National College of Ireland

Predicting Dietary B12 Deficiency in Vegetarians Using Interpretable Models

MSc Research Project
Master of Science in AI for Business

Felipe Rojas
Student ID: X23438941

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

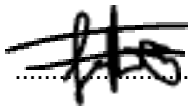
National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Felipe Rojas
Student ID:	X23438941
Programme:	Msc in AI for Business **Year:** 2025
Module:	Practicum
Supervisor:	Faithful Onwuegbuche
Submission Due Date:	11/08/2025
Project Title:	Predicting Dietary B12 Deficiency in Vegetarians Using Interpretable Models
Word Count:	7,052 **Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 08/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Dietary B12 Deficiency in Vegetarians Using Interpretable Models

Felipe Rojas
X23438941

Abstract

The health risks of vitamin B12 deficiency include anemia and neurological damage which primarily affect vegetarians due to their restricted access to B12 sources. The research develops an interpretable machine learning system which uses NHANES data to forecast B12 deficiency through dietary consumption and personal characteristics. The model analyzes vegetarian participants by applying SMOTE to handle class imbalance and Random Forest classification on B12 intake and demographic and nutritional characteristics. The research utilized SHAP and LIME interpretability tools to generate both general and specific explanations for each case. The study reveals essential nutrient-risk connections while showing how dietary-based predictive models can perform scalable non-invasive screenings. The method enables the early detection of vulnerable people while providing a clear nutritional public health tool.

1 Introduction

1.1 Background and Motivation

The human body requires Vitamin B12 to preserve neurological functions and produce red blood cells while synthetizing DNA. The body requires Vitamin B12 to prevent anaemia and fatigue and to avoid cognitive decline and permanent neurological damage (Allen, 2008; O'Leary and Samman, 2010). People who follow vegetarian or vegan diets are at higher risk because Vitamin B12 exists primarily in animal-based foods (Pawlak et al. 2013). Simultaneously, the diagnosis of B12 deficiency needs expensive clinical tests which demonstrates the necessity for non-invasive screening tools that are also scalable. Interpretable machine learning (ML) presents a promising path forward. The U.S. National Health and Nutrition Examination Survey (NHANES)[1] along with other large-scale public health surveys offer extensive self-reported data about diets and demographics. The analysis of B12 deficiency risk through predictive models becomes possible due to this data which it could mitigate the need for invasive diagnostics especially for vegetarians who need screening.

[1] U.S. Centers for Disease Control and Prevention. (n.d.). *National Health and Nutrition Examination Survey (NHANES)*. https://www.cdc.gov/nchs/nhanes/

1.2 Research Problem and Questions

Previous research has modelled B12 deficiency through biomarkers and clinical features, yet few studies have investigated predictive modelling with dietary recall data, particularly among vegetarian populations. This thesis looked for addressing that gap with the next questions:

1. Does the combination of food records and demographic information enable the prediction of Vitamin B12 deficiency in vegetarians without requiring clinical tests?

2. Which nutritional elements together with demographic characteristics play the most significant role in determining the risk of deficiency among vegetarian populations?

3. How can the implementation of interpretability tools such as SHAP and LIME enable better transparency to help in public health decision-making processes?

1.3 Contributions

The research demonstrates a repeatable machine learning approach which predicts Vitamin B12 deficiency by analyzing dietary information and demographic characteristics specifically for vegetarian populations who are typically excluded from nutritional risk assessment models. The Random Forest classifier received training from engineered features which included B12 intake per 1000 kcal and nutrient composition and demographic attributes while SMOTE handled the class imbalance of vegetarians in the training set. The integration of SHAP and LIME enabled both global and individual-level insights to promote interpretability. The combination of these elements establishes a new approach to dietary-based risk modeling which shows promise for public health screening at scale.

1.4 Thesis Structure

The following sections are structure as follows. Section two examines Vitamin B12 deficiency research together with NHANES dietary data modelling and explainable machine learning approaches. Section three explains the research methodology through data preprocessing steps and labelling strategy and model development methods. The system design is explored in Section 4 together with its design rationales. Section 5 explains implementation steps and tools. The model's predictive performance and interpretability outputs receive evaluation in Section 6. The final section of the thesis presents key findings and discusses limitations while suggesting research directions for the future.

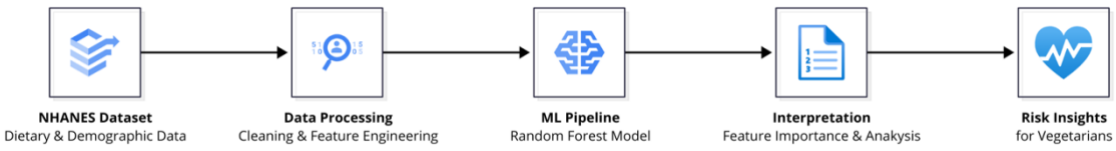


Figure 1: Overview of the Machine Learning Pipeline for B12 Deficiency Risk Modelling

2 Related Work

2.1 Vitamin B12 Deficiency in General and Vegetarian Populations

Vitamin B12 or cobalamin, plays an important role in neurological functions, DNA synthesis and red blood cell production. The research of Allen (2008), Henjum et al. (2023) and Pawlak et al. (2016) shows that B12 deficiency could lead to anemia along with cognitive problems and neuropathy. The worldwide distribution of B12 deficiency shows its highest rates among elderly people and vegetarians and vegans because they consume few animal-derived foods according to Pawlak et al. (2016).

The risk of B12 deficiency becomes higher for vegetarians who avoid dairy products and eggs unless they take supplements (Henjum et al. 2023). Research indicates that dietary fortification or supplementation can reduce these risks yet real-world compliance rates show significant variation. The gold standard for diagnosing B12 deficiency requires serum B12 tests along with methylmalonic acid (MMA) and homocysteine measurements yet these tests remain inaccessible in non-clinical environments (Allen et al. 2018). Self-reported dietary information could serve as a cost-effective prediction tool for B12 deficiency risk assessment.

The study by Tamune et al. (2020) along with other recent research attempted to detect B12 deficiency through machine learning models which used symptoms and laboratory tests but these models were restricted to elderly inpatient populations and did not generalize well.

In contrast, our research fills the existing knowledge gap through its focus on vegetarian populations while developing B12 deficiency risk models that use self-reported dietary and demographic information instead of clinical or biomarker data.

2.2 NHANES in Nutritional Epidemiology and Machine Learning

The National Health and Nutrition Examination Survey (NHANES) stands as one of the most extensive datasets which contains health information together with nutritional data and demographic statistics in the United States. The dataset serves as a primary source for epidemiological research to establish links between dietary consumption and disease results (Mineva et al. 2021). The dataset supports two recent applications which use machine learning models to forecast diabetes and obesity conditions (Samson et al. 2022; Guo et al. 2024).

The NHANES dataset provides essential population-level dietary and health information but researchers must acknowledge its limitations. The dietary intake data from 24-hour recalls depends on self-reported information which might result in recall errors and underreporting of micronutrient consumption. The interpretation of model outputs from this dataset requires consideration of these factors. The use of NHANES dietary data to predict micronutrient deficiencies through machine learning remains a relatively unexplored field despite its rich data resources. Most rely on biomarker data. The XGBoost model trained by Guo et al. (2024) utilized clinical and laboratory features to predict B12 deficiency while employing SHAP for interpretation. The model depends on MCV biomarkers for its operation which restricts its application in community-based settings.

