Predicting Healthy BMI

Kyle Dillon

20 December 2025

Synopsis

This report focuses on using data from the National Health and Nutrition Examination Survey (NHANES) on the CDC website to try and predict which adults have a healthy BMI primarily based on their nutrient intake, along with other demographic, laboratory, and questionnaire variables. Three machine learning models were used for this analysis. The findings show that nutrition variables alone may not be the best predictors of BMI, and this causes the models to display weak predictive power when predicting the positive class (Healthy BMI). This is important because it tells us there is more to BMI than just nutrition, and a new approach may be needed.

Introduction

Research Question

Can we predict which adults have a healthy BMI primarily based on their nutrient intake, along with other demographic, laboratory, and questionnaire variables from the National Health and Nutrition Examination Survey? Which nutrients play the biggest role in making this decision?

Background

Obesity is a growing epidemic happening all over the world. According to the World Health Organization (WHO), “In 2022, 1 in 8 people in the world were living with obesity” (World Health Organization, 2025). This gets even more concerning when looking at the United States of America. According to the Centers for Disease Control and Prevention (CDC), “In 2024, all U.S. states and territories had an obesity prevalence of 25% or higher (at least 1 in 4 adults). Overall, the Midwest (35.9%) and South (34.5%) had the highest prevalence of obesity, followed by the West (30.2%) and the Northeast (30.3%)” (Centers for Disease Control and Prevention, 2025). These statistics help to show a glimpse of how obesity is worsening.

Numerous studies have been targeted towards predicting obesity and the causes behind it. However, fewer studies have been conducted on which factors keep people at a healthy BMI. The proposed question can actually provide greater insight and be a more effective way to combat this epidemic. Putting the focus on nutrient intake is important because of the role it plays in health, and it can be changed almost immediately, unlike demographics. Looking at which nutrients are prominent in people who are considered healthy allows organizations, like the Department of Public Health,

to take action by implementing nutrition programs and initiatives to prevent people from becoming overweight or even obese. These programs can lead people to live healthier lifestyles and even save lives.

Hypothesis

Nutrition intake, along with demographic, questionnaire, and laboratory variables, will be able to predict if an adult has a healthy BMI because these variables are indicators of dietary behaviors, lifestyle patterns, and medical history, which affect a person's weight.

Prediction

Adults who have a balanced nutritional intake, a strong lifestyle, and a stable medical history will have a healthy BMI. We predict that a combination of macronutrients, including fat, protein, and carbs, will emerge as some of the variables that have the strongest effect on BMI, as exemplified by the machine learning model's feature importance.

Data

Data Acquisition

The datasets come from the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics (NCHS), found on the Centers for Disease Control and Prevention (CDC) website. The NHANES data is a substantial choice because it contains a variety of different studies with many relevant features to help answer the research question. Two different timeframes have been selected to conduct this analysis. The first timeframe is March 2017 to 2020 (Pre-COVID), and the second is from August 2021 to August 2023. These two

timeframes were selected because they give us enough observations to find meaningful impacts, and they are more recent, meaning the findings will be relevant to today. Variables from different areas of the NHANES dataset, including demographic, dietary, questionnaire, examination, and laboratory data, were used. These variables are listed in the data dictionary section within the appendix of this report. Dietary variables were chosen because they are the main aspect of what we are trying to look at in our research question. Demographic, questionnaire, examination, and laboratory variables were chosen to bring more context, helping to increase model performance. The dataset from March 2017 to 2020 contains 15,560 observations and 410 variables. The dataset from August 2021 to August 2023 contains 11,933 observations and 333 variables. Combined, they contain 27,493 observations and 436 variables.

Outcome Variable

The outcome variable will be BMI. This variable is located in the body measures dataset within the examination section of the NHANES data and is labeled as BMXBMI. It is calculated using the metric system, and the formula is as follows:

BMXBMI = weight (kg) / height (m)2**

A new binary BMI variable was created using a specified threshold. The variable is called BMI_Binary. The threshold for being classified as healthy will be adults with a BMI between 18.5 and 24.9. BMIs under and over this threshold will be classified as unhealthy. Since it is converted to binary, 1 is used to represent healthy observations, and 0 is used to represent unhealthy observations. Missing BMI values were removed. The Healthy threshold was determined by the CDC and can be found on their website.

Caveats

There were some important caveats about the original data collection that need to be mentioned. First, since we are using survey data, the data is self-reported. This means that the observations may not be completely accurate. Second, there is missing data. There are many observations where there are missing values, and values where the person answered with “Don’t know” or “Refused”. This again can have an effect on accuracy. Third, there was only a one-time collection of data. This means we only get a glimpse into these variables, and we cannot see them over a period of time. Finally, there is oversampling of groups. This means the data collected does not proportionally represent the population of the United States.

Cleaning

After the two datasets were combined, a Survey_Year variable was created to identify which observation came from what dataset. Next, we filtered for people over 18 and excluded pregnant women because the question is focused on adults, and pregnant women might skew the results. I then did an initial feature selection of all the variables I thought would be good for the research question. This was followed by converting all missing values, along with “Don’t know” and “Refused”, to the NA class. This was done to help reduce noise. Basic renaming of numerical columns was done to make the data cleaner, and categorical variables that were labeled as numeric were changed to factors, with their categories being specified. Duplicates were checked for, but every row was unique. A function was written to add the total nutritional intake variables with the total supplement use variables to get a new total variable. If one value is NA and the other has a value, the single value was used; otherwise, they were

summed. If both were NA, the total value was NA. The function was then applied to the necessary variables to make the new total variable, and the old variables were removed. This was chosen to capture the full nutritional intake of an adult. I then did an initial variable removal to remove redundant variables, variables that would not be good for modeling, and variables with substantial missing values. These data cleaning steps were taken to create a clean final dataset that is ready for data preprocessing and then followed by modeling.

One assumption with that cleaning is that the same survey was consistent for both of the timeframes. We assume that they are when combining them. We also assume that all the data is recorded correctly, and each observation is unique and correctly matches up with each of the sub-datasets. When removing the missing values for the outcome variable, we assume the missingness was random. A caveat arises when we change “Refused” and “Don’t know” to NA values. We are assuming that they are missing at random, and there is no underlying reason behind it. There could be reasons behind it, like not wanting to answer it out of embarrassment, but we just do not know.

Exploratory Data Analysis

A tibble: 49 × 5

Features <chr>	Value <chr>	Frequency <int>	Percentage <dbl>	Proportion <dbl>
Attempted_Weight_Loss_Past_Year	No	7773	52.16	0.52
Attempted_Weight_Loss_Past_Year	Yes	5931	39.80	0.40
Attempted_Weight_Loss_Past_Year	*NA*	1197	8.03	0.08
BMI	Healthy	5819	39.05	0.39
BMI	Unhealthy	9082	60.95	0.61
Blood_Cholesterol_Meds	No	4464	29.96	0.30
Blood_Cholesterol_Meds	Yes	1705	11.44	0.11
Blood_Cholesterol_Meds	*NA*	8732	58.60	0.59
Country_of_Origin	Born in United States	11171	74.97	0.75
Country_of_Origin	Others	3725	25.00	0.25
Country_of_Origin	*NA*	5	0.03	0.00
Education_Level	9-11th grade(12th with no diploma)	1370	9.19	0.09
Education_Level	College graduate or above	4147	27.83	0.28
Education_Level	High school graduate/GED	3266	21.92	0.22
Education_Level	Less than 9th grade	918	6.16	0.06
Education_Level	Some college or AA Degree	4512	30.28	0.30
Education_Level	*NA*	688	4.62	0.05
Gallstones	No	12673	85.05	0.85
Gallstones	Yes	1525	10.23	0.10
Gallstones	*NA*	703	4.72	0.05
Gender	Female	7813	52.43	0.52
Gender	Male	7088	47.57	0.48
Health_Insurance	No	1921	12.89	0.13
Health_Insurance	Yes	12944	86.87	0.87
Health_Insurance	*NA*	36	0.24	0.00
Liver_Condition	No	13426	90.10	0.90
Liver_Condition	Yes	775	5.20	0.05
Liver_Condition	*NA*	700	4.70	0.05
Marital_Status	Married/Living with partner	8014	53.78	0.54
Marital_Status	Never married	2830	18.99	0.19
Marital_Status	Widowed/Divorced/Seperated	3370	22.62	0.23
Marital_Status	*NA*	687	4.61	0.05
Race	Mexican American	1463	9.82	0.10
Race	Non-Hispanic Black	3051	20.48	0.20
Race	Non-Hispanic White	6625	44.46	0.44
Race	Other Hispanic	1519	10.19	0.10
Race	Other Race	2243	15.05	0.15
Taking_Treatment_Anemia	No	14137	94.87	0.95
Taking_Treatment_Anemia	Yes	740	4.97	0.05
Taking_Treatment_Anemia	*NA*	24	0.16	0.00
Thyroid_Problem	No	12401	83.22	0.83
Thyroid_Problem	Yes	1795	12.05	0.12
Thyroid_Problem	*NA*	705	4.73	0.05
Told_High_BP	No	9394	63.04	0.63
Told_High_BP	Yes	5491	36.85	0.37
Told_High_BP	*NA*	16	0.11	0.00
Told_High_BP_2	No	948	6.36	0.06
Told_High_BP_2	Yes	4511	30.27	0.30
Told_High_BP_2	*NA*	9442	63.36	0.63

Table 1: Frequency, percentage, and proportion of categorical variables.

Table 1 shows the frequency, percentage, and proportion of each category for each variable. We can see the distribution of each of the categorical variables, along with how many NA values each one has.

features <chr>	Mean <dbl>	Median <dbl>	Min <dbl>	Max <dbl>	Sd <dbl>	Skew <dbl>	Kurtosis <dbl>
Age	50.71	53.00	18.00	80.00	18.34	-0.15	1.87
Household_Size	2.62	2.00	1.00	7.00	1.48	1.02	3.52
Family_Income_to_Poverty_Ratio	2.71	2.43	0.00	5.00	1.65	0.15	1.60
Total_Energy	2048.98	1909.00	0.00	12501.00	1002.05	1.47	9.39
Total_Protein	77.34	69.59	0.00	545.20	41.84	1.75	10.14
Total_Carbs	235.93	216.21	0.00	1586.24	124.05	1.58	9.59
Total_Sugar	101.70	86.58	0.00	953.90	73.55	2.33	14.90
Total_Fiber	16.30	14.10	0.00	127.30	10.60	1.76	9.32
Total_Fat	85.12	77.01	0.00	567.96	48.80	1.62	9.36
Total_Sat_Fat	27.38	24.07	0.00	268.59	17.42	1.92	12.13
Total_Mono_Sat_Fat	29.17	25.82	0.00	205.38	17.54	1.75	9.98
Total_Poly_Sat_Fat	20.06	17.22	0.00	218.70	13.71	2.38	17.68
Total_Cholesterol	311.81	236.00	0.00	3598.00	259.67	1.94	10.32
Total_Lycopene	4494.31	1353.00	0.00	190742.00	8567.87	5.22	56.31
Total_Lutein_Zeaxanthin	1770.63	802.00	0.00	277797.00	4347.36	23.93	1299.40
Total_Thiamin	4.50	1.59	0.00	403.08	15.70	8.95	118.61
Total_Riboflavin	3.76	1.95	0.00	405.41	10.66	17.74	527.64
Total_Niacin	32.53	24.99	0.00	2036.16	52.62	17.99	487.07
Total_VitaminB6	4.49	2.08	0.00	503.11	13.53	17.10	487.51
Total_Folic_Acid	299.31	169.00	0.00	16084.00	382.64	9.10	269.40
Total_Folate_DFE	712.48	510.00	0.00	27525.00	676.08	8.16	231.19
Total_Choline	328.18	283.40	0.00	2749.60	204.46	1.75	10.03
Total_VitaminB12	113.08	5.22	0.00	10003.91	519.74	8.13	87.25
Total_VitaminC	173.05	81.78	0.00	30089.10	412.36	33.93	2208.06
Total_VitaminK	131.91	80.10	0.00	22942.10	319.96	45.25	2927.15
Total_VitaminD	25.09	7.07	0.00	2575.00	59.66	15.14	422.53
Total_Calcium	1000.54	880.00	0.00	9632.70	633.77	1.69	9.93
Total_Phosphorus	1296.85	1184.00	0.00	8226.80	686.89	1.63	9.51
Total_Magnesium	320.67	281.00	0.00	3653.00	191.32	2.25	16.71
Total_Iron	16.42	12.62	0.00	227.16	14.97	4.37	34.99
Total_Zinc	14.56	10.91	0.00	477.53	14.39	7.64	169.90
Total_Copper	1.36	1.12	0.00	43.25	1.11	10.62	302.71
Total_Sodium	3280.37	2946.00	0.00	25949.00	1803.93	1.82	11.33
Total_Potassium	2483.24	2291.00	0.00	14358.00	1294.07	1.43	8.14
Total_Selenium	118.61	103.72	0.00	2181.60	75.90	3.74	58.94
Total_Caffeine	142.75	96.00	0.00	4320.00	194.31	5.67	74.04
Vitamin_E_Alpha_Tocopherol	9.15	7.52	0.00	138.55	6.92	3.01	24.30
Added_Alpha_Tocopherol	0.83	0.00	0.00	81.64	3.62	7.63	91.95
Retinol	376.61	289.00	0.00	19374.00	448.42	13.89	442.69
Vitamin_A	586.24	452.00	0.00	20419.00	600.76	7.51	161.26
Alpha_Carotene	383.21	38.00	0.00	27509.00	1038.51	6.89	90.32
Beta_Carotene	2280.49	787.00	0.00	71772.00	4242.99	5.18	45.11
Beta_Cryptoxanthin	89.78	26.00	0.00	7381.00	229.94	11.50	243.60
Total_Folate	350.32	306.00	0.00	3752.00	217.96	2.21	14.88
Food_Folate	207.73	176.00	0.00	2064.00	141.25	2.35	14.73
Added_Vitamin_B12	0.75	0.00	0.00	78.34	2.81	8.47	119.87
Theobromine	35.90	0.00	0.00	1188.00	71.76	4.43	36.87
Alcohol	8.70	0.00	0.00	1152.80	27.03	10.35	288.58
Moisture	2878.65	2604.56	0.00	29329.36	1503.07	2.21	17.85
Total_Plain_Water	1236.00	987.00	0.00	26879.61	1223.63	2.77	27.61
Iodine	119.25	150.00	0.50	12500.00	258.10	39.26	1761.83
Plasma_Fasting_Glucose	111.37	102.00	47.00	561.00	35.93	4.19	26.82
Insulin	14.64	9.75	0.35	551.10	23.47	11.23	186.92
Glycohemoglobin	5.81	5.50	2.80	17.10	1.11	3.49	20.29
Sleep_Hours_Weekdays	7.64	8.00	2.00	14.00	1.64	-0.03	4.35
Sleep_Hours_Weekends	8.28	8.00	2.00	14.00	1.76	-0.05	4.05
Minutes_Moderate_LTPA	63.40	60.00	1.00	720.00	59.68	3.79	26.45
Minutes_Vigorous_LTPA	60.47	45.00	1.00	900.00	60.97	4.78	43.36
Minutes_Sedentary_Activity	346.34	300.00	0.00	1380.00	205.39	0.84	3.55

