cluster 0 and cluster 1, respectively, reduced instances from which the model could learn from. In support of this, cluster 0 (with fewer participants) had the lower score of the two.

A strength of the study includes the fact that NHANES is designed as a nationally representative sample of the American population. This means that results obtained should apply, theoretically, to the American adult population as a whole, and not just select subsets or groups. Having said that, the results using such a sample were not strong, and this reinforces the idea that PN research would benefit more from using smaller sample sizes but with higher resolution data than large sample sizes such as NHANES. However, having such a large sample size is also a strength in itself. Additionally, this large sample size allowed the removal of participants with missing data and thus meant imputation techniques could be avoided. Since there is no reason to believe the missing data followed any kind of pattern, such data removal can be considered consequence-free, as opposed to imputation which inherently makes assumptions on the data. The machine learning techniques applied to the data can also be considered some of the best currently available. Over recent years, RF has been dominant in performance on classification and regression tasks. Recently, XGBoost has taken a slight edge over RF and generally performs better, if not equally as good, on these tasks. SVR/SVM is also a high-quality algorithm. A final strength of the current work is that all results were averaged cross-validated results. Using cross-validation to evaluate results provides a more honest representation of the performance of the model by ensuring all data is used for training and testing. This prevents getting unusually high or low test scores simply due to the randomness of how the data was split.

There are also points for improvement. As discussed above, the main concern of the current study is that the NHANES dataset is designed as an epidemiological dataset and not a PN dataset. Thus, the quality and resolution of the data is

naturally lower than that required to produce meaningful predictions. For example, two leading papers in the area of PN are (Berry et al. 2020) and (Zeevi et al. 2015) who both predicted post-prandial glucose response. Although this presents a different challenge and requires different input variables, the resolution of the data was much higher. That is, data was collected more meticulously and in more detail on every individual, facilitating a high prediction accuracy. This type of high-resolution data will certainly be required to produce meaningful outcomes in the coming years of PN, though with this comes increased costs. In such cases, a cost-benefit trade-off becomes apparent, especially outside of the research setting and in situations of practical application, such as the clinic. Whether or not the benefits of high-resolution personalized approaches outweigh the costs is currently not known and will likely differ on a case-to-case basis within the various domains of health where PN approaches are applied. However, it is encouraging to know that not all PN approaches require high-resolution and expensive techniques to derive meaningful results (Kirk, Catal, and Tekinerdogan 2021).

The data being derived from a US population (i.e., NHANES) naturally restricts the generalizability of the results. Indeed, in populations outside of the US it is reasonable to expect that prediction accuracy would be even lower due to differences in personal and dietary data. Rather than aiming to develop data of high generalizability, however, as may be preferred in epidemiological approaches, work in the field of PN should look to use datasets relevant to the population in question. The NHANES has the advantages of being large in size, breadth of data collected, and of being accessible, meaning that researchers are not required to collect such data themselves. If such publicly available datasets could be utilized to generate relevant PN recommendations this would be beneficial for clinical practice by facilitating tailored recommendations for patients based on personal data that would be either already available or easily obtainable to the clinic. Future work should aim to find higher prediction accuracy using such data as it circumvents the need for variables obtained via potentially expensive and invasive means.

There may also be room for improvement in data processing. Log transformation of continuous variables was performed, and this improved results. Normalization and standardization were also investigated – both before and after log transformation – but none of these four combinations improved results. Regardless, it is possible that other data transformation techniques could have been more fruitful. Lastly, the present study used the calculated vitamin C intake for each participant provided by NHANES, which was then averaged across the two days of food data (labeled as TVC). However, it would be of interest to see if better accuracy could be obtained using the individual foods consumed instead of the calculated vitamin C amount. This should, theoretically, convey the same information on vitamin C consumption as using a calculated total, and, in fact, could be more accurate, since this circumvents points of error (e.g., measuring

error with measuring apparatus) in the processes that lead to these values in the United States Department of Agriculture (USDA) database from which the values are derived. Moreover, using individual foods could allow the capture of interaction effects between foods that may enhance or reduce vitamin C absorption. This information would not be captured using calculated vitamin C intake, as used in this study. This approach was trailed, though led to poor performance presumably because some foods were only consumed by a handful of people, ultimately meaning the model was low on training instances and could not learn the task properly. Remedies of this using participants with fixed or restricted diets may be of interest for future work.

