Motivation

This project investigates whether demographic and nutritional variables can be used to predict whether an individual uses dietary supplements. Using a classification approach, we examine how features such as age, income, education, and dietary intake relate to supplement use.

Understanding which groups are more likely to use supplements can help researchers describe health behavior patterns more clearly. It is also relevant from a business perspective, as supplement companies and digital health platforms may use this information to better understand their customer base and improve targeting or recommendations.

A previous research on university students in Warsaw found that supplement use was associated with socio-demographic and lifestyle factors such as gender, physical activity, smoking status, chronic disease, and dietary habits. The study also showed that people with higher nutritional knowledge and healthier diets were more likely to use supplements. Our project builds on this idea by testing whether similar patterns appear in our dataset and whether they can be captured through predictive modeling.

Dataset

The dataset used in this study comes from the National Health and Nutrition Examination Survey (NHANES), a national survey of health and nutrition on participants of all ages in the United States. We primarily used data from the 2021 - 2023 survey cycle for model training and testing, and additionally used data from the 2017 - March 2020 pre-pandemic cycle to further evaluate model generalization.

It includes demographic variables such as age, gender, and dietary intake variables such as total energy intake, protein, fat, and carbohydrates. These components were merged using a unique participant identifier so that each observation contains information for the same individual.
This dataset is appropriate for our research question because it includes the main variables needed to study the relationship between demographic factors, diet, and supplement use. Its large and nationally representative sample also makes it suitable for both statistical analysis and predictive modeling.

Exploratory Data Analysis

The preprocessing steps began by merging the demographic, dietary, and supplement datasets using the participant identifier (SEQN). We then created macronutrient proportions variables, encoded categorical variables, and defined a binary outcome variable indicating whether an individual used dietary supplements. We have also done median imputation for numeric variables, mode imputation for categorical variables to address missing data.

The dataset is a structured tabular dataset in which each row represents one participant and each column represents a variable. It contains a relatively large number of observations, with far more rows than columns. The variables include a mix of categorical variables and numerical variables. The outcome variable, supplement count, is strongly right-skewed, with many participants reporting zero or only a small number of supplements.

Initial visualizations, such as boxplots of macronutrient proportions by supplement use, showed only small differences between groups, suggesting that macronutrient intake alone may not strongly distinguish supplement users from non-users. In addition, clustering analysis suggested that individuals tended to group more strongly by demographic and lifestyle characteristics than by nutritional intake alone, indicating that supplement use may be more behaviorally driven than nutritionally driven.

Modeling Approach

We first implemented a dummy classifier that always predicts the majority class. This served as a reference point to determine whether other models were capturing meaningful patterns rather than benefiting only from class imbalance.

We then used logistic regression as our primary benchmark model. This choice was motivated not only by the binary outcome, but also by the goal of assessing whether a relatively simple, linear decision boundary could capture the relationship between predictors and supplement use. For improvement, we used class-weight balancing to account for class imbalance and applied a preprocessing pipeline that included imputation, standardization for numerical variables, and one-hot encoding for categorical variables.

To complement this, we trained a decision tree classifier to capture possible nonlinear relationships and interactions between variables. While logistic regression assumes a linear relationship between the predictors and the log-odds of outcome, supplement use may depend on threshold-based or interaction effects. The decision tree allowed us to test whether relaxing the linearity assumption would improve predictive performance.

In addition to the binary classification task, we also experimented with a multiclass random forest model to predict supplement-use groups (none, 1–5, or 5+). This was meant to extend the analysis beyond a simple yes/no outcome and test whether the intensity of supplement use could also be predicted.

We evaluated performance using a combination of metrics, including accuracy, balanced accuracy, precision, recall, F1-score, and ROC AUC. This combination was important because no single metric fully captures model performance. In particular, accuracy alone can be misleading when class proportions are uneven. In addition to numerical metrics, we used visualization tools such as ROC curves, confusion matrices, and comparative metric plots to better understand differences between models and identify patterns in misclassification.

Finally, we performed an external-style validation step by applying the trained models to a previous NHANES cycle dataset. This step allowed us to assess whether model performance generalizes beyond the original dataset and remains stable across different population samples.

