Predicting BMI Category from NHANES 2021–2023 Using Multinomial Logit and Gradient Boosting Trees

Group members: Brinda Mehra, Vivian O'Brien, Stella Guan
Github Repo: https://github.com/StellaTGuan/nhanes-obesity-classification

1. Introduction

Obesity is a major public health concern in the United States and is strongly associated with chronic conditions such as cardiovascular disease and type 2 diabetes. Beyond individual behaviors, body weight is also shaped by socioeconomic constraints that influence daily habits. This motivates a practical question: how well can a small set of demographic, socioeconomic, and lifestyle variables predict whether an adult falls into a higher-risk BMI category?

Using NHANES 2021–2023, we predict adult BMI class using a three-category outcome—UnderNormal (BMI < 25), Overweight (25 ≤ BMI < 30), and Obese (BMI ≥ 30)—constructed from measured BMI. Predictors include age, sex, education, marital status, poverty–income ratio, MVPA-equivalent minutes per week, average sleep hours, alcohol intensity (drinks/day), and Day 1 dietary totals (calories, fiber, saturated fat) when available.

We fit an interpretable multinomial logistic regression baseline and a gradient boosting tree model to capture potential non-linearities and interactions. Both models use the same stratified train/test split and are tuned via cross-validation for comparability.

2. Background & Data Description

2.1 Data Source

NHANES (National Health and Nutrition Examination Survey) is a widely used public health dataset designed to be nationally representative of the U.S. population. It contains standardized measurements, survey-based demographics, and health behavior questionnaires. We used NHANES 2021–2023 because it provides a recent snapshot of adult health behaviors and measured BMI, and it includes both lifestyle and dietary intake modules that plausibly relate to BMI status. Our dataset merges multiple NHANES components using the respondent identifier (SEQN), including demographics (DEMO), body measures (BMX), physical activity (PAQ), alcohol use (ALQ), sleep (SLQ), and day-1 dietary totals (DR1TOT).

2.2 Preprocessing & Feature Engineering

We restrict the analysis to non-pregnant adults because BMI has a different interpretation during pregnancy, and pregnancy status could introduce systematic differences that are not the focus of this classification task. Accordingly, we include participants aged 20 years and older and exclude individuals confirmed pregnant at the examination. We then define the three-class BMI outcome using standard adult cutoffs:

- UnderNormal: BMI < 25
- Overweight: 25 ≤ BMI < 30
- Obese: BMI ≥ 30

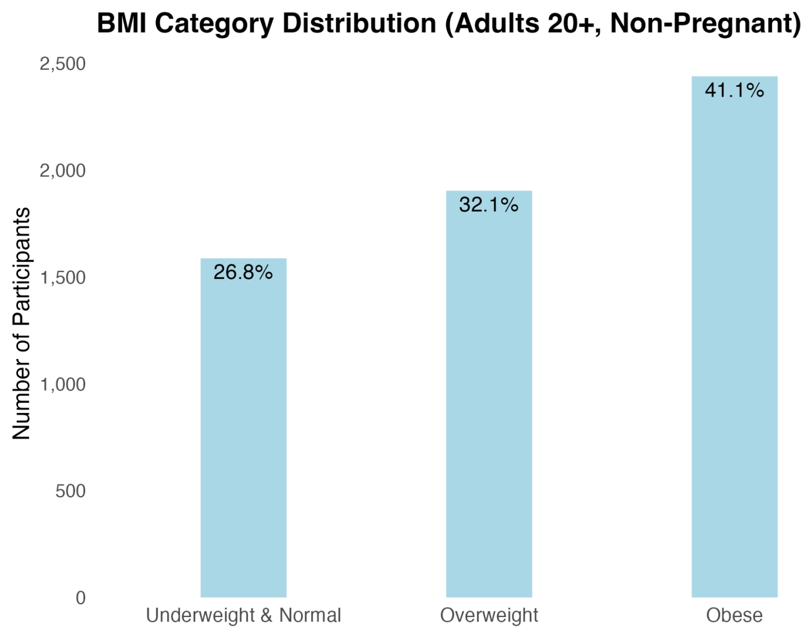

Figure 1: BMI Category Distribution

Preprocessing now follows a single workflow: dataset construction, feature engineering, and complete-case selection. We retain only the relevant variables from each NHANES component file, merge sources using the participant identifier, and convert survey responses into analysis-ready numeric and categorical formats. We apply broad sanity checks by recoding clearly impossible values to missing rather than dropping entire observations. For example, reported sleep duration is constrained to 0–24 hours, physical activity session duration is constrained to 0–720 minutes, average alcoholic drinks per day is constrained to 0–50, and total daily energy intake (Day 1) is constrained to 200–8,000 kcal/day. We also convert NHANES special missing codes (e.g., 777 and 999 in alcohol measures) into standard missing values so they are handled consistently. After these checks and recodes, we drop all remaining missing values.