Conclusion

In conclusion, using data from the NHANES dataset likely to be readily available from a hypothetical health app user, plasma vitamin C was predicted with a maximum accuracy of R-sq equal to 0.3. Whilst this is evidence of some predictive capability, it cannot be considered high enough to be meaningful. This low predictive quality is likely owed to the fact that dietary data – a major predictor in plasma vitamin C status – was both low in resolution and had too low a frequency of collection. For effective PN research, higher quality data will be required. Of all of the regression algorithms investigated, XGBoost showed the best performance, closely followed by RF and then SVR. The superiority of these algorithms is in line with prior expectations since these algorithms have excellent reputations for solving regression tasks. Neither clustering nor PCA improved prediction accuracy. The current study adds to the PN body of literature by showing that data of a higher resolution than the NHANES provides is required for PN research.

Acknowledgements

Open Access funding provided by the Qatar National Library.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Daniel Kirk http://orcid.org/0000-0001-7738-7686
Cagatay Catal http://orcid.org/0000-0003-0959-2930
Bedir Tekinerdogan http://orcid.org/0000-0002-8538-7261

References

Archer, E., G. Pavela, and C. J. Lavie. 2015. The inadmissibility of what we eat in america and NHANES dietary data in nutrition and obesity research and the scientific formulation of national dietary guidelines. *Mayo Clinic Proceedings* 90 (7):911-926. doi:10.1016/j.mayocp.2015.04.009

Awad, M., R. Khanna, M. Awad, and R. Khanna. 2015. Support Vector Regression. In *Efficient Learning Machines*, Jeffrey Pepper; Steve Weiss; Patrick Hauke, edited by, 67-81. New York, US: Apress. doi: 10.1007/978-1-4302-5990-9_4.

Baek, J. W., J. C. Kim, J. Chun, and K. Chung. 2019. Hybrid clustering based health decision-making for improving dietary habits. *Technology and Health Care* 27 (5):459–72. doi:10.3233/THC-191730.

Berry, S. E., A. M. Valdes, D. A. Drew, F. Asnicar, M. Mazidi, J. Wolf, J. Capdevila, G. Hadjigeorgiou, R. Davies, H. Al Khatib, et al. 2020. Human postprandial responses to food and potential for precision nutrition. *Nature Medicine* 26 (6):964–73. doi:10.1038/s41591-020-0934-0.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.

Bzdok, D., N. Altman, and M. Krzywinski. 2018. Points of significance: statistics versus machine learning. *Nature Methods* 15 (4):233–34. doi:10.1038/NMETH.4642.

Carr, A. C., and S. Rowe. 2020. Factors affecting vitamin c status and prevalence of deficiency: A global health perspective. *Nutrients* 12(7): 1963. Switzerland. MDPI AG. doi:10.3390/nu12071963.

Cecil, J. E., and K. L. Barton. 2020. Inter-individual differences in the nutrition response: From research to recommendations. *Proceedings of the Nutrition Society* 79 (2):171–73. doi:10.1017/S0029665119001198.

Celis-Morales, C., K. M. Livingstone, C. F. M. Marsaux, A. L. Macready, R. Fallaize, C. B. O'Donovan, C. Woolhead, H. Forster, M. C. Walsh, S. Navas-Carretero, et al. 2017. Effect of personalized nutrition on health-related behaviour change: Evidence from the Food4Me European randomized controlled trial. *International Journal of Epidemiology* 46 (2):578–88. doi:10.1093/IJE/DYW186.

Chen, T., and C. Guestrin 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California. doi:10.1145/2939672.2939785

de Toro-martín, J., B. J. Arsenault, J. P. Després, and M. C. Vohl. 2017. Precision nutrition: A review of personalized nutritional approaches for the prevention and management of metabolic syndrome. In *Nutrients*, vol. 9 (8): 913. Switzerland: MDPI AG. doi:10.3390/nu9080913.

Dehghan, M., N. Akhtar-Danesh, C. R. McMillan, and L. Thabane. 2007. Is plasma vitamin C an appropriate biomarker of vitamin C intake? A systematic review and meta-analysis. *Nutrition Journal* 6. doi:10.1186/1475-2891-6-41.

Hagel, A. F., H. Albrecht, W. Dauth, W. Hagel, F. Vitali, I. Ganzleben, H. W. Schultis, P. C. Konturek, J. Stein, M. F. Neurath, et al. 2018. Plasma concentrations of ascorbic acid in a cross section of the German population. *Journal of International Medical Research* 46 (1):168–74. doi:10.1177/0300060517714387.

Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2 (3):283–304. doi:10.1023/A:1009769707641.

Jollife, I. T., and J. Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374:20150202. doi:10.1098/rsta.2015.0202.