Our research stands as a pioneering effort which applies NHANES 2017–2023 dietary information to create an explainable B12 deficiency prediction model for vegetarian subgroups.

2.3 Machine Learning for Health: Class Imbalance and Interpretability

The main issue in clinical datasets involves underrepresentation of positive cases such as B12 deficiency. The Synthetic Minority Oversampling Technique (SMOTE) stands as a popular solution for handling class imbalance problems especially when used in healthcare ML applications (Chawla et al. 2002; Fernández et al. 2018). Research in nutritional datasets demonstrates the effectiveness of combining SMOTE with random forest classifiers to boost minority class recall (Blagus and Lusa, 2013).

The first step in our pipeline involved stratifying the NHANES dataset before selecting 5,000 representative samples to reduce sampling bias. The model gained better understanding of minority class patterns through SMOTE-based augmentation of vegetarian participants in the training data. The health ML literature supports this dual approach because it enhances model fairness and representation (Fernández et al. 2018).

The importance of interpretability surpasses accuracy requirements. SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al. 2016) are widely accepted explainability methods. The SHAP method explains model predictions through both global and local feature attributions yet LIME explains individual predictions by approximating the decision boundary locally. The clinical ML field uses these methods to support patient and practitioner decision-making (McKelvey et al. 2018).

This study used a two-stage process which started with dataset reduction to 5,000 stratified rows followed by SMOTE-based augmentation of the vegetarian subgroup and ended with Random Forest model training. The integration of SHAP and LIME addressed both the imbalance problems and explainability gaps which previous research had identified.

2.4 Gaps in the Literature and Contribution

Numerous limitations exist in AI-driven nutrition research studies according to the present academic literature. The majority of predictive studies for vitamin B12 deficiency depend on biomarkers including serum B12 and MMA levels but these diagnostics prove expensive and impractical for broad community-based applications. The exclusive use of clinical markers in the existing research makes it difficult to apply their findings to public health programs.

The second drawback stems from the scarcity of predictive modeling research which utilizes population-based dietary data from NHANES for micronutrient deficiency identification. Studies based on NHANES data tend to examine chronic diseases and macronutrient patterns but rarely investigate specific nutrient deficiencies.

Many current models demonstrate insufficient attention to interpretability within their frameworks. The research places performance metrics at the forefront but fails to develop models which deliver explainable or transparent predictions because such transparency remains crucial for healthcare decision-making and nutritional guidance.

The majority of previous research fails to provide detailed explanations about their feature engineering steps. The absence of domain-aware feature creation methods reduces

reproducibility and makes model outputs more difficult to interpret. The development of explainable and actionable models requires feature engineering that adheres to nutritional science principles.

The authors of Tamune et al. (2020) and Guo et al. (2024) performed B12 prediction using ML but used dietary intake as the primary predictor while depending on clinical datasets. The research failed to implement SMOTE in combination with stratified undersampling and dual explainability methods.

The study establishes itself through its implementation of interpretable domain-grounded features that include a vegetarian flag created from USDA[2] food codes and RDA-based deficiency labels following NIH[3] guidelines. The transparent engineering framework helps achieve ethical ML design by maintaining dietary science alignment with model logic.

The research stresses both fairness and ethical modeling as essential components. Researchers treated NHANES indicators of race, gender and socioeconomic status with precision. Our implementation of SHAP and LIME techniques promotes ethical transparency while detecting bias patterns that align with requirements for explainable techniques in sensitive domains (Doshi-Velez and Kim, 2017; Barocas et al. 2019).

Our system integrates NHANES dietary data exclusively with SMOTE class balancing and SHAP and LIME interpretability techniques to develop a scalable risk assessment tool for B12 deficiency identification. The proposed methodology resolves both theoretical and practical shortcomings that exist in the existing research literature.

Table 1: Comparative Summary of Related Work

Study	Dataset Type	Target Population	Predictors Used	Interpretability	Notes
Tamune et al. (2018)	Geriatric Hospital Data	Elderly inpatients	Syptoms and labs	None	Clinical scope only
Guo et al. (2024)	Routine Clinical Data	General clinical	Labs (e.g., MCV)	SHAP	Biomarker dependent
This study	NHANES Dietary Data	Vegetarians (subset of NHANES)	Food intake and demos	SHAP and Lime	Diet-only inputs, transparent feature engineering

The current models for B12 deficiency prediction depend on biomarkers and clinical assessment or lack interpretability. The project fills these gaps by using a transparent diet-based ML pipeline trained on NHANES data. The following section explains the methodological choices that support this approach.

[2] U.S. Department of Agriculture, *FoodData Central*, https://fdc.nal.usda.gov

[3] National Institutes of Health, Office of Dietary Supplements, *Vitamin B12 Fact Sheet for Health Professionals*, https://ods.od.nih.gov/factsheets/VitaminB12-HealthProfessional/

3 Research Methodology

This section describes the complete machine learning process from data collection to model understanding. The structured pipeline starts with data sourcing followed by preprocessing and exploratory data analysis (EDA) before moving to feature engineering and label creation and imbalance correction and feature selection and model training with interpretability.

3.1 Dataset Overview

The research draws its data from the National Health and Nutrition Examination Survey (NHANES) which functions as a major health monitoring system under the U.S. Centers for Disease Control and Prevention (CDC) (CDC NHANES Documentation, n.d.). The research analysed two NHANES cycles by merging pre-pandemic dietary data from the NHANES 2017–March 2020 Pre-Pandemic Dataset with post-COVID data from the NHANES 2021–2023 Dataset.

3.1.1 Initial and Final Sample

The merged dataset contained 16,245 participants who received unique SEQN codes at the beginning. The NHANES modules DR1IFF and DR2IFF and DEMO provided data for the 24-hour dietary recall (Days 1 and 2) and demographic attributes (e.g., gender, age). The dataset included three categories of information: nutrient intake data and energy totals and demographic attributes. The dataset underwent extensive filtering to achieve data reliability and completeness through the removal of participants with missing or invalid values and the selection of individuals who provided complete data for B12 intake and energy consumption and core demographic information. The cleaning process resulted in 16,244 participants becoming eligible for the full dataset.

A stratified random sample of 5,000 participants was selected from this population to achieve better representation across vegetarian status. The model training became more efficient through this reduction while maintaining essential distribution patterns. The vegetarian subset in this reduced sample remained underrepresented so SMOTE was used for augmentation (see Section 3.4). The following features were constructed based on the available data.

- b12_density
- vegetarian_flag
- nutrient averages (including iron_avg, folate_avg, choline_avg)
- energy intake
- gender and age (one-hot encoded and normalized)

3.1.2 Deficiency Label

The main outcome variable B12 deficiency received its binary label through NIH RDA thresholds which are presented in Table 2 for different age and gender. The labeling system allows risk detection without biomarker dependence as described by Guo et al. (2024).

Table 2: Dietary thresholds for B12 deficiency labeling by age and gender, according to NIH.