Specifically, our predictor set combines demographics, socioeconomic status, and behavioral/lifestyle measures. Demographic predictors include sex (sex; coded as male vs female), age in years (age), education level (educ; collapsed into three levels: HS or less, Some college/AA, College+), and marital status (marital). Socioeconomic status is captured by the poverty–income ratio (pir), a standardized measure of household income relative to the federal poverty threshold. Lifestyle and health-behavior predictors include average sleep duration (hours) (sleep_hours_avg), computed as a weighted weekly average of weekday and weekend sleep hours; alcohol consumption (drinks/day) (alcohol_drinks_per_day), defined as 0 for never drinkers or those reporting no drinking in the past year, and otherwise set to reported average drinks per day; and physical activity volume (MVPA-equivalent minutes/week) (mvpa_eq_min_wk), constructed by converting reported moderate and vigorous activity frequencies into sessions per week, multiplying by minutes per session to get minutes per week, and then combining them to reflect higher intensity for vigorous activity. Dietary predictors come from the Day 1 total nutrient intake file and include total energy intake (kcal) (kcal_day1), dietary fiber intake (g) (fiber_day1), and saturated fat intake (g) (satfat_day1).

Because the analysis dataset is restricted to complete cases, we do not perform modeling-time missingness handling (e.g., “Unknown” categories, median imputation, or missingness indicators). Instead, the processed dataset written to disk contains no missing values, and both the train–test split and all downstream modeling steps operate on the same NA-free feature set.

$$Avg.\ sleep\ hrs\ =\ \frac{5 \cdot weekdays + 2 \cdot weekends}{7}$$

Equation 1: Average Sleep Hours

$$Physical\ Activity\ Volume\ =\ ModerateMinPerWeek\ +\ 2 * VigorousMinPerWeek$$

Equation 2: Physical Activity Volume (MVPA-equivalent minutes/week)

3. Models

3.1 Multinomial Logistic Regression: statistical specification, training, and tuning

Our first model is multinomial logistic regression, which is appropriate when the outcome is categorical with more than two unordered classes. We let $Y \in \{1, 2, 3\}$ denote BMI class for individuals, corresponding to groups UnderNormal, Overweight, and Obese; we let x be the vector of predictors (demographics, lifestyle variables, and dietary totals). The multinomial logit model specifies class probabilities using a softmax form. Taking **UnderNormal** as the baseline class, for each non-baseline class $k \in \{Overweight, Obese\}$, the model assumes:

$$log(\frac{Pr(Y_i = k \mid x_i)}{Pr(Y_i = UnderNormal \mid x_i)}) = \beta_{k0} + x_i^T \beta_k$$

Equation 3: Multinomial Logit

Equivalently, the class probabilities are:

$$Pr(Y_i = k \mid x_i) = \frac{exp(\beta_{k0} + x_i^T \beta_k)}{1 + \sum_{j \in \{Overweight, Obese\}} exp(\beta_{j0} + x_i^T \beta_j)}, \ Pr(Y_i = UnderNormal \mid x_i) = \frac{1}{1 + \sum_j exp(\beta_{j0} + x_i^T \beta_j)}$$

Equation 4: Multinomial Logit Class Probabilities

We estimate the multinomial logistic regression parameters by choosing coefficients that minimize the negative log-likelihood on the training data. Since multinomial logistic regression is interpretable, we also examine model coefficients. For each class **k** relative to the reference category, a coefficient for predictor **j** represents the change in the log-odds of being in class **k** associated with a one-unit increase in that predictor, holding other predictors fixed. Exponentiating the coefficient yields the corresponding odds ratio, ($exp(\beta_{kj})$). We provide the full set of estimated multinomial logistic regression coefficients and corresponding odds ratios in the Appendix (**Table 4**). This table supports detailed, class-specific interpretation and allows readers to identify the strongest predictors for each BMI category.