Table 2: Mean, median, minimum, maximum, standard deviation, skewness, and kurtosis for numerical variables

Table 2 shows the Mean, median, minimum, maximum, standard deviation, skewness, and kurtosis for numerical variables. This helps to see how the numerical variables are distributed and if there are any outliers that need to be dealt with.

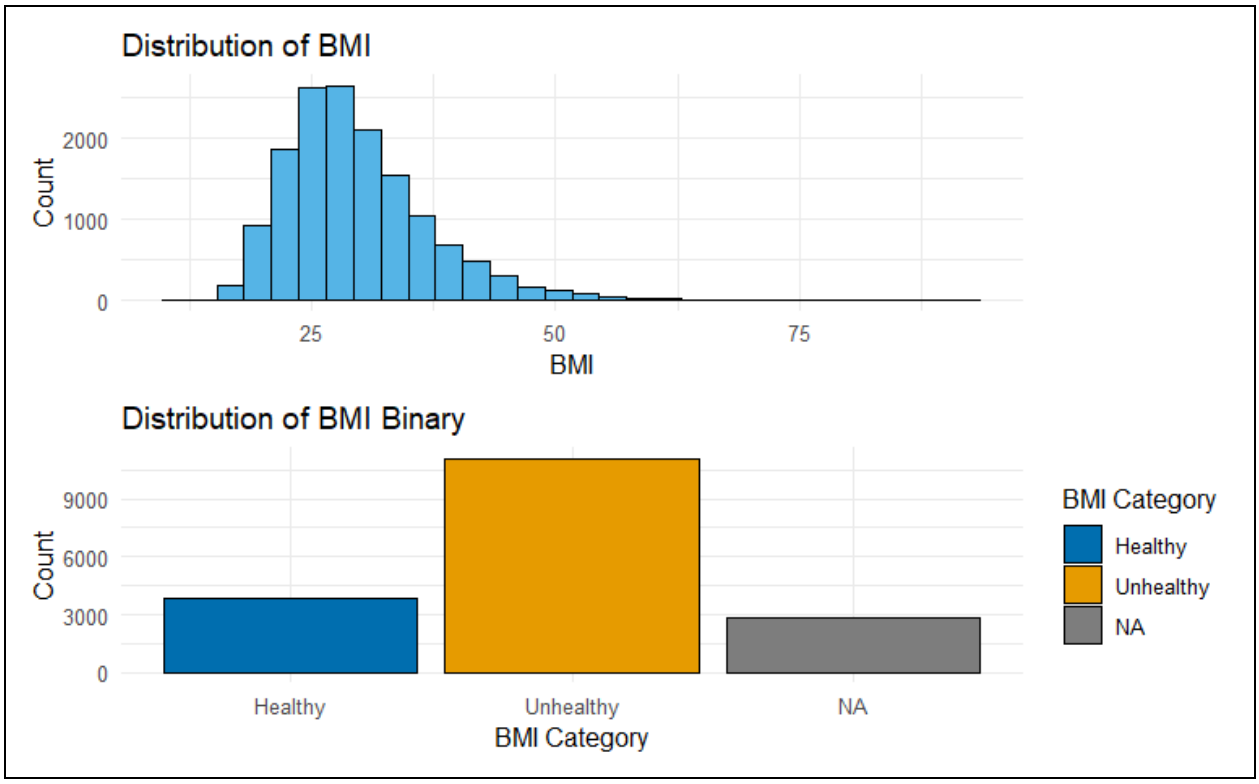

Figure 1: Distribution of the numeric BMI variable and the binary BMI variable

Figure 1 above looks at the distribution of values of the numeric BMI variable BMXBMI and the binary BMI variable. Most of the BMI values seem to be between 20 and 40 in the numeric BMI graph. In the binary BMI graph, we can see that more adults are classified as unhealthy than healthy.

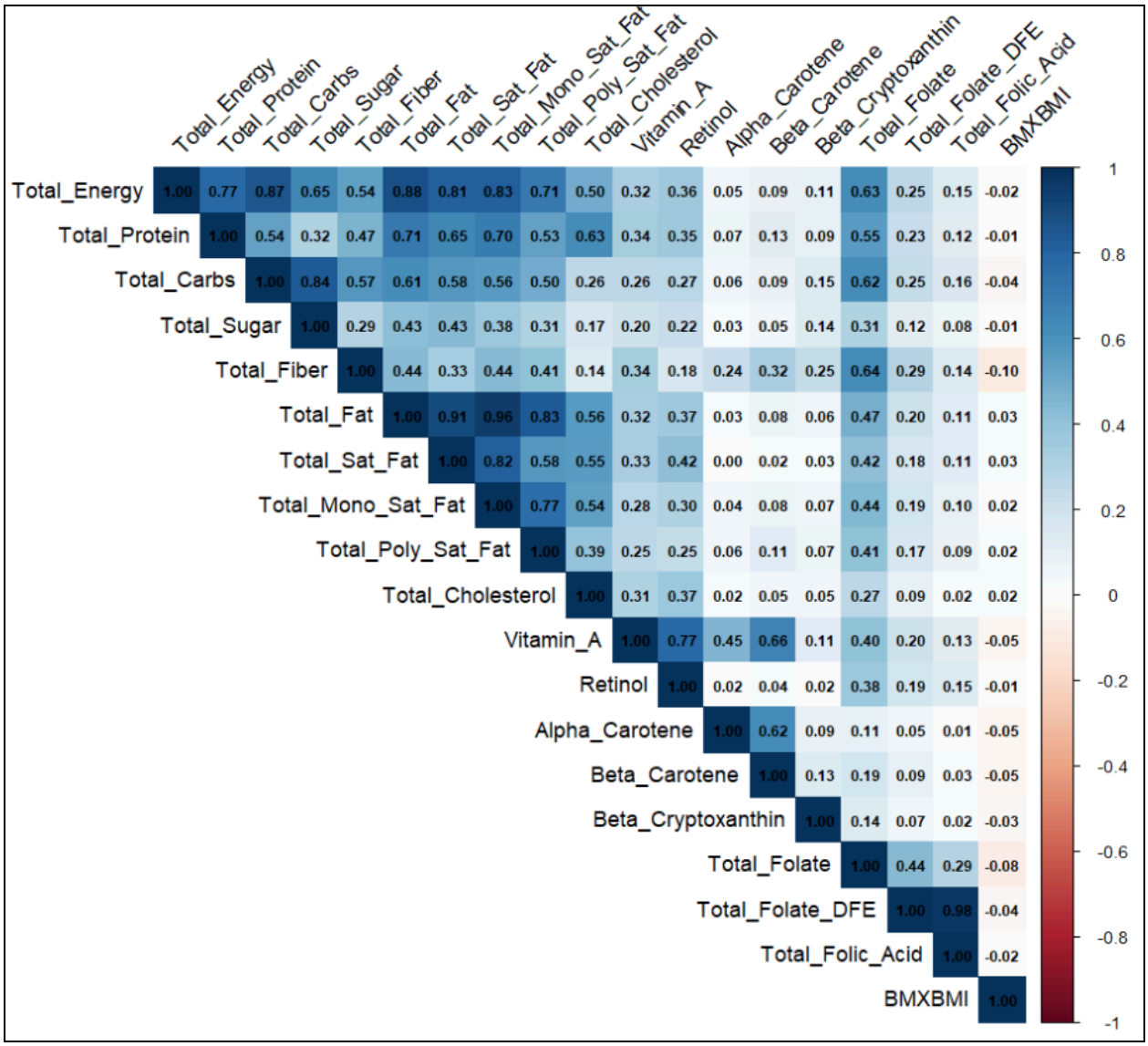

Figure 2: Corrplot with micro and macro nutrients

Figure 2 displays an initial correlation graph to see how correlated some macro and micro nutrients are. Some examples of variables that are highly correlated include Total_Folate_DFE and Total_Folic_Acid, as well as Total_Fat and Total_Mono_Sat_Fat. This is good to know because this brings up multicollinearity concerns and is something that needs to be addressed before modeling.

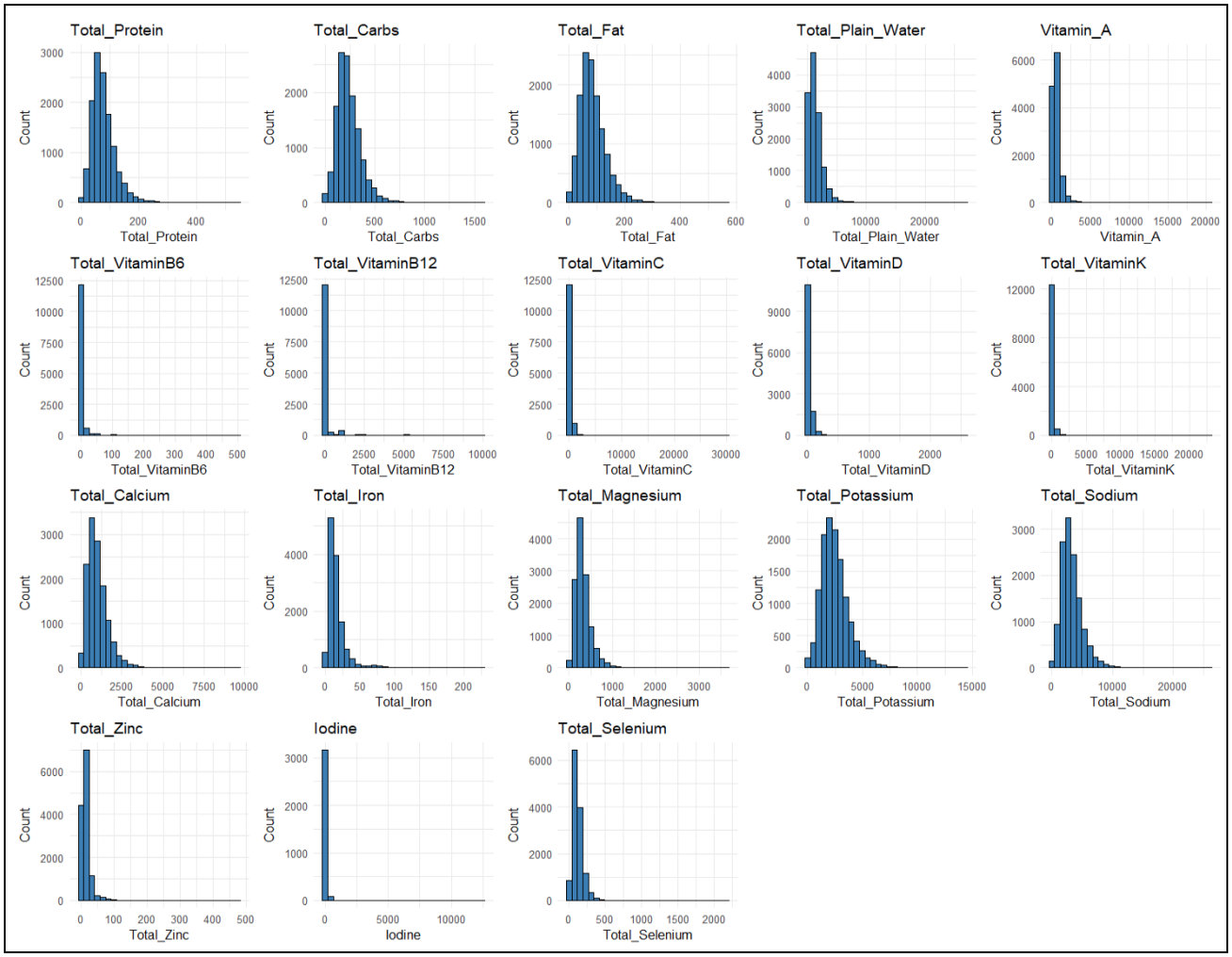

Figure 3: Distribution of some Micro and Macro nutrients

Figure 3 shows numerous graphs that display the distribution of various micro and macro variables. These graphs help us see where most of the values are, as well as which features have outliers that need to be handled so they do not affect the models.