Age	Male	Female	Pregnancy	Lactation
Birth to 6 months*	0.4 mcg	0.4 mcg		
7–12 months*	0.5 mcg	0.5 mcg		
1–3 years	0.9 mcg	0.9 mcg		
4–8 years	1.2 mcg	1.2 mcg		
9–13 years	1.8 mcg	1.8 mcg		
14–18 years	2.4 mcg	2.4 mcg	2.6 mcg	2.8 mcg
19+ years	2.4 mcg	2.4 mcg	2.6 mcg	2.8 mcg
*** Adequate Intake (AI)**				

3.1.3 Vegetarian Status Label

The absence of meat, poultry and fish consumption during the 24-hour dietary recall led to the conclusion that the participant followed a vegetarian diet. The dietary survey method used to evaluate plant-based deficiency risk matches the approach described in Pawlak et al. (2016). The nutrient distribution among flagged participants showed enough validity to use this label as a predictive variable despite the method's limitations.

The b12_deficiency label originated from dietary thresholds before sampling and the vegetarian_flag emerged from the absence of meat, poultry and fish during both recall days. The cleaned dataset underwent a random reduction to 5,000 participants for easier handling. The application of SMOTE during training created additional vegetarian cases to address dietary class imbalance. The new dataset contained approximately 9,923 participants after this process.

3.1.4 Class Distribution and Modelling Challenge

The 16,244 dataset showed an unbalanced distribution of vegetarians compared to non-vegetarians. The model training process became difficult because of the limited number of plant-based diet participants. The training phase included SMOTE to generate additional vegetarian records which improved the robustness of the subgroup model (see Section 3.4).

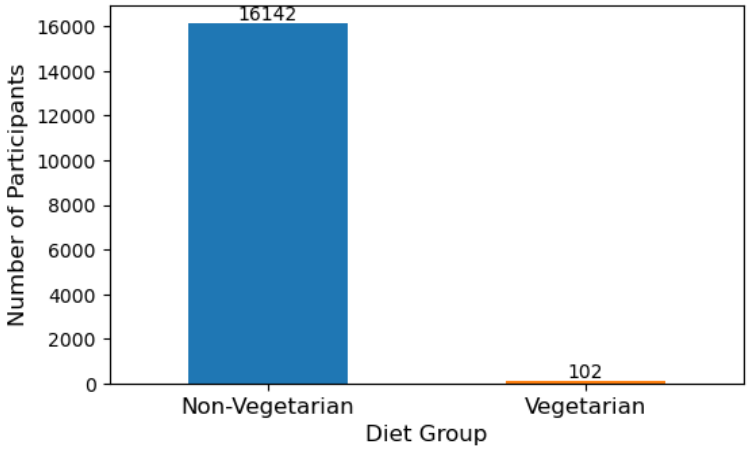


Figure 2: Bar Chart of Diet Group Distribution in the Full Sample (Vegetarian vs. Non-Vegetarian)

3.2 Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) was performed on the final cleaned dataset of 16,244 participants, each with complete nutritional, demographic, and engineered features. The goal was to examine distributional patterns, detect skewness and outliers, explore correlations, and determine which features are related to the binary Vitamin B12 deficiency label. All analysis and visualization were performed in Python using standard data visualization libraries (see Section 5.1).

To describe nutrient distributions, histograms and boxplots were created for key features such as avg_b12, iron_avg, folate_avg, and b12_density. These variables were calculated as two-day averages from NHANES Day 1 and Day 2 dietary recalls. While most nutrient values were reported as absolute daily intake (in micrograms per day), b12_density was explicitly normalized per 1000 kcal to adjust for caloric differences (Drewnowski, 2005). Several of the nutrient features, particularly B12 and iron, displayed right-skewed distributions, indicative of a few individuals with high intake. Outliers were retained intentionally to reflect the full range of dietary behavior present in population-level intake data (Wilcox, 2012).

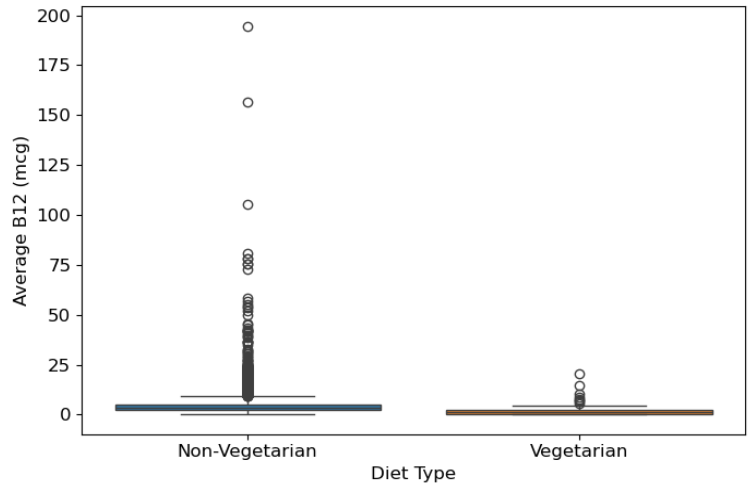


Figure 3: Boxplot of average B12 intake by vegetarian status

We evaluated how different nutrients distributed among vegetarian and non-vegetarian participants. The boxplots showed that vegetarians consumed less B12 and iron and folate than non-vegetarians (see Figure 3). The B12 intake of vegetarians showed a narrow distribution that stayed below the deficiency threshold which indicates this group faces a higher risk of deficiency (Pawlak et al. 2016; Henjum et al. 2023).

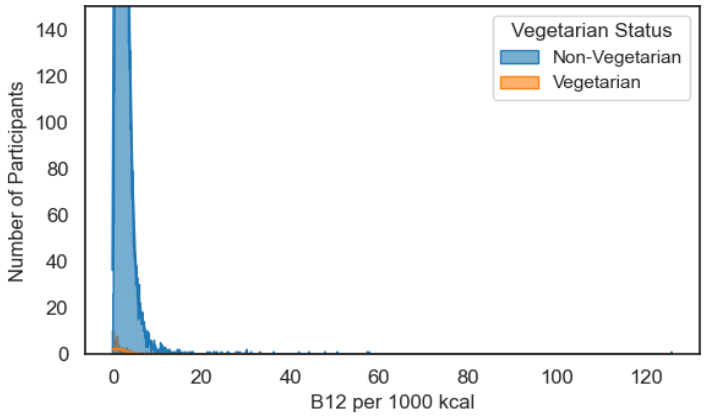


Figure 4: Histogram of B12 density (µg/1000 kcal)

The histogram of b12_density demonstrated a distinct pattern of low-density values in the deficient class which confirmed its predictive value (Obeid et al. 2019).

A Pearson correlation matrix was used to perform correlational analysis. The analysis showed moderate positive relationships between iron and folate and between total energy intake and most nutrients. The analysis did not show any critical multicollinearity that would require feature removal, but high correlations were noted for interpretation (Dormann et al. 2013).