Finally, we run two simple stability diagnostics to determine whether more aggressive penalization is necessary. We compute the difference between best cross-validated accuracy and test accuracy **(Table 5)** as a rough overfitting signal, and we compute the maximum absolute coefficient magnitude as a rough indicator of potential separation or unstable estimates. In our fitted model, the maximum absolute coefficient is 0.507, which is not suggestive of extreme separation. Together with the small CV–test difference of -0.0161, this supports treating the baseline multinomial model as adequate without moving to a more complex penalized multinomial framework for this project's scope.

3.2 Gradient boosting trees (GBM): statistical specification, training, and tuning

Our second model is gradient boosting trees, a powerful ensemble method that builds an additive model of many shallow decision trees. In boosting, the prediction function is constructed as a sum of "weak learners." For multi-class classification, the model can be understood as learning class score functions ($F_k(x)$) and mapping them to probabilities through a softmax:

$$Pr(Y_i = k \mid x_i) = \frac{exp(F_k(x_i))}{\sum_{j=1}^{3} exp(F_j(x_i))}$$

Equation 5: Boosting Multiclass Probability Mapping

The functions are built iteratively, starting from an initial set of scores, boosting adds trees to reduce the loss function, typically the multinomial deviance (cross-entropy). At iteration (m), the algorithm fits a decision tree to approximate the negative gradient of the loss with respect to the current model predictions (hence "gradient" boosting). The update takes the form:

$$F_k^{(m)}(x) = F_k^{(m-1)}(x) + v \cdot h_{km}(x)$$

Equation 6: Gradient Boosting Additive Update

where ($h_{km}(x)$) is the fitted tree at iteration (m) for class (k), and (\nu) is the learning rate (shrinkage) controlling step size. Key hyperparameters govern model complexity and generalization: the number of trees (B) (n.trees), the depth of each tree (interaction.depth), the learning rate (shrinkage), and a minimum node size (n.minobsinnode) that controls how small terminal nodes may become. Larger depth and more trees increase flexibility, while smaller learning rates often require more trees but can generalize better.

We train GBM using the same stratified 75/25 split approach as in the logistic model, ensuring a fair comparison. Hyperparameter tuning is performed via 5-fold cross-validation on the training set, selecting the configuration that maximizes cross-validated accuracy. The full tuning grid and results are saved in **Table 6**, and the selected best tuning parameters are shown in **Table 1**, indicating that the best configuration is n.trees = 100, interaction.depth = 1, shrinkage = 0.05, and n.minobsinnode = 20. We visualize how cross-validated error changes with number of trees under the selected hyperparameter combination using Figure 2, which helps confirm that performance is not being driven by an obviously under- or over-fit number of boosting iterations.