Findings

Model Performance

Among the models, logistic regression achieved the best overall performance, with accuracy and balanced accuracy of approximately 0.69 and a ROC AUC of around 0.75. The decision tree performed slightly worse, with accuracy around 0.66 and ROC AUC around 0.72. This suggests that while nonlinear relationships may exist, they do not provide a strong advantage over a simpler linear model for this task. Overall, these results indicate that supplement use can be predicted with moderate accuracy using demographic and dietary variables.

As an extension, we also trained a random forest model for a multiclass supplement-use outcome (none, 1–5, and 5+). Its performance was weaker than the binary models, which is expected since distinguishing among three classes is more difficult than a binary classification task. The model failed to identify the “5+” group, likely because this class had very few observations relative to the others. This strong class imbalance limited the model’s ability to learn reliable patterns for high-frequency supplement users.

In addition, both models showed reasonably good generalization ability when evaluated on data from an earlier NHANES cycle. Logistic regression achieved an accuracy of about 0.69 and a ROC AUC of 0.74 on the previous-cycle data, while the decision tree reached an accuracy of about 0.67 and a ROC AUC of 0.71. This suggests that the identified relationships are not limited to a single survey period and remain fairly stable across cycles.

Interpretation of Results

Overall, the results suggest that dietary supplement use is associated with both demographic and dietary factors, but demographic variables provide the strongest and most consistent signal. This pattern was already visible in the exploratory analysis: boxplots of macronutrient proportions showed only small differences between supplement users and non-users, and clustering suggested that supplement-use behavior aligns more closely with demographic and lifestyle patterns than with macronutrient intake alone.

The modeling results support this interpretation. The decision tree showed that feature importance concentrated in a small number of predictors, especially age, indicating that a few strong variables explain much of the binary classification signal. In contrast, the random forest distributed importance across a broader set of features, including both dietary and socioeconomic variables, suggesting that multiple factors contribute jointly when the prediction task is more complex.

One reason for this difference is that dietary variables tend to provide weaker and more subtle signals. A single decision tree prioritizes the strongest predictors early, such as age, and may ignore these weaker effects. On the other hand, the random forest, by averaging across many trees, is better able to capture and accumulate the contribution of these smaller signals. Additionally, because the random forest is applied to a multiclass task (none, 1–5, 5+), it must make finer distinctions, which further increases the importance of dietary features.

These findings are not contradictory. Clustering identifies broad population structure, while random forest importance reflects how useful variables are for predicting a specific outcome. Taken together, the results suggest that supplement use is shaped primarily by demographic and lifestyle patterns, with dietary variables providing additional predictive value rather than serving as the main driver. More broadly, this indicates that supplement use may reflect general health-oriented behavior rather than being solely a response to nutritional deficiencies.

Model Issues and Limitations

One potential concern in this analysis is overfitting, particularly for the decision tree model. Decision trees can easily become overly complex and fit noise in the training data. To address this, we constrained the model using a maximum depth and a minimum leaf size, which helped improve generalization and maintain interpretability. In addition, the overall model performance suggests that there is still unexplained variability in supplement use. This indicates that important factors (e.g. health beliefs, access to healthcare, or personal preferences) may not be captured in the dataset.

Future Improvements

If more time were available, several improvements could be explored. First, further investigation into interaction effects between demographic and dietary variables could help uncover more complex relationships. Second, incorporating additional data sources (e.g. alcohol use, health status, medical history, etc.) could improve model performance and provide a more complete understanding of supplement-use behavior.

Division of Labor

Our team divided the project by major components while working together on key decisions and final outputs. One member took the lead on the main modeling task, including implementing the logistic regression and decision tree models. Another member focused on exploratory and extended analyses, including EDA, clustering, and a random forest model for a multiclass version of the problem. The rest of the work was completed collaboratively.

References

Sicinska, E., Madej, D., Szmidt, M. K., Januszko, O., & Kaluza, J. (2022). Dietary Supplement Use in Relation to Socio-Demographic and Lifestyle Factors, including Adherence to Mediterranean-Style Diet in University Students. *Nutrients, 14*(13), 2745. https://doi.org/10.3390/nu14132745

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, https://wwwn.cdc.gov/nchs/nhanes/default.aspx