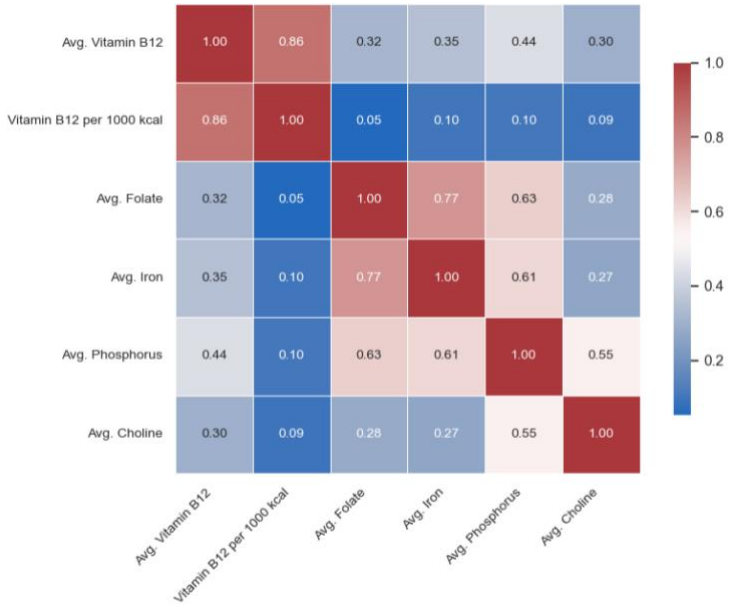


Figure 5: Pearson Correlation Heatmap across nutrient and energy intake variables

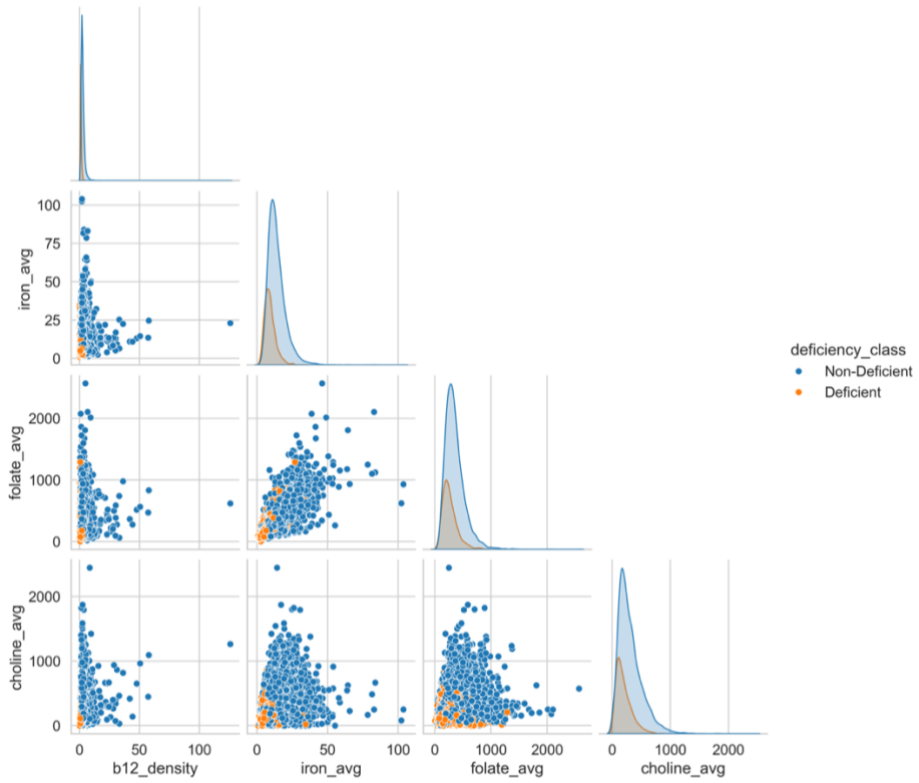


Figure 6: Pairplot of Key Nutrient Features by Deficiency Class

The figure demonstrates the connections between four essential nutrient variables (b12_density, iron_avg, folate_avg and choline_avg) across different deficiency classes. The orange group representing Vitamin B12 deficiency shows a clear pattern of low nutrient consumption especially for B12 and iron which supports their importance in the model.

The distinction between classes became most visible when examining b12_density which validated its status as a main model feature according to Guo et al. (2024).

The EDA results demonstrated significant distributional patterns and distinct feature spaces which validated the use of supervised learning methods for B12 dietary deficiency prediction.

3.3 Exploratory Data Analysis (EDA)

The NHANES raw data underwent transformation to create a structured feature matrix which was appropriate for machine learning applications. The process of feature engineering involved both normalization of variables and transformation of categorical data into numerical formats.

The binary outcome variable b12_deficiency was created through NIH Recommended Dietary Allowances for age and sex (see Section 3.1) which allowed supervised classification between deficient (class 1) and non-deficient (class 0) categories.

The final feature set included:

- **b12_density:** B12 intake per 1000 kcal, normalizing for energy consumption.
- **vegetarian_flag:** Binary indicator from the absence of meat, poultry and fish consumption.
- **Nutrient averages:** B12, Iron, folate, choline, and other micronutrient was calculated from both dietary recall days.
- **Energy intake:** Total caloric intake.
- **Demographics:** Age and gender (encoded for modeling).
- **B12 deficiency:** Target variable; defined using Recommended Dietary Allowance (RDA) thresholds from the National Institutes of Health (NIH).
- **rda_b12:** Reference thresholds for recommended daily intake, based on NIH guidelines.
- **SEQN:** Unique identifier for each participant of NHANES.

All continuous variables were normalized, and categorical variables encoded to ensure compatibility with the model and enhance interpretability. These features formed the basis for model training and evaluation.

3.4 Handling Class Imbalance with SMOTE

The dataset showed underrepresentation of vegetarian participants who formed the core population for studying B12 deficiency in plant-based diets according to Section 3.1. The uneven distribution of participants between vegetarian and non-vegetarian groups created doubts about how well the model would perform in predicting deficiency risk for vegetarian populations.

The application of Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the vegetarian class, enhancing the number of vegetarian participants in the training data instead of balancing the B12 deficiency label. The analytical unit needed balance through this methodological shift which treated the vegetarian subgroup as the analytical unit requiring balance.

We reduced the training dataset to a representative subset of approximately 5,000 samples before applying SMOTE for computational efficiency. The application of SMOTE occurred after splitting the dataset into training and test sets to maintain synthetic records within the training set while preserving test set validity.

The training dataset contained 9,922 participants who received equal representation between vegetarians and non-vegetarians. The model gained better capability to detect patterns related to deficiency risk in vegetarian populations after this improvement.

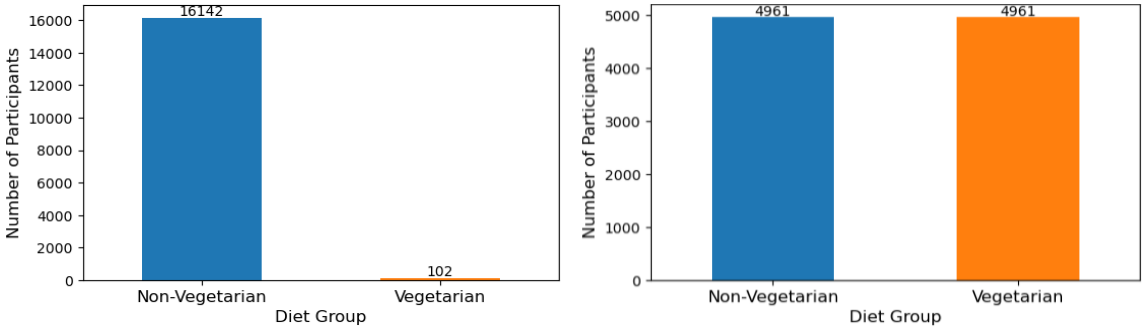


Figure 7: Distribution of vegetarian_flag before and after SMOTE, bar chart showing subgroup oversampling.