n.trees	interaction.depth	shrinkage	n.minobsinnode
100	1	0.05	20

Table 1: GBM Best Tune Result

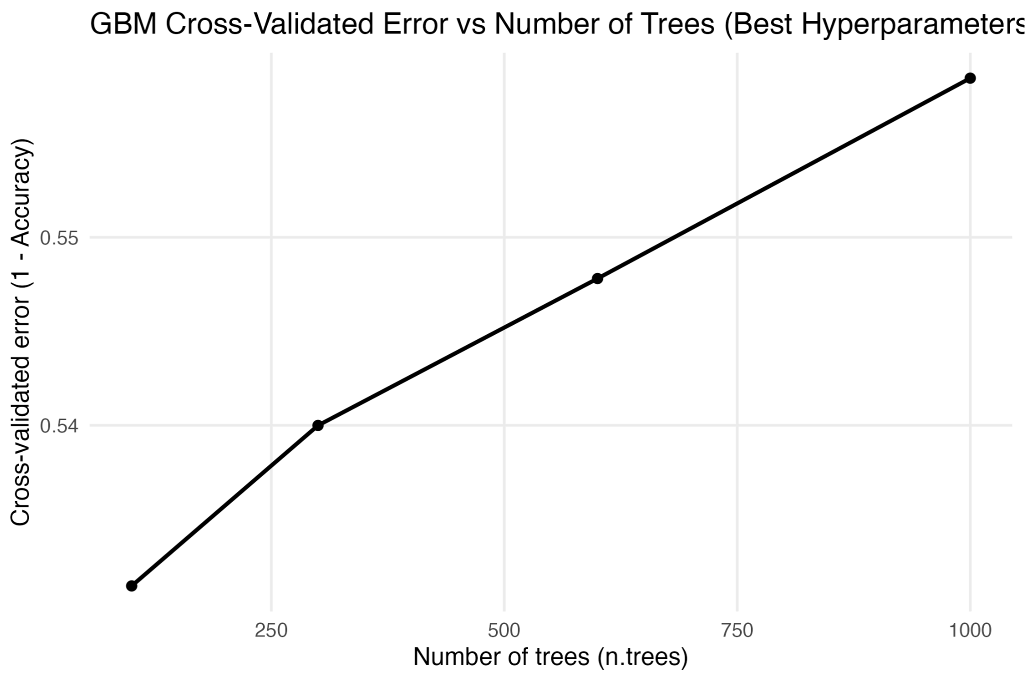

Figure 2: GBM CV curve

Since boosting is less directly interpretable than logistic regression, we summarize variable influence using variable importance. These importance values reflect how often and how effectively each feature contributes to splitting decisions across the ensemble, which is useful for understanding which predictors the model relies on most, even though it does not provide causality.

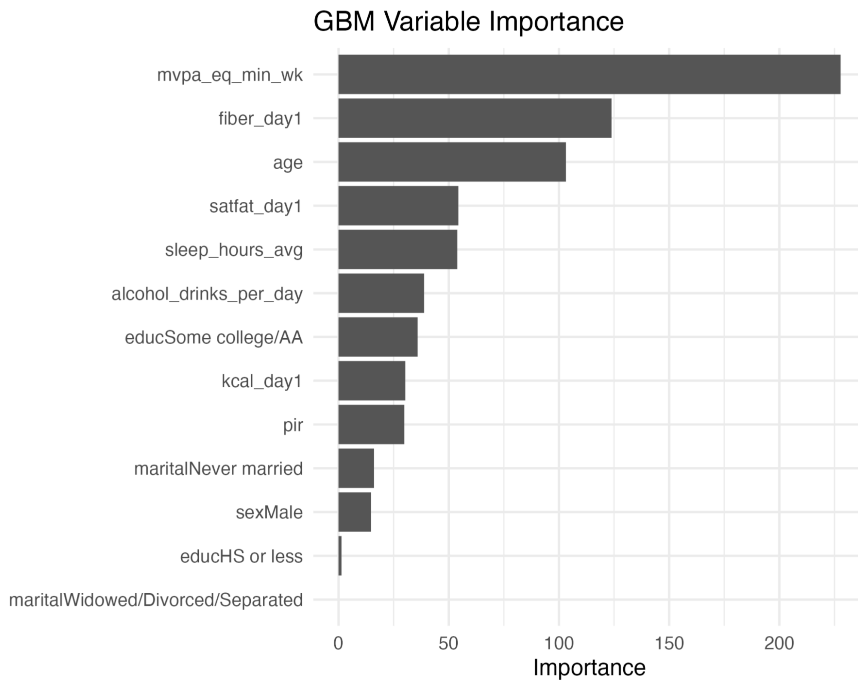

Figure 3: GBM Variable Importance

4. Results

We evaluate both models on the held-out test set using accuracy and macro-F1, and we present confusion matrices to understand class-specific performance. The multinomial logistic regression achieves test accuracy 0.465 and macro-F1 0.401, while GBM achieves test accuracy 0.460 and macro-F1 0.407. Overall, predictive performance is very similar across both models. Given the multinomial model's slightly higher test accuracy and substantially greater interpretability, we chose logistic regression as the preferred approach for this feature set.

accuracy	macro_f1
0.4647435897	0.4010637179

Table 2: Multinom Test Metrics

accuracy	macro_f1
0.4604700855	0.4068444967

Table 3: GBM Test Metrics

The confusion matrices provide a clearer picture of model behavior than accuracy alone. The multinomial logistic regression has relatively high sensitivity for the Obese class (0.7846) but very low sensitivity for UnderNormal (0.2190) and Overweight (0.25). This indicates that the model tends to concentrate predictions in the majority class, capturing many obese individuals while failing to correctly identify a large portion of non-obese individuals.