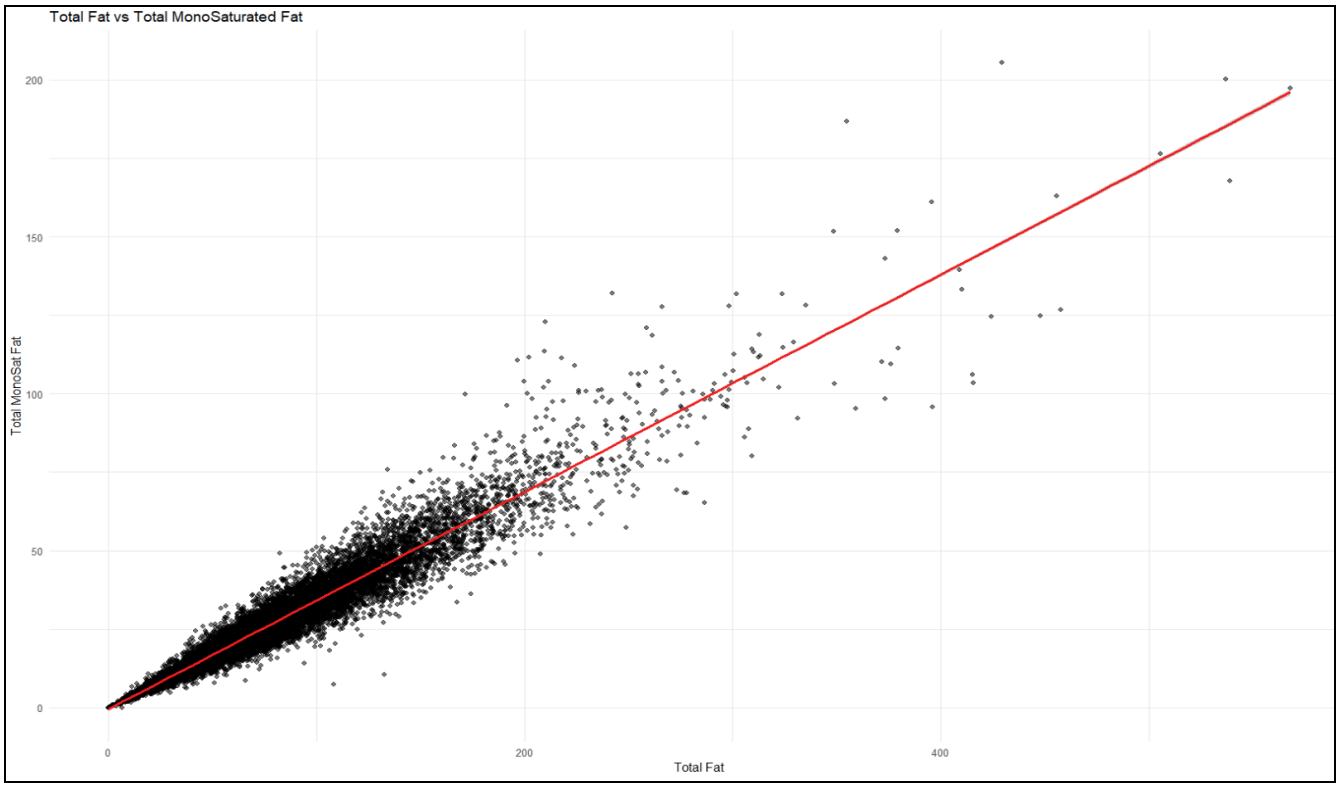

Figure 4: Scatter plot showing Total Fat vs Total Monounsaturated Fats

Figure 4 displays a scatter plot of Total Fat vs Total Monounsaturated Fats. Total Fat is on the x-axis, and Total Monounsaturated Fats is on the y-axis. Looking at the graph, we can see that there seems to be a strong correlation between these two variables, meaning multicollinearity might be the case here. This will hurt the model's performance, so it is something that needs to be addressed.

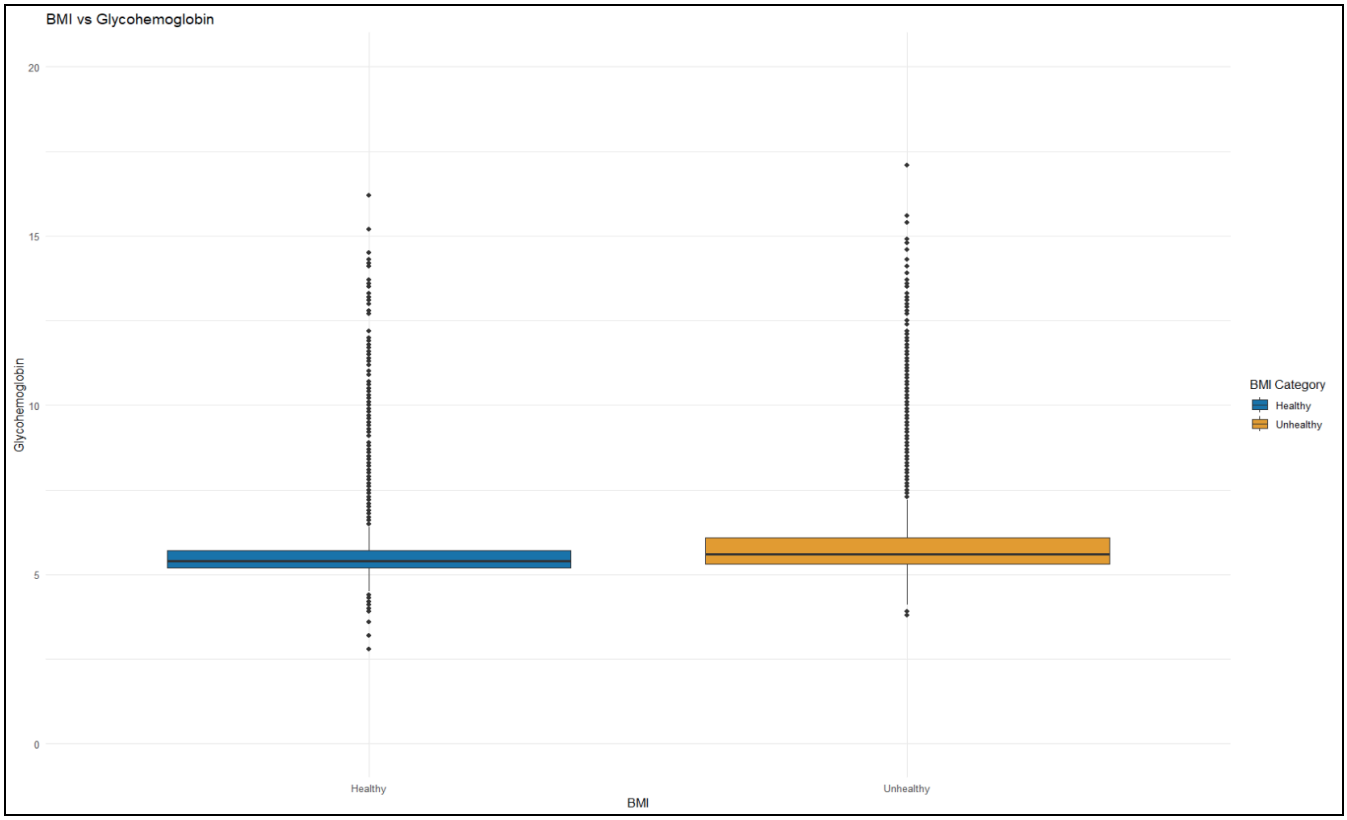

Figure 5: Box plot showing the distribution of Glycohemoglobin compared to BMI

Figure 5 displays a box plot with binary BMI on the x-axis and Glycohemoglobin on the y-axis. We can see that the Glycohemoglobin is higher for people who are unhealthy and lower for healthier people. This may indicate that Glycohemoglobin is a good predictor of BMI.

Models

Pre-processing and Dimensionality Reduction

Data pre-processing is an essential part of the pre-modeling processes. These steps were taken to help enhance the performance of the models. First, a train/ test split calculation was used to find the optimal train test split ratio. With the size of my dataset and the number of features provided, the train/ test ratio I was provided with was 0.88:0.12. Considering this, I decided to go with an 80/20 split. Once I established the ratio, the initial_split() function within the tidymodels framework was utilized to set my training and testing sets. I also stratified by the target variable (BMI_Binary). Next, categorical variables were dummy encoded using fastDummies. This is important because it allows variables with categories to be used by the machine learning models. Data imputation was conducted with the use of multiple imputation chained equations (MICE). We assume that the missing values in the survey data are missing at random, and MICE can perform well under this assumption. After inputting, a final correlation matrix was created to observe variables that have a high correlation. This can be seen in Figure 16 in the appendix. Variables with high correlation were examined and appropriately removed. The variables removed include Total Folic Acid, Total Energy, Total_Fat, total Carbs, Total Choline, Total Phosphorus, and Food Folate. These variables were removed because the ones chosen instead provide better insight, which helps to answer the research question. This was done to reduce the risk of multicollinearity, which can negatively affect a model's performance.

One of the last steps taken was setting the recipe for the models. The recipe set BMI_Binary as the target with all other features as predictors. Step_rm was used to

remove Survey_year from the recipe. This was done because Survey_Year serves as a variable to differentiate the datasets the observations came from, and not as a predictive component. Step_log was set for selected variables with high skewness. As seen in Table 2 of the EDA, there were many variables with high skewness, so this is put in place to try and combat that. Step_normalize was set for all numeric predictors to set the means to 0 and the standard deviations to 1. This is to prevent features with large ranges from controlling the model. Step_smote and step_adasyn were used individually to help with class imbalance. Finally, a PCA test was conducted to inspect dimensionality. The PCA test identified 63 PCs, with the first PC making up 14 percent of the variance. The top 41 PCs make up 90 percent of the variation. This can be seen in Figure 15 in the appendix. Below is a SCREE plot to help visualize the variance each of the top PCs holds.

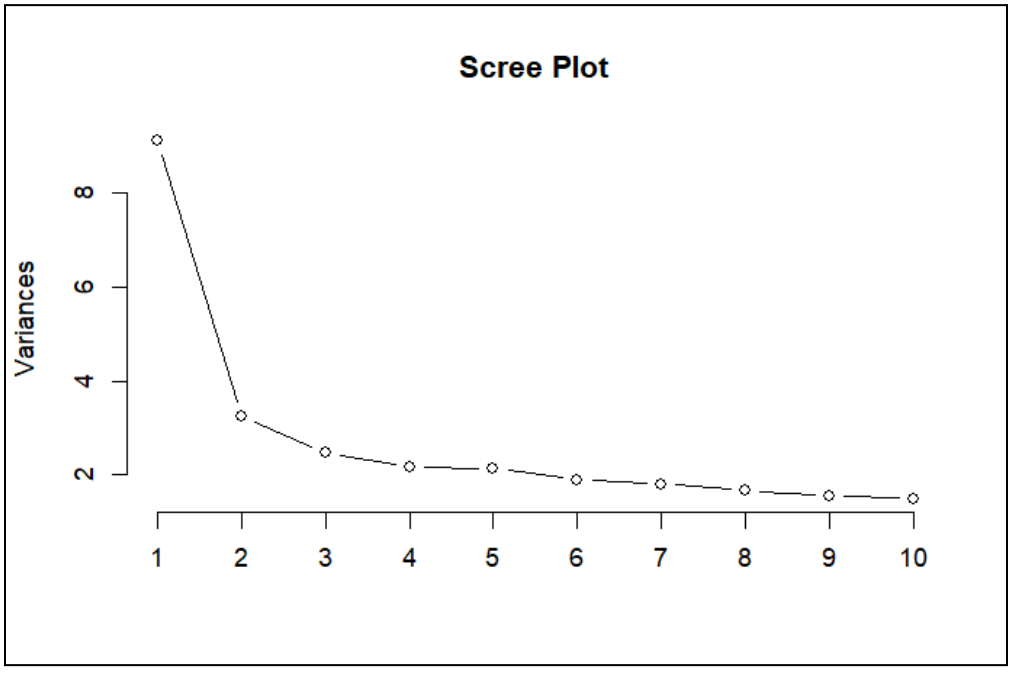

Figure 6: SCREE Plot for PCA

Algorithms Selection

Three supervised learning algorithms have been chosen to conduct this analysis. The three models include Logistic Regression, Random Forest, and XGBoost. Logistic Regression was chosen as a baseline model to compare the performance of the other models to it. Given that our target variable is heavily imbalanced, we expect accuracy and recall to perform poorly. This being said, we will look at other metrics like PR-AUC and F1 score to get the full picture. Random Forest was chosen because of its ability to handle non-linear relationships, help reduce overfitting, and be robust to feature interactions. Finally, XGBoost was chosen because of the potential gain in performance metrics from tuning and the advantages of boosting. Boosting is where each new model that is made learns from the errors of the old ones. This allows for better accuracy, better handling of complex relationships, and better handling of misrepresented cases, which we have.