3.5 Feature Selection

The feature selection process followed class rebalancing and feature engineering by using statistical and model-based importance measures. All engineered features remained in the dataset except for those which directly encoded the target variable. The variable rda_b12 which shows whether participants reached their RDA threshold was eliminated because it directly related to the label thus causing label leakage. The participant identifier SEQN was excluded from the analysis.

The selection process received support from both model-based and statistical methods. The feature importances values from Random Forest received comparison with F-scores obtained through SelectKBest with ANOVA (Pedregosa et al. 2011). The two methods produced their respective rankings which are shown in Table 3.

The exploratory nature of this study along with Random Forest interpretability strengths led to the retention of all remaining features (Section 4.1 explains the model selection). The Random Forest model gained the ability to detect patterns through this approach which avoided duplicate information. The SHAP values (Lundberg and Lee, 2017) confirmed the model-based and statistical rankings that were presented in section 6.2.

Table 3: Top 9 Features Ranked by Random Forest and SelectKBest

Feature	Random Forest Rank	KBest Rank
Avg. Vitamin B12	1 (0.2872)	1 (1475.5)
Vitamin B12 per 1000 kcal	2 (0.2854)	2 (722.6)
Avg. Energy Intake (kcal)	3 (0.1104)	5 (14.3)
Avg. Iron	4 (0.0927)	3 (1073.1)
Avg. Phosphorus	5 (0.0776)	4 (196.4)
Avg. Choline	6 (0.0473)	6 (212.8)
Avg. Folate	7 (0.0314)	7 (9.5)
Age	8 (0.0662)	8 (824.5)
Gender	9 (0.0019)	9 (7.5)

3.6 Model Training

We evaluated model performance through two experimental approaches: (1) training on the complete 16,244 participant dataset with no oversampling (imbalanced) and (2) training on a reduced stratified subset of 5,000 participants followed by SMOTE to enhance the vegetarian class (balanced).

The data received stratified sampling for training and test set creation to maintain the original class distribution in both configurations.

We applied SMOTE to enhance the vegetarian subgroup within the training data instead of using it for deficiency labels. Section 3.4 explains how synthetic augmentation increased the number of vegetarian participants to enhance model generalization in this underrepresented dietary group. The test set remained unaltered to prevent evaluation data contamination and performance metric inflation.

The training data received SMOTE augmentation which resulted in a total of 9,923 samples with enhanced vegetarian representation. The model training process used the SMOTE-augmented training data as its input. The implementation used Python with open-source libraries for execution.

The model received training from SMOTE-augmented data before being evaluated on the original test data. The evaluation of performance relied on three main metrics:

- **Accuracy:** Overall correctness
- **Recall:** Sensitivity for the deficient class
- **F1-score:** Armonic mean of precision and recall

The assessment focused primarily on recall for the deficient class because false negative reduction stands as a top clinical priority in deficiency screening.

Section 6.1 presents detailed information about model performance metrics along with visual evaluation results including confusion matrices and precision-recall comparisons.

3.7 Model Interpretability: SHAP and LIME

This research employed two model-agnostic interpretability methods called SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to improve model transparency and support clinical validation. The selection of these methods with their scope and their conceptual differences are explained in Section 4.3. SHAP analysed feature contributions at both the global and local levels and LIME provided instance-level explanations for particular test cases. These tools allowed to obtain both population-level and participant-level insights. The results of the implementation together with visualizations and findings are displayed in Sections 6.2 and 6.3. SHAP was used to assess global and local feature contributions, while LIME was applied to specific test instances for instance-level explanation. These tools enabled both population-level and participant-level insights. Detailed implementation results, visualizations, and findings are presented in Sections 6.2 and 6.3.

4 Model Architecture and Explainability Design

This section explains the fundamental logic and technical structure of the machine learning framework which this research used. The system focused on making predictions while keeping the model interpretable and reducing bias, ensuring reproducibility since these are vital elements in health prediction tasks. We describe our preferred algorithm alongside explainability tool implementations (SHAP and LIME) and explain the fundamental design principles of the pipeline.

4.1 Model Choice Rationale

The task requires predicting Vitamin B12 deficiency risk from dietary and demographic characteristics while maintaining accurate and understandable predictions. Two model types were selected for the purpose of this study:

- **The Random Forest Classifier (RFC)** was chosen for its robustness to noise and ability to detect nonlinear relationships and its built-in feature importance metrics. RFCs work efficiently on tabular data with mixed types (numerical and categorical) which matches the structure of typical nutritional datasets (Breiman, 2001; McKelvey et al. 2018). The research used RFC on a SMOTE-augmented subset of vegetarian participants to analyze deficiency patterns that are unique to this dietary group.
- **Explainable Models (SHAP and LIME-compatible):** The application of the model to nutritional health screening required models which support SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). These tools enable healthcare providers to understand the reasoning behind their predictions instead of treating the algorithm as a mystery (Lundberg and Lee, 2017; Ribeiro et al. 2016).

The interpretability becomes essential for risk analysis in subgroups such as vegetarians since their clinical guidelines differ. The decision to prioritize interpretability over raw accuracy was made after considering both ethical requirements and practical limitations. Models used in public health need to be auditable and understandable while also being actionable according to Doshi-Velez and Kim (2017).

4.2 Model Pipeline

The machine learning pipeline was implemented in Python along with open-source libraries which included scikit-learn (Pedregosa et al. 2011), imbalanced-learn (Lemaître et al. 2017), pandas, SHAP (Lundberg and Lee, 2017), and LIME (Ribeiro et al. 2016). The modular design integrates all steps from data ingestion and preprocessing to model training and interpretability, as shown in Figure 8.

This structure supports transparency, reproducibility, and component substitution, for instance, replacing the classifier or balancing method while maintaining alignment with explainable AI best practices for health applications.