Jungert, A., and M. Neuhäuser-Berthold. 2015. The lower vitamin C plasma concentrations in elderly men compared with elderly women can partly be attributed to a volumetric dilution effect due to differences in fat-free mass. *British Journal of Nutrition* 113:859–64. doi:10.1017/S0007114515000240.

Juraschek, S. P., E. Guallar, L. J. Appel, and E. R. Miller. 2012. Effects of vitamin c supplementation on blood pressure: A meta-analysis of randomized controlled trials. *American Journal of Clinical Nutrition* 95 (5):1079–88. doi:10.3945/ajcn.111.027995.

Kanungo, T., D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. 2002. An efficient k-means clustering algorithms: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (7):881-892. doi:10.1109/TPAMI.2002.1017616

Kim, J. C., and K. Chung. 2020. Knowledge-based hybrid decision model using neural network for nutrition management. *Information Technology and Management* 21 (1):29–39. doi:10.1007/s10799-019-00300-5.

Kirk, D., C. Catal, and B. Tekinerdogan. 2021. Precision nutrition: A systematic literature review. *Computers in Biology and Medicine* 133. doi:10.1016/J.COMPBIOMED.2021.104365.

Kraemer, C. M. 2020. Vitamin C (Ascorbic Acid): Reference Range, Interpretation, Collection and Panels. Reference Range, Interpretation, Collection and Panels. Retrieved from https://emedicine.medscape.com/article/2088649-overview (Accessed on 20 June 2021).

Mannor, S., X. Jin, J. Han, X. Jin, J. Han, X. Jin, J. Han, and X. Zhang. 2011. K-Means Clustering. In *Encyclopedia of Machine Learning*, 563–64. US: Springer. doi:10.1007/978-0-387-30164-8_425.

McDonald, G. C. 2009. Ridge regression. *Wiley Interdisciplinary Reviews. Computational Statistics* 1 (1):93–100. doi:10.1002/wics.14.

Myint, P. K., R. N. Luben, N. J. Wareham, and K. T. Khaw. 2011. Association between plasma vitamin C concentrations and blood pressure in the European prospective investigation into cancer-Norfolk population-based study. *Hypertension* (3):372–79. doi:10.1161/HYPERTENSIONAHA.111.171108.

Ness, A. R., D. Chee, and P. Elliott. 1997. Vitamin C and blood pressure - an overview. *Journal of Human Hypertension* 11 (6):343–50. doi:10.1038/sj.jhh.1000423.

NHANES. 2015. Measuring guides for the dietary recall interview. accessed 2020. https://www.cdc.gov/nchs/nhanes/measuring_guides_dri/measuringguides.htm

NHANES. 2017. About the national health and nutrition examination survey. accessed 2020. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

NIH. 2020. Vitamin C - fact sheet for health professionals. accessed https://ods.od.nih.gov/factsheets/VitaminC-HealthProfessional/#:%7E:text=Vitamin%20C%20is%20required%20for,vital%20role%20in%20wound%20healing

Olive, D. J. 2017. Linear regression. *Linear Regression*. doi:10.1007/978-3-319-55252-1.

Ordovas, J. M., L. R. Ferguson, E. S. Tai, and J. C. Mathers. 2018. Personalised nutrition and health. *British Medical Journal* 361. doi:10.1136/bmj.k2173.

Ramyaa, R., O. Hosseini, G. P. Krishnan, and S. Krishnan. 2019. Phenotyping women based on dietary macronutrients, physical activity, and body weight using machine learning tools. *Nutrients* 11 (7):1681. doi:10.3390/nu11071681.

Rowe, M. 2019. An Introduction to Machine Learning for Clinicians. *Academic Medicine* 94 (10):1433–36. doi:10.1097/ACM.0000000000002792.

Shiao, S. P. K., J. Grayson, A. Lie, and C. H. Yu. 2018. Personalized nutrition—genes, diet, and related interactive parameters as predictors of cancer in multiethnic colorectal cancer families. *Nutrients* 10 (6):795. doi:10.3390/nu10060795.

Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. doi:10.1111/j.2517-6161.1996.tb02080.x.

Travica, N., K. Ried, A. Sali, I. Hudson, A. Scholey, and A. Pipingas. 2019. Plasma Vitamin C concentrations and cognitive function: A cross-sectional study. *Frontiers in Aging Neuroscience* 11. doi:10.3389/fnagi.2019.00072.

Zeevi, D., T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, et al. 2015. Personalized Nutrition by Prediction of Glycemic Responses. *Cell* 163 (5):1079–94. doi:10.1016/j.cell.2015.11.001.