Five-fold stratified cross-validation was applied to the training set to help diagnose overfitting and observe the effect of class imbalance on model performance. Cross-validation was stratified using the target variable (BMI_Binary). All three of the models used tune and tune grid to find the optimal hyperparameters to allow for the best performance metrics. Every hyperparameter combination is applied to each fold, where it then gets averaged out from the five folds, and the best metrics are displayed. We then use the best hyperparameters to test on unseen data. For the Logistic Regression model, penalty and mixture were used as hyperparameters. The engine was set to glmnet, and the mode was classification. For the Random Forest model, trees, mtry, and min_n were used as hyperparameters. Within the engine, ranger was set, and

importance was set to impurity. The importance is to evaluate the features with the most predictive power after running the test set. The mode was set to classification. Finally, for the XGBoost model, trees, tree_depth, learn_rate, loss_reduction, sample_size, and mtry were used as hyperparameters. The engine was set to xgboost, and the mode was set to classification.

Each of the three models have assumptions that need to be addressed and tested for. For Linear regression, one assumption is that observations are independent of one another. We assume that each observation in the dataset is an individual person, and an individual was not sampled more than once. One thing to keep in mind is that we are using datasets from two different timeframes, so the same person could have been sampled for both time periods, but with two different unique identifiers. This is something we just can't know. A second assumption is that there is no strong collinearity between variables. This was tested using the correlation plot seen in Figure 16 in the appendix. A third assumption is that there is a sufficient data size. This was tested and acted upon by adding more observations from another timeframe. A fourth assumption is that the relationship between the log odds of the outcome variable and the independent variables needs to be linear. Random Forest and XGBoost models have the same assumptions except for the log odds assumption.

To control for overfitting, cross-validation metrics were compared to the test set metrics. Cross-validation and test set metrics were similar, meaning the models generalize well.

Final Models

Logistic Regression

Cross-Validation Metrics

Method	Balanced Accuracy	Recall	Specificity	Percision	F1 Score	PR_AUC
None (baseline)	0.525	0.074	0.976	0.511	0.129	0.407
SMOTE	0.631	0.660	0.602	0.362	0.468	0.406
ADASYN	0.625	0.668	0.582	0.354	0.462	0.400

Table 3: Logistic Regression cross-validation metrics using different class imbalance techniques. The results are the average across five folds of the best hyperparameter combination.

Test Set Metrics

Method	Balanced Accuracy	Recall	Specificity	Percision	F1 Score	PR_AUC
None (baseline)	0.5	0	1	NA	NA	0.628
SMOTE	0.5	0	1	NA	NA	0.628
ADASYN	0.5	0	1	NA	NA	0.628

Table 4: Logistic Regression test set metrics of different class imbalance techniques using the best hyperparameter combination

Confusion Matrix

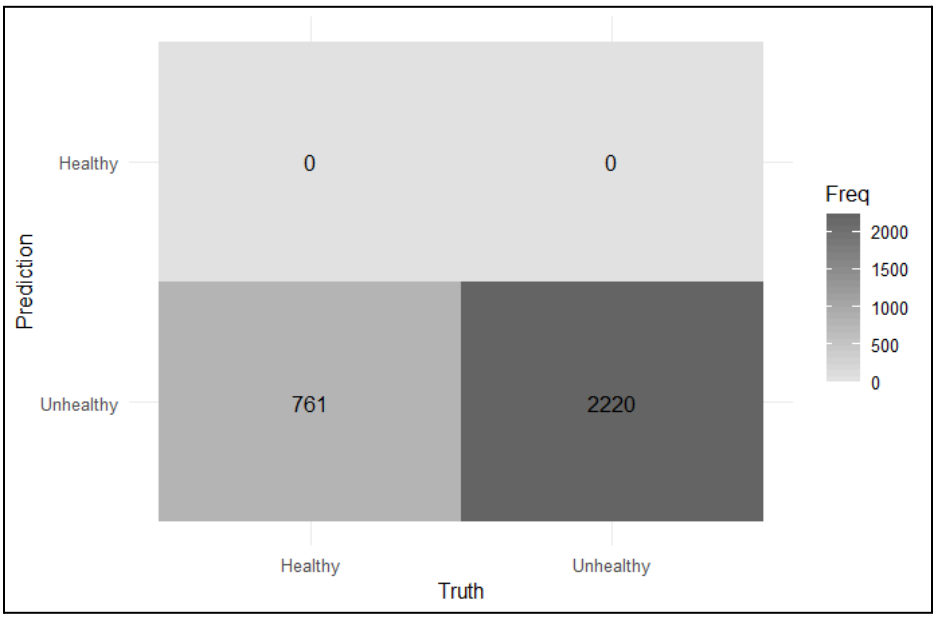

Figure 7: Confusion matrix for Logistic Regression. The results were the same for all of the methods.

ROC Curve

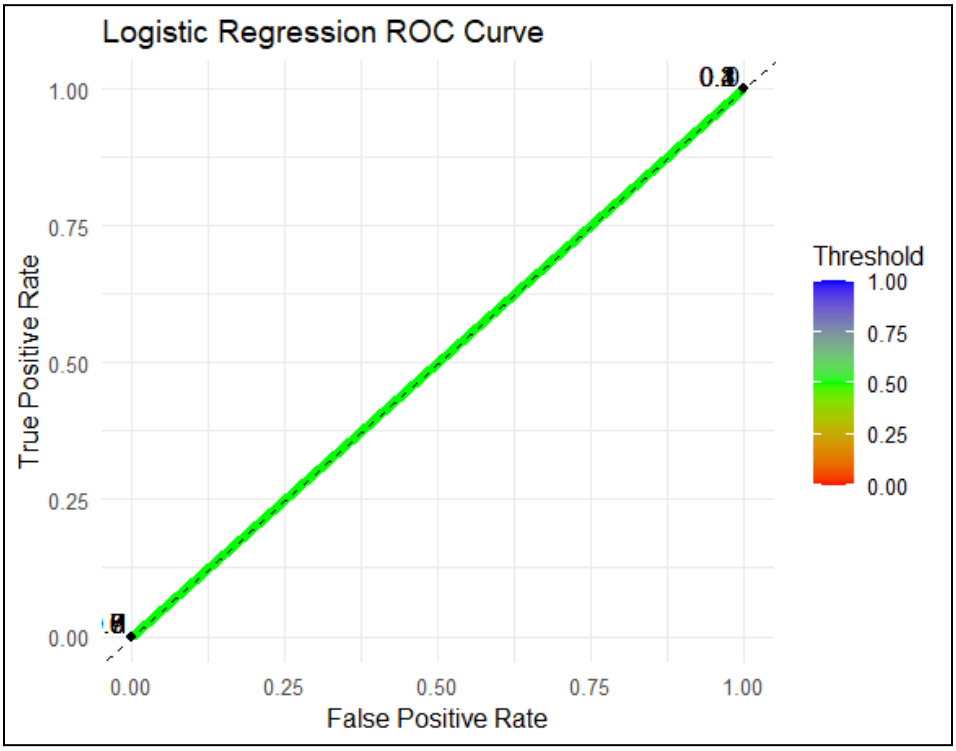

Figure 8: ROC Curve for Logistic Regression. The results were the same for all of the methods.

Overview of Logistic Regression

Logistic Regression showed some signs of decent performance on the cross-validation set using class imbalance techniques, but failed to learn anything at all. As seen in Table 4, recall, precision, and F1 are all 0 or NA, indicating that the model did not predict any of the majority class. This is reinforced in Figures 7 and 8, with 0 for the minority classes and an ROC line that follows the random guessing line. The model is essentially just predicting the majority class for every observation.

Random Forest

Cross-Validation Metrics

Method	Balanced Accuracy	Recall	Specificity	Percision	F1 Score	PR_AUC
None (baseline)	0.501	0.001	0.1.00	0.722	0.004	0.394
SMOTE	0.568	0.239	0.898	0.446	0.311	0.398
ADASYN	0.563	0.215	0.910	0.451	0.291	0.395

Table 5: Random Forest cross-validation metrics using different class imbalance techniques. The results are the average across five folds of the best hyperparameter combination.

Test Set Metrics

Method	Balanced Accuracy	Recall	Specificity	Percision	F1 Score	PR_AUC
None (baseline)	0.507	0.023	0.991	0.462	0.045	0.416
SMOTE	0.573	0.255	0.891	0.445	0.324	0.401
ADASYN	0.560	0.219	0.901	0.433	0.291	0.398

Table 6: Random Forest test set metrics of different class imbalance techniques using the best hyperparameter combination

Confusion Matrix SMOTE

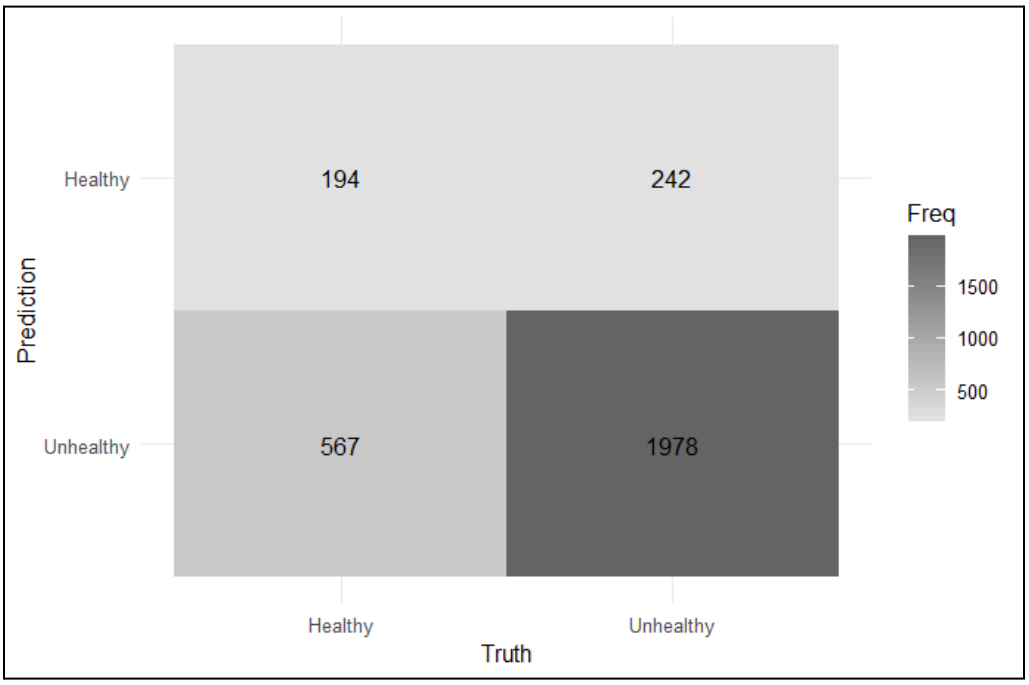

Figure 9: Confusion matrix for Random Forest.

ROC Curve SMOTE

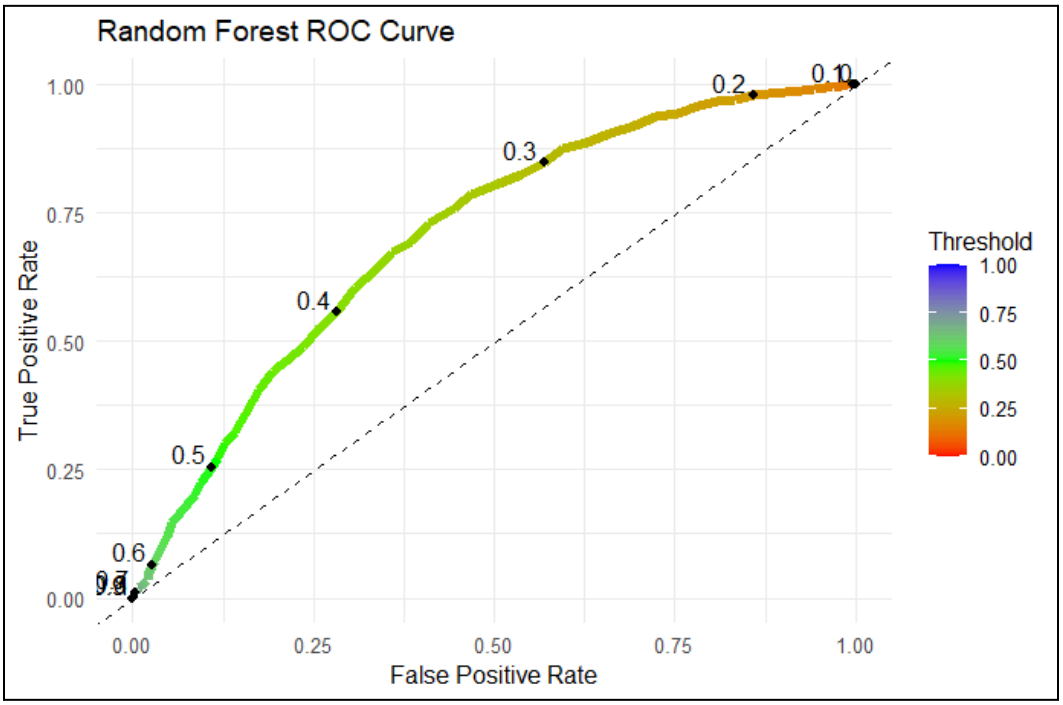

Figure 10: ROC Curve for Random Forest.