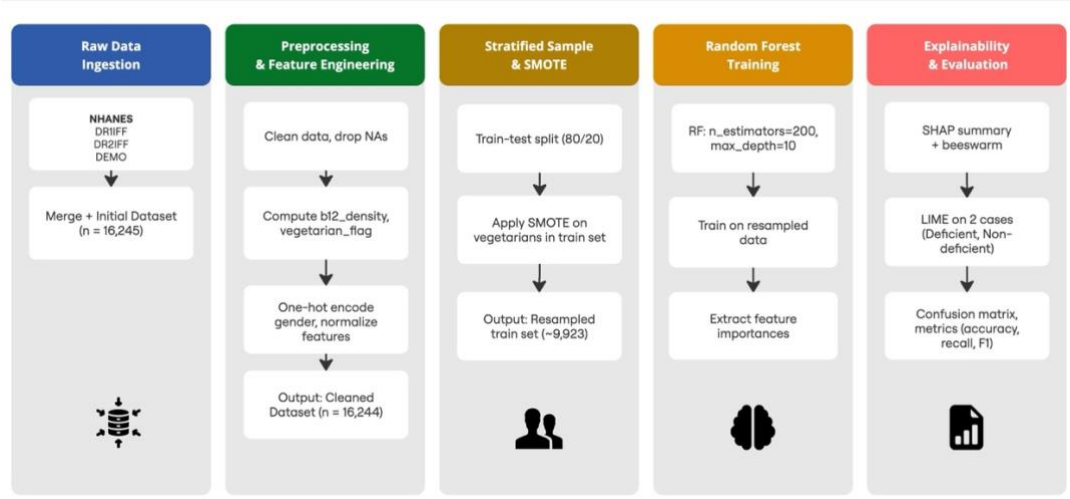


Figure 8: Modular Pipeline for Dietary Data Processing, B12 Deficiency Prediction, and Model Explainability

4.3 Interpretability Framework

The trained Random Forest model was analysed through SHAP and LIME to achieve transparency and interpretability. The SHAP method delivered both general and specific feature attribution values which helped to detect overall patterns and understand individual cases (Lundberg and Lee, 2017). The LIME method provided interpretable prediction explanations for specific instances including false negatives (Ribeiro et al. 2016).

The selected methods proved effective for dietary and epidemiological modelling because they let domain experts assess prediction drivers at both population and individual levels. The pipeline integration in Table 4 enables ethical AI principles through auditability and explainability as described by Doshi-Velez and Kim (2017).

Table 4: SHAP vs. LIME Interpretability Comparison

Method	Scope	Output	Primary Use
SHAP	Global + Local	Feature importance values	Understanding overall model behavior
LIME	Local only	Local surrogate explanation	Inspecting edge cases and individual predictions

The dual method was considered to conduct population-level and participant-level audits. SHAP provided a broad understanding of model reasoning, while LIME gave more detailed insights into individual predictions. For specific visualizations and analysis, see Sections 6.2 and 6.3. The tools provided transparency and interpretability throughout the model development process and evaluation phase.

5 Implementation

This section describes the practical deployment of the Vitamin B12 deficiency prediction system by detailing how the methodological pipeline was technically implemented. The following section demonstrates how each component was implemented in code using Python-based libraries and Jupyter notebooks (Kluyver et al. 2016).

5.1 Tools and Environment

The project was developed in Python 3.10 in Jupyter Notebook due to its easy experimentation capabilities along with visualization tools and documentation features. The data processing tasks were accomplished through pandas and numpy libraries while scikit-learn (Pedregosa et al. 2011) handled modelling and evaluation tasks. The implementation of SMOTE resampling used imbalanced-learn (Chawla et al. 2002) and matplotlib/seaborn (Hunter, 2007; Waskom, 2021) performed the visualization tasks. The project utilized SHAP and LIME (Lundberg and Lee, 2017; Ribeiro et al. 2016) for interpretability purposes. The experiments were conducted on a local Intel i7, 16GB RAM workstation.

5.2 Pipeline Implementation and Configuration

The cleaned dataset contained 16,245 participants together with engineered features including nutrient averages and B12 density and vegetarian flag and a target variable based on NIH RDA thresholds. The categorical features received encoding while the continuous variables maintained their original measurement units for better interpretability. The analysis excluded participants who had missing nutrient information.

The 80/20 train-test split maintained the original class distribution of the data. The training set vegetarian subset received SMOTE treatment which expanded its sample size to approximately 9,923. The Random Forest Classifier (n_estimators=200, max_depth=10) received training to extract feature importance for interpretability purposes.

5.3 Ethical Considerations

The research study adhered to ethical standards for working with health data. The NHANES dataset exists as public domain information which contains no personally identifiable details. The study used SMOTE to address vegetarian underrepresentation while excluding sensitive variables including income and geography and selecting only relevant demographics for RDA. The selection of Random Forest as the model was based on its interpretability features and its ability to work with SHAP and LIME for both global and case-specific explanations. The evaluation process focused on achieving high recall rates for deficient classes together with F1-score performance. The pipeline operated with open-source Python libraries while maintaining full documentation to support reproducibility.

6 Evaluation

The following section provides an extensive evaluation of machine learning models which predict Vitamin B12 deficiency. The evaluation focuses on quantitative performance metrics together with interpretability outputs to determine both model performance and decision-making processes.

6.1 Performance Metrics

The Random Forest classifier received evaluation through standard classification metrics which included Accuracy, Precision, Recall and F1-score (Sokolova and Lapalme, 2009; Powers, 2011). The model required special attention to Recall performance for deficient class participants because of the significant class imbalance between deficient and non-deficient participants. The model received two performance assessments through evaluation of the original imbalanced dataset followed by evaluation of the SMOTE-balanced version of the training data on the untouched test set. The model achieved its best results in terms of recall and F1-score for the deficient class when using the SMOTE configuration while maintaining high precision and AUC.

Table 5. Performance metrics for the Random Forest model on both the imbalanced and SMOTE-balanced datasets.

	Accuracy	**Balanced Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Imbalanced	0.9524	0.9643	1.0000	0.9286	0.9630	0.9898
SMOTE	0.9869	0.9877	0.9765	0.9928	0.9846	0.9972

The confusion matrices (see Figure 9, below) show a significant improvement in the detection of deficient individuals after applying SMOTE, as evidenced by a reduction in false negatives. This improvement was achieved without a significant loss in overall accuracy, which supports the model's suitability for screening applications where missed deficiencies could have clinical consequences.

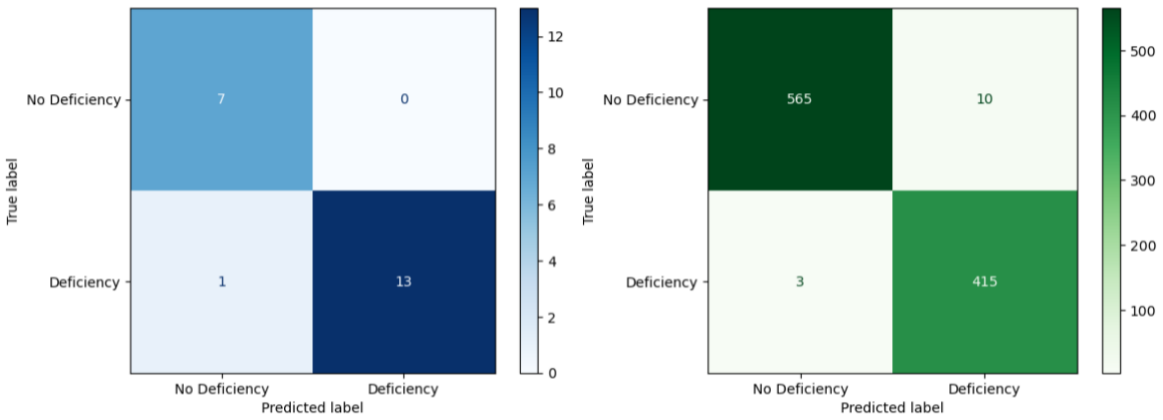


Figure 9: Confusion matrix comparison (Pre-SMOTE vs Post-SMOTE)

6.2 Feature Importance (Global)

The Random Forest Gini importance scores and SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) were used to evaluate global feature importance. The internal scores indicated b12_density and avg_b12 as the most important features but SHAP provided a better explanation framework that considers feature interactions and the model's global behavior.

The SHAP summary plot in Figure 10 shows the top nine features that affect model predictions. The domain knowledge was supported by the strong associations between B12 deficiency and nutrient-specific variables including folate, iron and vegetarian_flag.

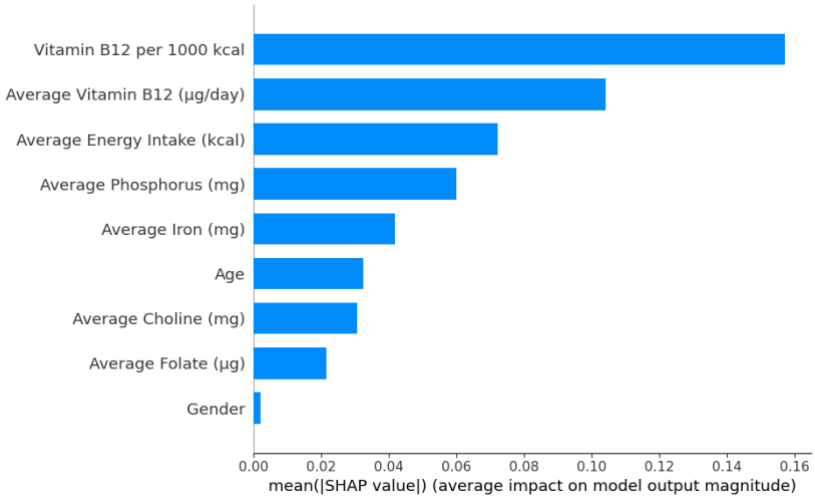


Figure 10: SHAP summary plot (top 9 features)

The SHAP beeswarm plot (Figure 11) was used to analyze how each feature affects individual predictions. The plot shows the direction and distribution of feature effects across all samples and demonstrates complex behaviors like how lower b12_density or avg_b12 values always push predictions toward the deficient class. High and low feature values are colored, revealing class-separating patterns (e.g., low B12 intake in deficient class).

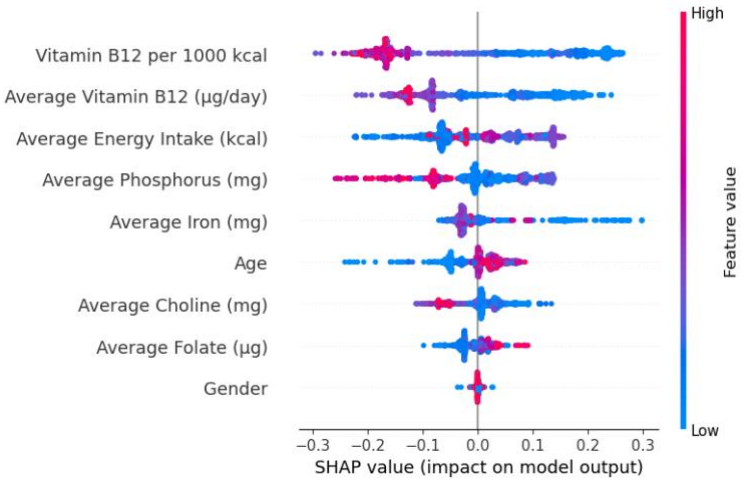


Figure 11: SHAP beeswarm plot showing feature impact distribution across all samples.

6.3 Case-Level Interpretability (Local)

SHAP provides global feature analysis, but LIME (Local Interpretable Model-agnostic Explanations) generates case-specific explanations about model decisions (Ribeiro et al. 2016). The method creates human-understandable explanations through local approximation of model behavior around specific predictions.

The LIME explanations in Figures 12 and 13 demonstrate the model's decision-making process for a B12-deficient person and a non-deficient person. The local feature contributions (e.g., avg_b12, vegetarian_flag, folate) receive positive or negative weights based on their impact on the predicted class in each case.

The model gains additional transparency through localized insights which prove essential for clinical and public health applications because understanding decision reasons matters equally to the decisions themselves.

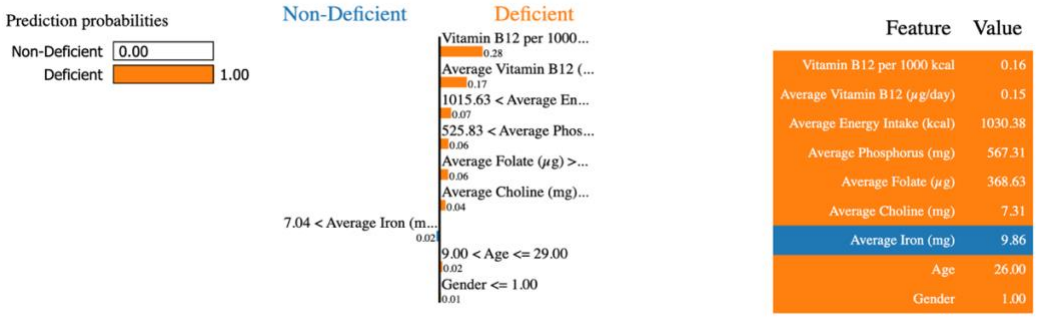


Figure 12: LIME explanation plot for a B12-deficient case.

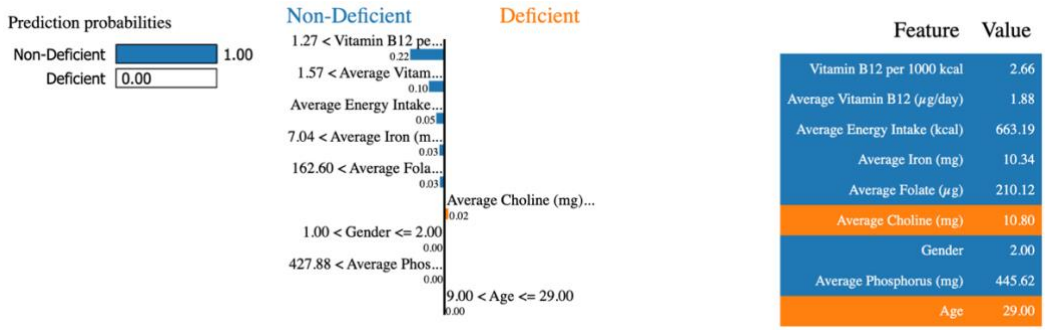


Figure 13: LIME explanation plot for a non-deficient case.

In the deficient case, low B12 density and average intake are key drivers. In contrast, the non-deficient prediction is supported by sufficient B12 and other micronutrient values, which shift the prediction decisively.

6.4 Reflections on Model Behaviour

The evaluation results demonstrate several essential behavioural insights. B12 density proved to be the most important factor in the analysis because it serves as a quality indicator for nutrients (Allen, 2008) and vegetarianism became more significant when combined with low B12 Intake (O'Leary and Samman, 2010). The SHAP explanations revealed that choline and

folate along with other micronutrients made significant contributions to the results. (Clarke et al. 2004).