Feature importance SMOTE

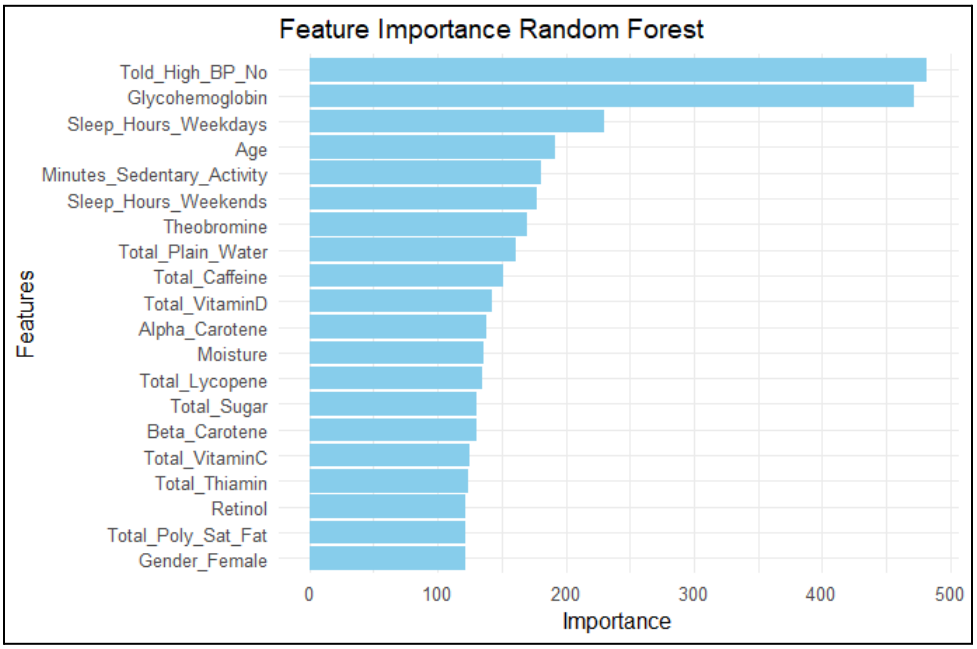

Figure 11: Feature importance for Random Forest model

Overview of Random Forest

When looking at Tables 5 and 6, it can be concluded that SMOTE was the best class imbalance technique used when compared to the other methods. It performed much better than the baseline and performed slightly better than ADASYN. It performed consistently from the cross-validation to the test set. Recall and precision show that it can identify the minority class, but not at a high level. This is reinforced by the confusion matrix and ROC curve. Looking at Figure 11, we can see that nutrients do not have a strong signal when predicting BMI, and demographic, laboratory, and questionnaire variables have the most importance. The confusion matrix and ROC curves for the baseline and ADASYN can be found in the appendix.

XGBoost

Cross-Validation Metrics

Method	Balanced Accuracy	Recall	Specificity	Percision	F1 Score	PR_AUC
None (baseline)	0.5	0	1	NA	NA	0.412
SMOTE	0.613	0.518	0.708	0.378	0.437	0.397
ADASYN	0.612	0.519	0.705	0.375	0.435	0.393

Table 7: XGBoost cross-validation metrics using different class imbalance techniques. The results are the average across five folds of the best hyperparameter combination.

Test Set Metrics

Method	Balanced Accuracy	Recall	Specificity	Percision	F1 Score	PR_AUC
None (baseline)	0.536	0.101	0.972	0.55	0.171	0.444
SMOTE	0.61	0.306	0.895	0.501	0.380	0.425
ADASYN	0.59	0.292	0.889	0.473	0.361	0.416

Table 8: XGBoost test set metrics of different class imbalance techniques using the best hyperparameter combination

Confusion Matrix SMOTE

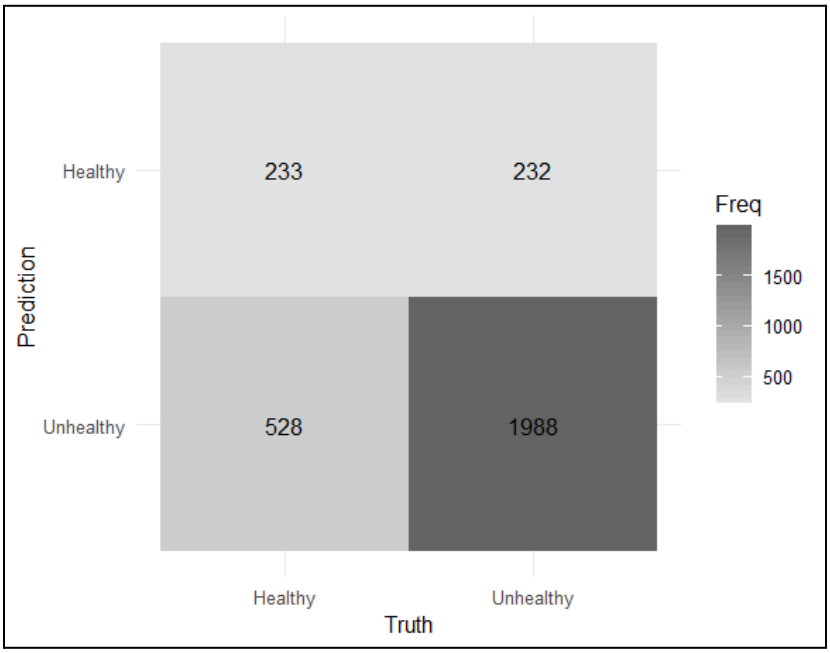

Figure 12: Confusion matrix for XGBoost.

ROC Curve SMOTE

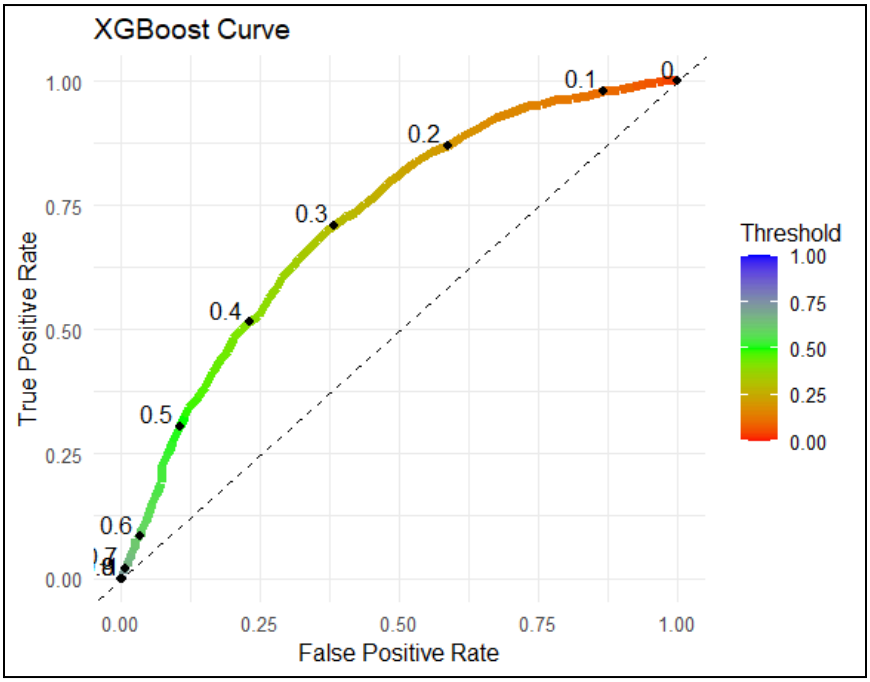

Figure 13: ROC Curve for XGBoost.

Feature Importance SMOTE

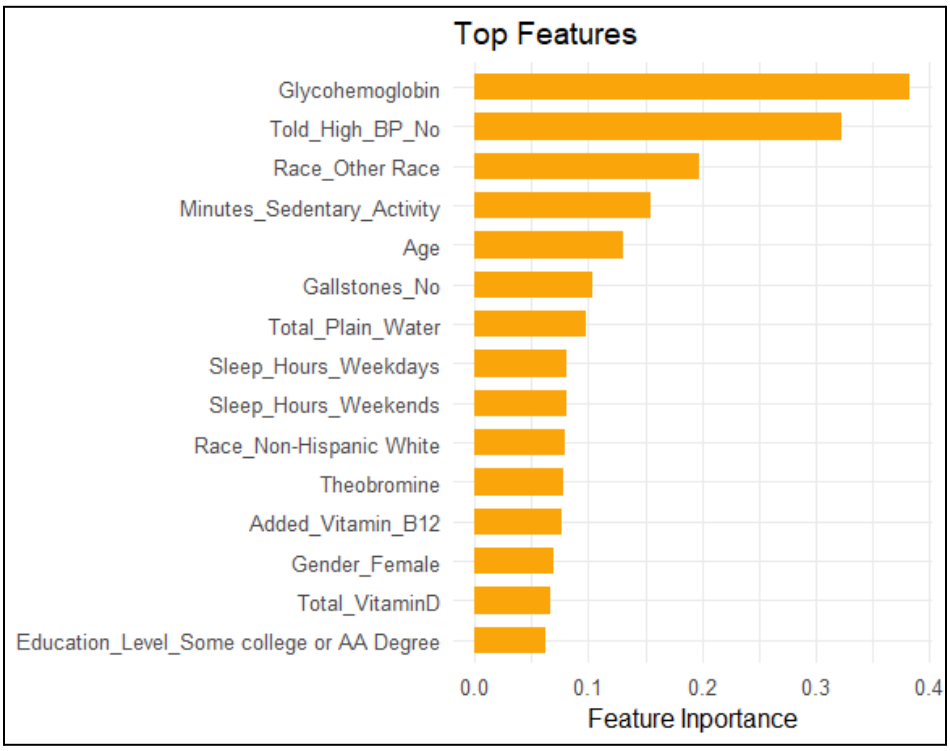

Figure 14: Feature importance for XGBoost model.

Overview of XGBoost

When looking at Tables 7 and 8, it can be concluded that SMOTE was again the best class imbalance technique used compared to the other methods. Like the Random Forest model, it performed much better than the baseline and slightly better than ADASYN. Recall and precision show us that it can identify the minority class, but not at a high level. The confusion matrix reinforces this, with 233 true positives and 232 false positives. The model has a hard time identifying the minority class. Looking at Figure 14, we can see that the top features are not nutrients, and it seems that nutrients do not play that big a role in predicting BMI. The confusion matrix and ROC curves for the baseline and ADASYN can be found in the appendix.

Best Model

Comparing SMOTE from Random Forest and XGBoost

Metric	Random Forest	XGBoost
Balanced Accuracy	0.573	0.61
Recall	0.255	0.306
Specificity	0.891	0.895
Precision	0.445	0.501
F1 Score	0.324	0.380
PR_AUC	0.401	0.425

Table 9: Comparing the best class imbalance method (SMOTE) from Random Forest and XGBoost to see the best model

Overview of Best Model

To preface, Logistic Regression has not been included in this analysis due to the model having little to no predictive power. That being said, looking at Table 9, we can see that XGBoost using SMOTE has better results for every performance metric, making it the best model. However, it still has a limited ability to predict the minority class.

Decisions and Next Steps

Key Takeaways

This analysis attempted to answer the proposed question of whether primarily nutrient, along with demographic, laboratory, examination, and questionnaire variables, could predict whether an adult has a healthy BMI, and which nutrients cause this. After completing this analysis, many key takeaways can be concluded. First, the model that performed the best was the XGBoost model using SMOTE. The recall, precision, and F1 score tell us that it has some predictive power when predicting healthy BMI. However, this power is very limited. The models seemed to have a tough time identifying the minority class, which could be due to many reasons. Class imbalance and features with limited signal could be a couple of reasons. Looking at feature importance, we can conclude that nutrients play a small role in predicting BMI, with demographic, laboratory, examination, and questionnaire variables being more important. The findings disprove my hypothesis and indicate that the variables chosen can not reliably predict healthy BMI. The findings also prove my prediction wrong, indicating that nutrients have very limited predictive power and may not be the best to predict alone.

Recommendations and Future Decisions

Many recommendations can be made for the future. Even though the models had limited power in predicting healthy BMI, we do think we should move forward with this, but with a different approach. Instead of using primarily nutrient variables, which we have seen do not have the strongest predictive power, we would recommend shifting the focus to more demographic, laboratory, examination, and questionnaire variables.

We also think adding more observations from previous years would help with things like class imbalance. After adding more observations and switching the approach, another thing we would recommend is trying more class imbalance techniques, retuning hyperparameters, and trying new models. Even though Logistic Regression showed little to no predictive power, the other two models, using class imbalance techniques, showed some sign of being able to predict the positive class. This leaves us with some hope, and with the new approach, the findings could be a lot different.

Caveats and Concerns

Many caveats have already been explained in the caveats section within the data section of this report. To recap, it talked about the caveats of the original data. This includes self-reported data, missing data, one-time collection of data, and oversampling of groups. However, there are some cautions and limitations of the completed analysis. One is that, even though the models have limited predictive power, the results should not be used to assess a single person's health. It should be used at a population level to look at patterns and trends. To go along with this, the results are predictions, meaning we can not conclude whether a person is healthy or unhealthy. We are just looking at likelihood. Another limitation is that everyone is different. The same variables may not be indicators of healthy or unhealthy for everyone. Also, some people work out, eat a balanced diet, and take care of themselves, but may not fit within the healthy BMI range. Another caution is that this analysis focuses primarily on nutrients, but there are numerous different factors that make up BMI. Finally, as explained throughout this report, the models displayed limited predictive power.

Appendix

Github Repo: https://github.com/Dillonkw/Data-Science-Capstone

Sources:

“Adult BMI Categories.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/bmi/adult-calculator/bmi-categories.html. Accessed 21 Dec. 2025.

“Adult Obesity Prevalence Maps.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/obesity/data-and-statistics/adult-obesity-prevalence-maps.html. Accessed 21 Dec. 2025.

“Nhanes Questionnaires, Datasets, and Related Documentation.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, wwwn.cdc.gov/nchs/nhanes/. Accessed 21 Dec. 2025.

“Obesity and Overweight.” *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Accessed 21 Dec. 2025.

Code Book

Variable Name	Type	Description
BMI_Binary	Binary	(Target Variable)
Demographic Variables		
RIDAGEYR	int	Age in years at screening
RIAGENDR	factor	Gender
RIDRETH1	factor	Race/Hispanic origin
DMDEDUC2	factor	Education level
DMDBORN4	factor	Country of birth
DMDEDUC2	factor	Marital status
DMDHHSIZ	int	Total number of people in the household
INDFMPIR	con	Ratio of family income to poverty
Dietary (Total Nutrition Intake)		
DR1TKCAL	cont	Energy (kcal)
DR1TPROT	cont	Protein (gm)
DR1TCARB	cont	Carbohydrate (gm)
DR1SUGR	cont	Total Sugar (gm)
DR1TRIBE	cont	Dietary fiber (gm)
DR1TTFAT	cont	Total fat (gm)
DR1TSFAT	cont	Total saturated fatty acid (gm)
DR1TMFAT	cont	Total monounsaturated fatty acid (gm)
DR1TPFAT	cont	Total polyunsaturated fatty acid (gm)
DR1TCHOL	cont	Cholesterol (mg)
DR1TLYCO	cont	Lycopene (mcg)
DR1TATOA	cont	Added alpha tocopherol

DR1TRET	cont	Retinol
DR1TTLZ	cont	Lutein + Zeaxanthin (mcg)
DR1TATOC	cont	Vitamin E (mg)
DR1TVARA	cont	Vitamin A, RAE (mcg)
DR1TVB1	cont	Thiamin (Vitamin B1) (mg)
DR1TVB2	cont	Riboflavin (Vitamin B2) (mg)
DR1TNIAC	cont	Niacin (mg)
DR1TVB6	cont	Vitamin B6 (mg)
DR1TFDFE	cont	Folate, DFE (mcg)
DR1TFA	cont	Folic acid (mcg)
DR1TFF	cont	Food folate (mcg)
DR1TCHL	cont	Total choline (mg)
DR1TACAR	cont	Alpha carotene (mc)
DR1TBCAR	cont	Beta carotene (mcg)
DR1TCRYP	cot	Beta crytoxanthin (mcg)
DR1TVB12	cont	Vitamin B12 (mcg)
DR1TVC	cont	Vitamin C (mcg)
DR1TVK	cont	Vitamin K (mcg)
DR1TVD	cont	Vitamin D (D2 + D3) (mcg)
DR1TICAL	cont	Calcium (mg)
DR1TPHOS	cont	Phosphorus (mg)
DR1TMANGN	cont	Magnesium (mg)
DR1TIRON	cont	Iron (mg)
DR1TZINC	cont	Zinc (mg)
DR1TCOPP	cont	Copper (mg)

DR1TSODI	cont	Sodium (mg)
DR1TPOTA	cont	Potassium (mg)
DR1TSELE	cont	Selenium (mg)
DR1TCAFF	cont	Caffeine (mg)
DR1TTHEO	cont	Theobromine
DR1TALCO	cont	Alcohol (gm)
DR1TMOIS	cont	Moisture (gm)
DR1T_32OZ	cont	Total plain water drank yesterday (gm)
Total Dietary Supplements		
DSQTKCAL	cont	Energy (kcal)
DSQTPROT	cont	Protein (gm)
DSQTCARB	cont	Carbohydrate (gm)
DSQTSUGR	cont	Total sugar (gm)
DSQTSFAT	cont	Total saturated fatty acid (gm)
DSQTMFAT	cont	Total monounsaturated fatty acid (gm)
DSQTPFAT	cont	Total polyunsaturated fatty acid (gm)
DSQTFIBE	cont	Dietary fiber (gm)
DSQTTFAT	cont	Total fat (gm)
DSQTCHOL	cont	Cholesterol (mg)
DSQTYCO	cont	Lycopene (mcg)
DSQTILZ	cont	Lutein + Zeaxanthin (mcg)
DSQTVB1	cont	Thiamin (Vitamin B1) (mg)
DSQTVB2	cont	Riboflavin (Vitamin B2) (mg)
DSQTNIAC	cont	Niacin (mg)
DSQTVB6	cont	Vitamin B6 (mg)

DSQTIFA	cont	Folic acid
DSQTFDFE	cont	Folate, DFE (mcg)
DSQTCHL	cont	Total choline (mg)
DSQTVB12	cont	Vitamin B12 (mcg)
DSQTVC	cont	Vitamin C (mcg)
DSQTVK	cont	Vitamin K (mcg)
DSQTVD	cont	Vitamin D (D2 + D3) (mcg)
DSQTCALC	cont	Calcium (mg)
DSQTPHOS	cont	Phosphorus (mg)
DSQTMAGN	cont	Magnesium (mg)
DSQTIRON	cont	Iron (mg)
DSQTZINC	cont	Zinc (mg)
DSQTCOPP	cont	Copper (mg)
DSQTSODI	cont	Sodium (mg)
DSQTPOTA	cont	Potassium (mg)
DSQTSELE	cont	Selenium (mg)
DSQTCAFF	cont	Caffeine (mg)
Laboratory		
URXPREG		Urine Pregnancy Result
LBXGLU	cont	Fasting Glucose (mg/dl)
LBXIN	cont	Insulin (uU/mL)
LBXGH	cont	Glycohemoglobin (%)
Questionnaire		
HIQ011	factor	Covered by health insurance
MCG160M	factor	Ever told you have a thyroid problem

MCQ550	factor	Has DR ever said you have gallstones
MCQ160L	factor	Ever told you had any liver condition
MCQ053	factor	Taking treatment for anemia/past 3 months
BPQ101D	factor	Taking meds to lower blood cholesterol
BPQ020	facor	Ever told you had high blood pressure
BPQ030	factor	Told had high blood pressure 2+ times
SLD012	cont	Sleep hours weekdays or workdays
SLD013	cont	Sleep hours weekends
PAD800	int	Minutes moderate LTPA
PAD820	int	Minutes of vigorous LTPA
PAD680	int	Minutes of sedentary activity
WGQ070	factor	Tried to lose weight in the past year

PCA Test

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25
Standard deviation	3.0215	1.80420	1.56810	1.4722	1.45827	1.3726	1.34021	1.28994	1.24251	1.2193	1.19746	1.16623	1.14944	1.13848	1.09900	1.08179	1.06860	1.06117	1.04048	1.02491	1.02134	1.0103	0.98617	0.97379	0.95958
Proportion of Variance	0.1449	0.05167	0.03903	0.0344	0.03375	0.0299	0.02851	0.02641	0.02451	0.0236	0.02276	0.02159	0.02097	0.02057	0.01917	0.01858	0.01813	0.01787	0.01718	0.01667	0.01656	0.0162	0.01544	0.01505	0.01462
Cumulative Proportion	0.1449	0.19658	0.23562	0.2700	0.30377	0.3337	0.36219	0.38860	0.41311	0.4367	0.45946	0.48105	0.50203	0.52260	0.54177	0.56035	0.57847	0.59635	0.61353	0.63020	0.64676	0.6630	0.67840	0.69345	0.70807

	PC26	PC27	PC28	PC29	PC30	PC31	PC32	PC33	PC34	PC35	PC36	PC37	PC38	PC39	PC40	PC41	PC42	PC43	PC44	PC45	PC46	PC47	PC48	PC49	PC50
Standard deviation	0.94711	0.94425	0.93660	0.92469	0.91019	0.90699	0.89970	0.88188	0.8586	0.85244	0.84127	0.83186	0.82251	0.80862	0.79289	0.78951	0.76596	0.72356	0.7142	0.70391	0.67137	0.65764	0.65180	0.61885	0.60057
Proportion of Variance	0.01424	0.01415	0.01392	0.01357	0.01315	0.01306	0.01285	0.01234	0.0117	0.01153	0.01123	0.01098	0.01074	0.01038	0.00998	0.00989	0.00931	0.00831	0.0081	0.00786	0.00715	0.00687	0.00674	0.00608	0.00573
Cumulative Proportion	0.72230	0.73646	0.75038	0.76395	0.77710	0.79016	0.80301	0.81535	0.8270	0.83859	0.84982	0.86081	0.87155	0.88192	0.89190	0.90180	0.91111	0.91942	0.9275	0.93538	0.94254	0.94940	0.95615	0.96222	0.96795

	PC51	PC52	PC53	PC54	PC55	PC56	PC57	PC58	PC59	PC60	PC61	PC62	PC63
Standard deviation	0.58165	0.51304	0.50861	0.47084	0.4348	0.40544	0.37280	0.35909	0.32427	0.28559	0.26128	0.24662	0.009018
Proportion of Variance	0.00537	0.00418	0.00411	0.00352	0.0030	0.00261	0.00221	0.00205	0.00167	0.00129	0.00108	0.00097	0.000000
Cumulative Proportion	0.97332	0.97750	0.98160	0.98512	0.9881	0.99073	0.99294	0.99499	0.99666	0.99795	0.99903	1.00000	1.000000

Figure 15: PCA test

Correlation Matrix for Feature Removal

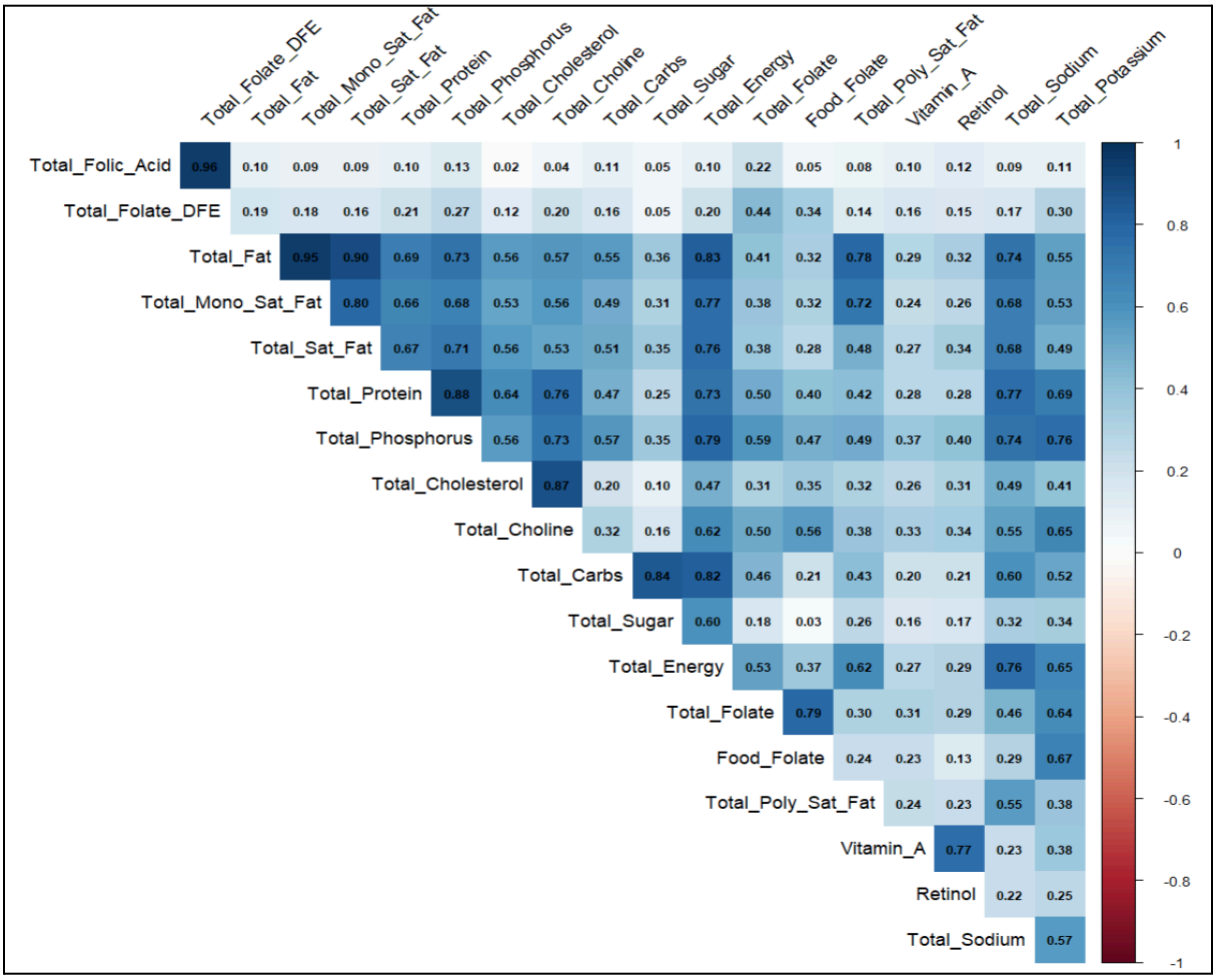

Figure 16: Correlation matrix showing highly correlated variables. Used for feature removal inthe preprocessing section