These findings align with known nutritional science, echoing prior evidence about the dietary correlates of B12 deficiency. The integration of SHAP and LIME ensured transparency remained central throughout model evaluation. Together, the combination of statistical performance and interpretable outputs provides a solid foundation for trustworthy deployment in dietary screening contexts.

6.5 Discussion

The Random Forest model together with SHAP and LIME successfully identified between sufficient and deficient Vitamin B12 intake levels. The leading predictive factors b12_density and avg_b12 confirm established nutritional knowledge about absolute and relative intake importance. The high recall rate on the SMOTE-balanced dataset proves its effectiveness for public health screening because it minimizes false negatives.

However, limitations remain. The use of self-reported dietary information leads to recall bias and the lack of follow-up data prevents researchers from studying long-term effects. The use of SMOTE to handle class imbalance produced synthetic data that could potentially contain noise so researchers should investigate different balancing methods. The exclusion of medication use and gastrointestinal conditions and socioeconomic status variables could decrease the model's ability to detect indirect deficiency risks.

The study findings support existing research about how vegetarian diets and nutrient density levels with multiple micronutrient relationships affect B12 status. Future research should validate findings on separate datasets while integrating real-time dietary tracking tools for developing scalable non-invasive deficiency monitoring systems.

7 Conclusion and Future Work

The research showed that interpretable machine learning models can identify Vitamin B12 deficiency by analysing dietary patterns and demographic information while maintaining high accuracy in predicting vegetarian cases. It also maintained transparency through SHAP and LIME methods which delivered both general and specific explanations. The most important factors that predicted the results were vegetarianism together with folate and nutrient density.

This research depends on self-reported dietary information while omitting specific covariates from its analysis. The integration of data from popular dietary tracking applications such as MyFitnessPal would enable real-time non-invasive deficiency risk monitoring in future studies. The model's generalizability would improve through the addition of lifestyle and medication data and by testing it with different population datasets.

A dashboard that combines SHAP/LIME insights would help clinical and nutritional professionals make better decisions. The proposed pipeline provides a reproducible ethical framework which can be adapted to other micronutrient deficiencies (e.g., iron, folate, vitamin D) to advance explainable AI in public health nutrition.

References

Ahmad, M.A., Eckert, C. and Teredesai, A., 2018. Interpretable machine learning in healthcare. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, pp.447–448. https://doi.org/10.1145/3233547.3233667

Allen, L. H., 2008. Causes of Vitamin B12 and Folate Deficiency. Food and Nutrition Bulletin, 29(2_suppl1), pp.S20–S34. https://doi.org/10.1177/15648265080292S105

Allen, L.H., Miller, J.W., de Groot, L., Hampel, D. and Green, R., 2018. Biomarkers of Nutrition for Development (BOND): Vitamin B-12 review. *The Journal of Nutrition*, 148(suppl_4), pp.1995S–2027S. https://doi.org/10.1093/jn/nxy201

Barocas, S., Hardt, M. and Narayanan, A., 2019. *Fairness and machine learning: Limitations and opportunities*. Cambridge, MA: MIT Press.

Blagus, R. and Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14, p.106. https://doi.org/10.1186/1471-2105-14-106

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

CDC NHANES Documentation. n.d. *Official explanation of your data source and survey design*. [online] https://wwwn.cdc.gov/nchs/nhanes/Default.aspx

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp.321–357. https://doi.org/10.1613/jair.953

Clarke, R. et al., 2004. Vitamin B12 and folate deficiency in later life. *Age and Ageing*, 33(1), pp.34–41. https://doi.org/10.1093/ageing/afg109

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a call for a more informed use. *Ecography*, 36(1), pp.27–46.

Drewnowski, A., 2005. Concept of a nutritious food: toward a nutrient density score2. *The American journal of clinical nutrition*, 82(4), pp.721-732.

Fernández, A., Garcia, S., Herrera, F. and Chawla, N.V., 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, pp.863–905.

Guo, J., He, Q. and Li, Y., 2024. Machine learning-based prediction of vitamin D deficiency: NHANES 2001–2018. *Frontiers in Endocrinology*, 15, p.1327058. https://doi.org/10.3389/fendo.2024.1327058

Henjum, S. et al., 2023. Adequate vitamin B12 and folate status of Norwegian vegans and vegetarians. *British Journal of Nutrition*, 129(12), pp.2076–2083. https://doi.org/10.1017/S0007114522002987

Hunter, J. D., 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), pp.90–95. https://doi.org/10.1109/MCSE.2007.55

Kluyver, T. et al., 2016. Jupyter Notebooks–a publishing format for reproducible computational workflows. In: *Positioning and power in academic publishing: Players, agents and agendas*. IOS press, pp.87-90.

Lemaître, G., Nogueira, F. & Aridas, C. K., 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), pp.1–5.

Lundberg, S.M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp.4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Mineva, E. M. et al., 2021. Fewer US adults had low or transitional vitamin B12 status based on the novel combined indicator of vitamin B12 status compared with individual, conventional markers, NHANES 1999–2004. *The American Journal of Clinical Nutrition*, 114(3), pp.1070–1079. https://doi.org/10.1093/ajcn/nqab122

NIH Office of Dietary Supplements. n.d. *Vitamin B12 fact sheet for health professionals*. [online] https://ods.od.nih.gov/factsheets/VitaminB12-HealthProfessional/

Obeid, R., Fedosov, S. N. & Herrmann, W., 2019. Vitamin B12 intake from animal foods, biomarkers, and health aspects. *Frontiers in Nutrition*, 6, p.93. https://doi.org/10.3389/fnut.2019.00093

O'Leary, F. and Samman, S., 2010. Vitamin B12 in health and disease. *Nutrients*, 2(3), pp.299–316.

Pawlak, R., Lester, S. E. & Babatunde, T., 2014. The prevalence of cobalamin deficiency among vegetarians assessed by serum vitamin B12: A review of literature. *European Journal of Clinical Nutrition*, 68(5), pp.541–548. https://doi.org/10.1038/ejcn.2014.46

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Powers, D.M.W., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pp.37–63. (Reposted 2020 on arXiv:2010.16061).

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York: ACM, pp.1135–1144. https://doi.org/10.1145/2939672.2939778

Samson, M.E., Shrestha, S., He, J., Hao, Y. and Dai, C., 2022. Vitamin B-12 malabsorption and renal function are critical considerations in studies of folate and vitamin B-12 interactions in cognitive performance: NHANES 2011–2014. *The American Journal of Clinical Nutrition*, 116(1), pp.74–85. https://doi.org/10.1093/ajcn/nqac065

Sokolova, M. & Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), pp.427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tamune, H., Aoki, T., Naiki, T., Moriwaki, Y., Naruse, H., Mori, N. and Iino, Y., 2019. Efficient prediction of vitamin B deficiencies via machine-learning using routine blood test results in patients with psychiatric disorders. *Frontiers in Psychiatry*, 10, p.1029. https://doi.org/10.3389/fpsyt.2019.01029

Waskom, M. L., 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), p.3021. https://doi.org/10.21105/joss.03021

Wilcox, R.R., 2012. *Introduction to robust estimation and hypothesis testing*. 3rd ed.