Feature Importance Initial Models

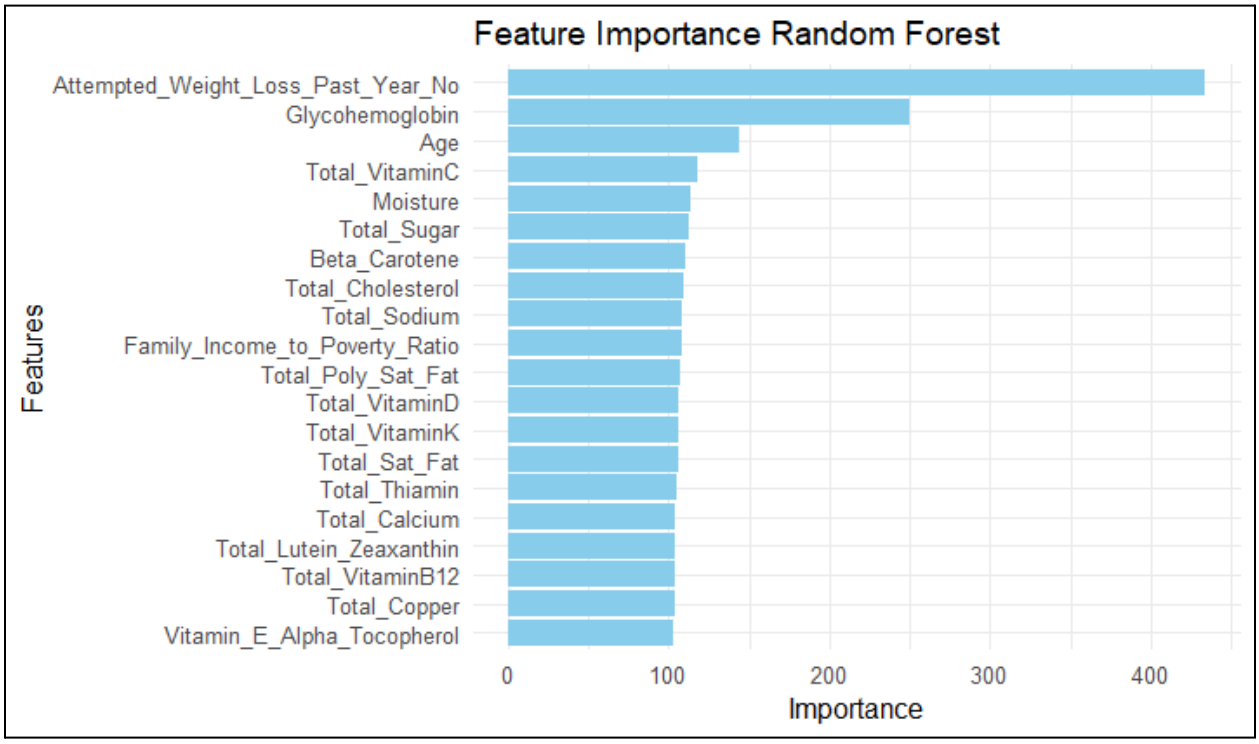

Figure 17: Initial models feature importance

This was the original feature importance for the initial model selection. After review, it was determined that Attempted_Weight_Loss_Past_Year_No was to be removed. This is because it has so much importance and could negatively affect the model. The model could be relying too much on this feature. It may be too correlated to the target variable, and this could cause data leakage.

Random Forest Baseline Confusion Matrix

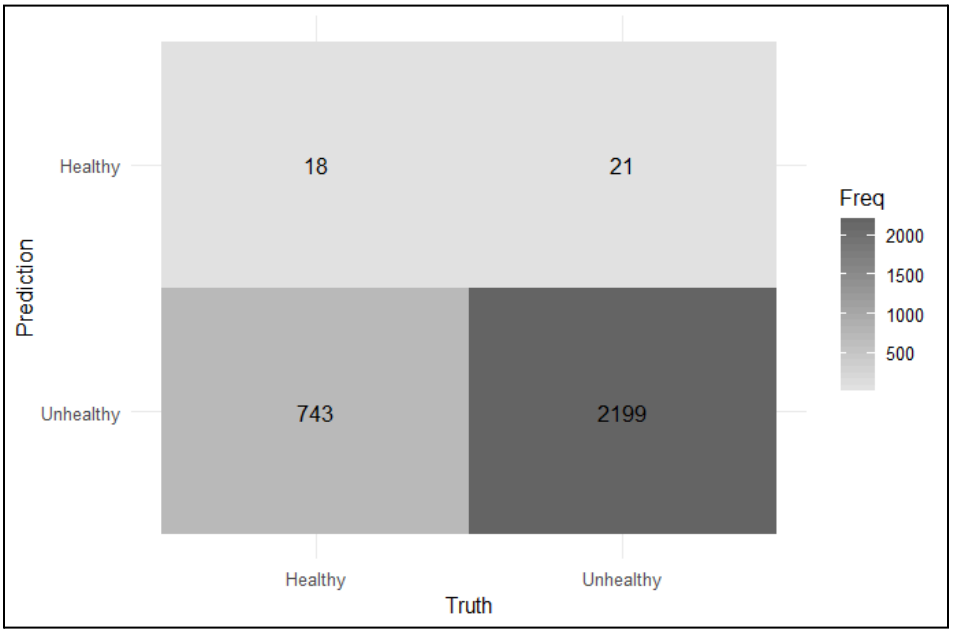

Figure 18: Confusion Matrix Random Forest

Random Forest Baseline ROC Curve

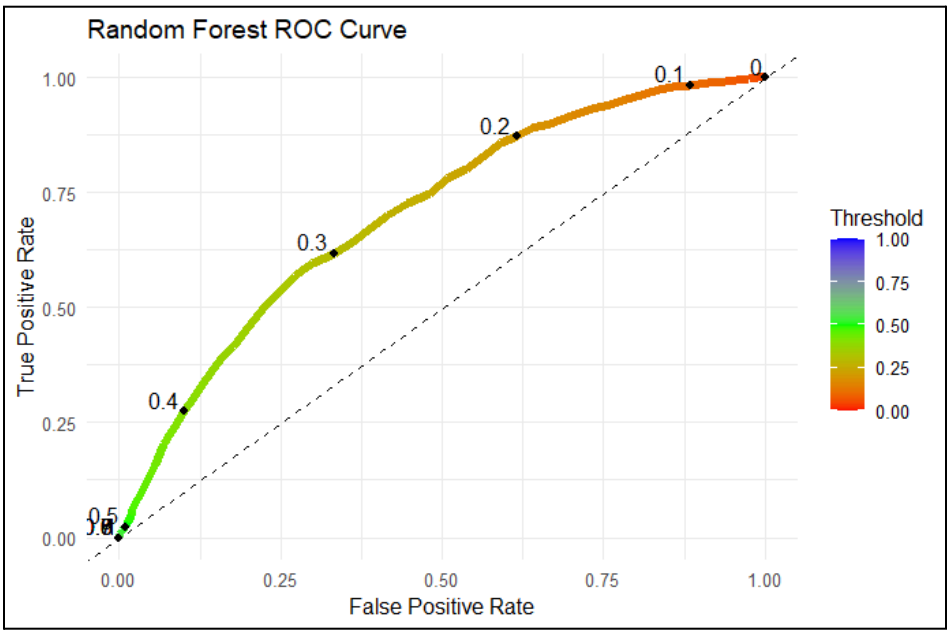

Figure 19: Random Forest baseline ROC curve

Random Forest Baseline Feature Importance

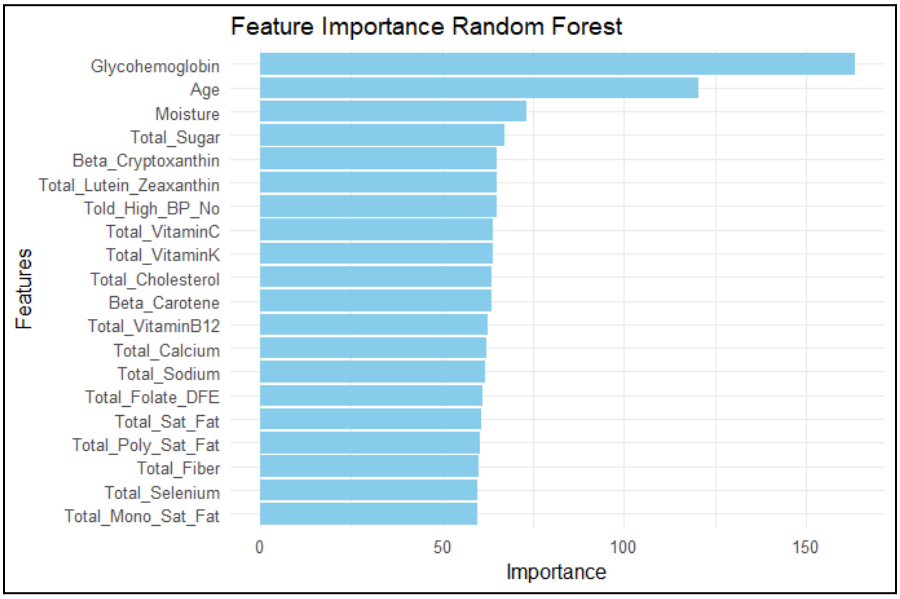

Figure 20: Random ForestBaseline feature importance

Random Forest ADASYN Confusion Matrix

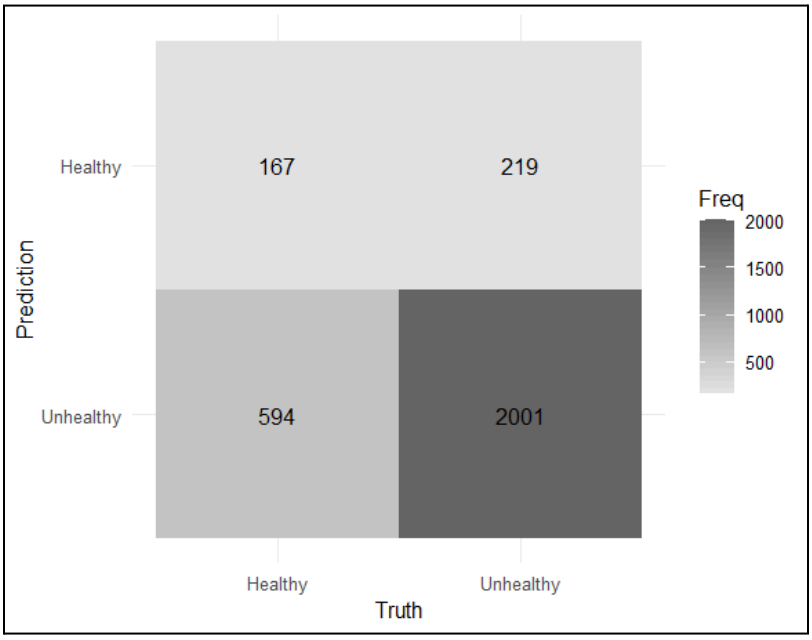

Figure 21: Random Forest ADASYN confusion matrix

Random Forest ADASYN ROC Curve

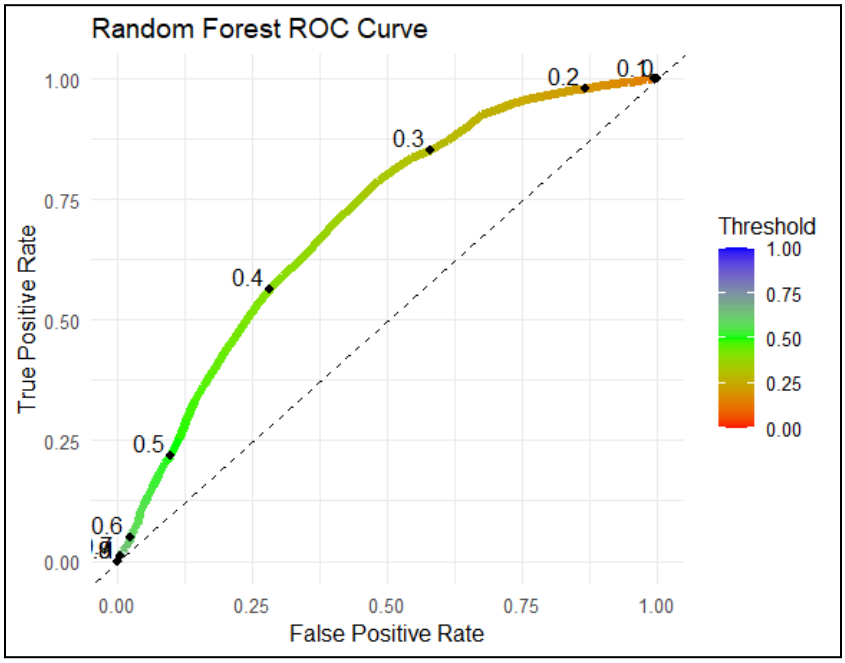

Figure 22: Random Forest ADASYN ROC curve

Random Forest ADASYN Feature Importance

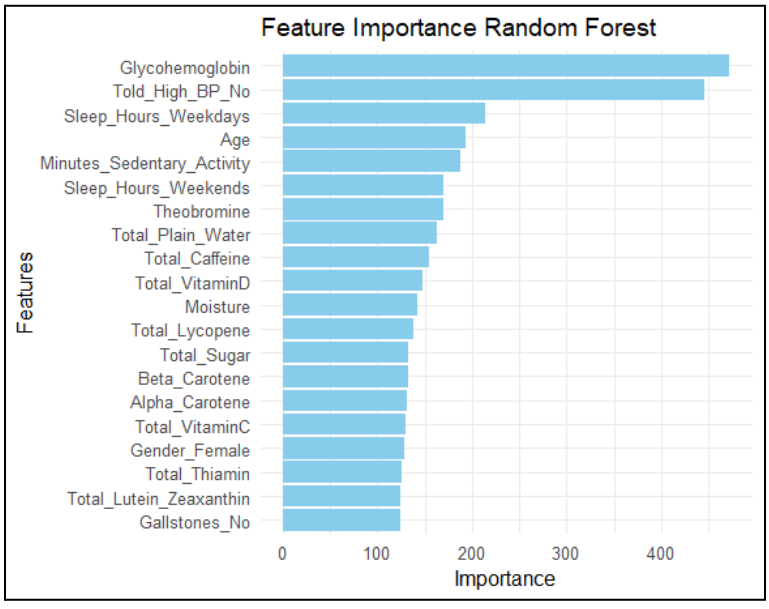

Figure 23: Random Forest ADASYN feature importance

XGBoost Baseline Confusion Matrix

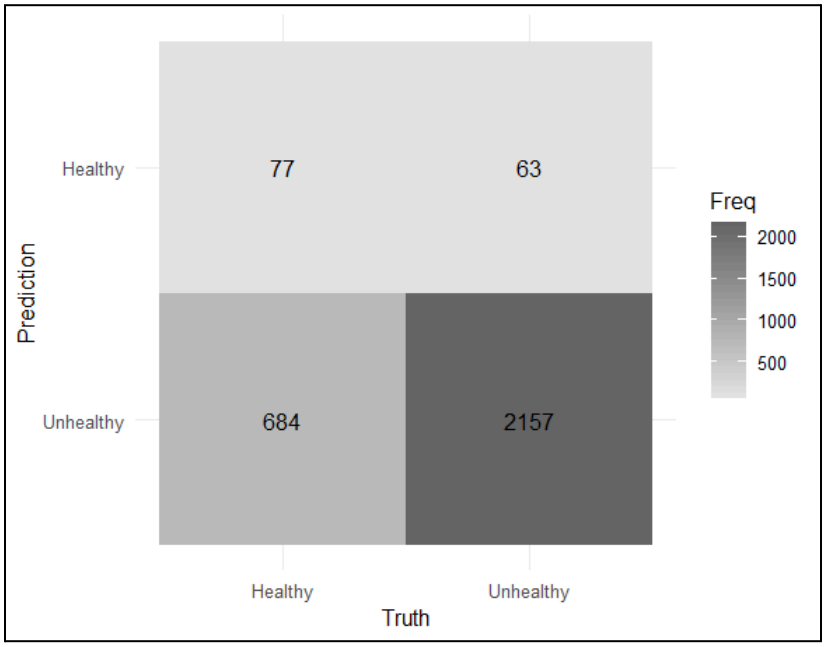

Figure 24: XGBoost baseline confusion matrix

XGBoost Baseline ROC Curve

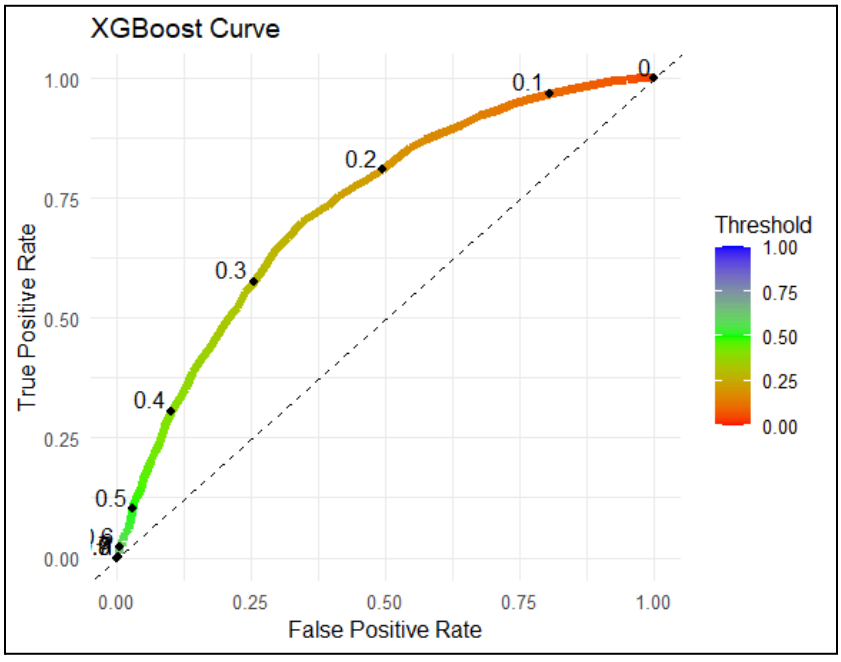

Figure 25: XGBoost baseline ROC curve

XGBoost Baseline Feature Importance

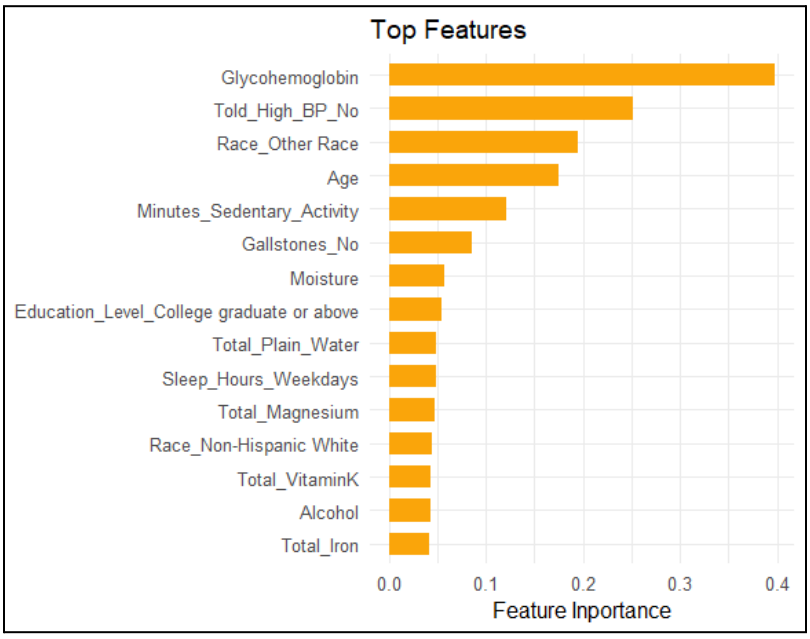

Figure 26: XGBoost baseline feature importance

XGBoost ADASYN Confusion Matrix

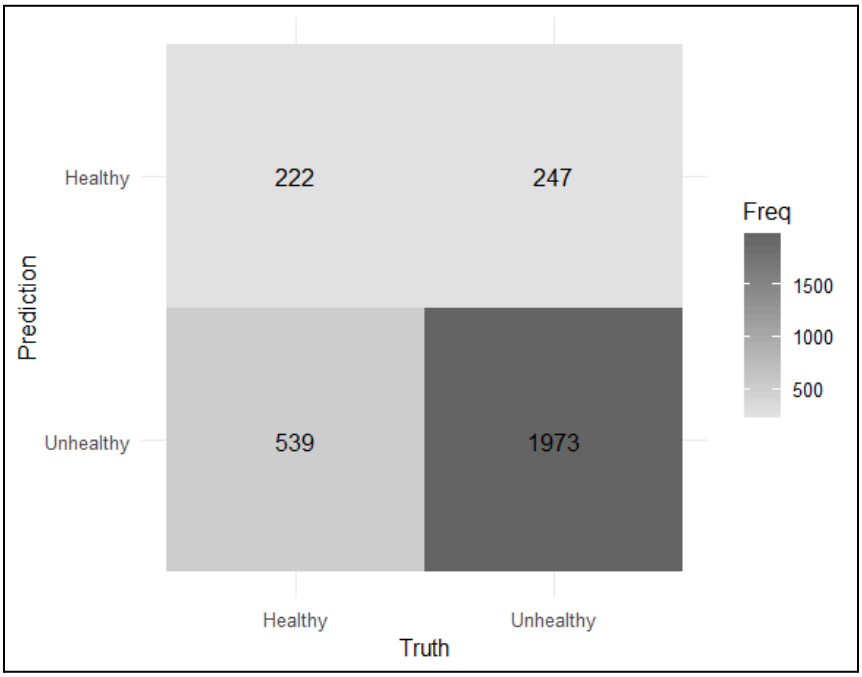

Figure 27: XGBoost ADASYN confusion matrix

XGBoost ADASYN ROC Curve

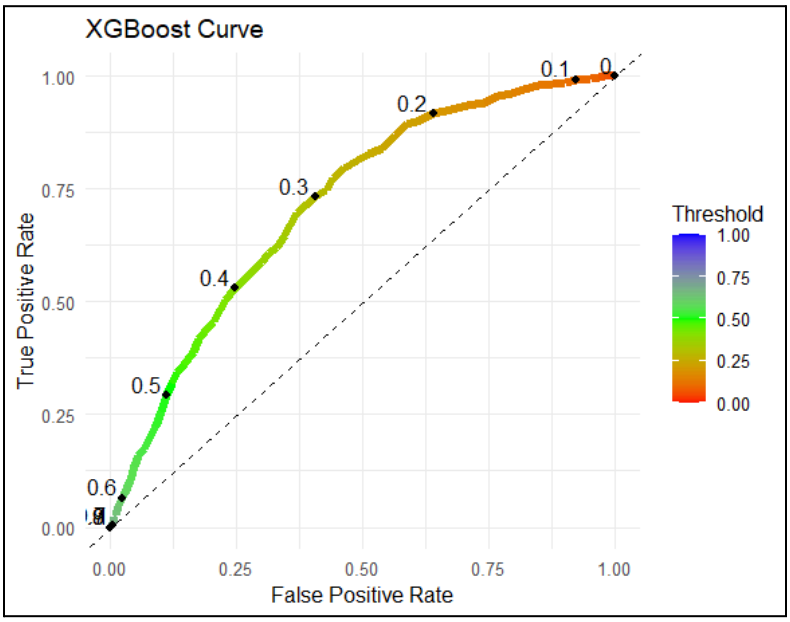

Figure 28: XGBoost ADASYN ROC curve

XGBoost ADASYN Feature Importance

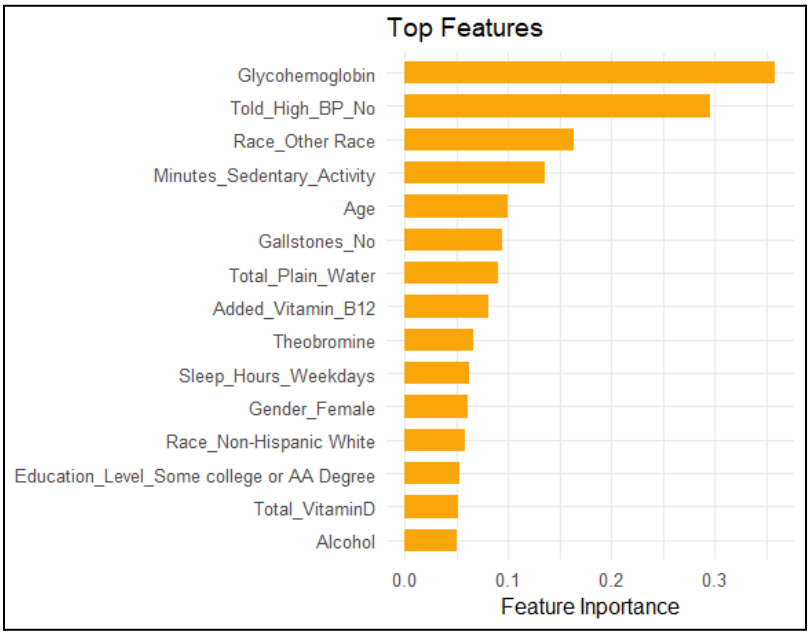

Figure 29: XGBoost ADASYN feature importance