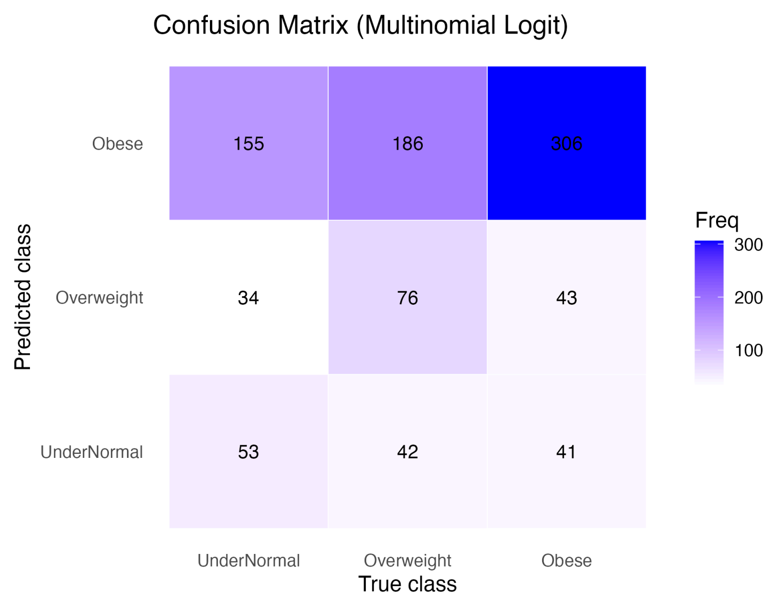

Figure 4: Multinomial Logit Confusion Matrix

GBM shows a slightly different tradeoff across classes. Relative to multinomial logistic regression, GBM modestly increases recall for UnderNormal (0.252 vs 0.219) and Overweight (0.257 vs 0.250), but this comes at the cost of lower recall for Obese (0.749 vs 0.785). In other words, boosting shifts away from concentrating predictions in Obese and recovers some additional non-obese cases, yet overall performance remains very similar—so it does not deliver a clear headline gain over the multinomial logit model for this feature set.

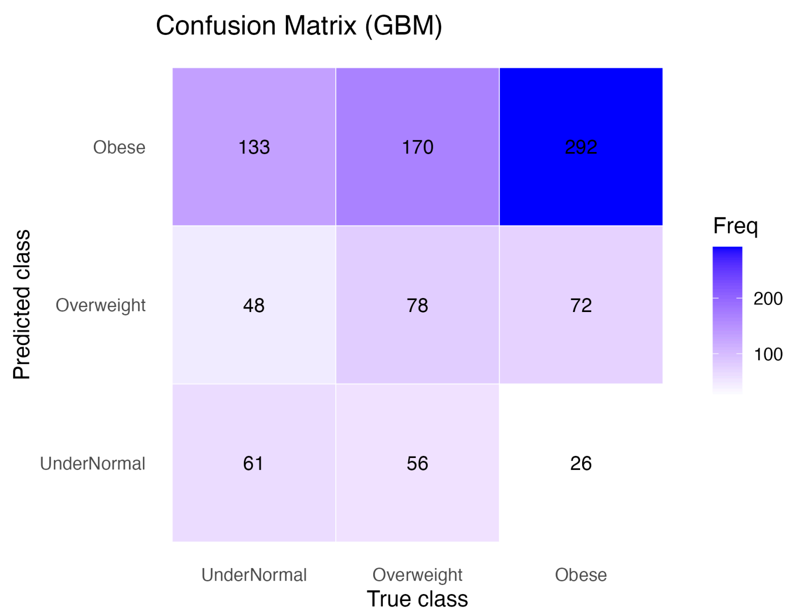

Figure 5: GBM confusion matrix

Interpretability differs sharply between the two methods. Multinomial logistic regression yields coefficients that directly represent log-odds differences between BMI categories. The complete set of estimated coefficients and corresponding odds ratios are in **Table 4**. For any predictor, the sign and magnitude of its coefficient indicate how that predictor shifts the odds of being Overweight or Obese relative to UnderNormal, holding all other covariates fixed. For instance, age has a positive coefficient for both Overweight and Obese, consistent with higher age being associated with higher odds of heavier BMI categories in this specification. In contrast, MVPA-equivalent minutes per week has a negative coefficient for Obese, consistent with greater physical activity being associated with lower odds of obesity, conditional on the other predictors. These coefficient interpretations reflect conditional associations in the observed data and should not be interpreted as causal effects.

GBM does not yield coefficients, but its variable importance results still highlight which predictors are most useful for classification in the complete-case sample (i.e., after dropping all observations with missing values). Physical activity (mvpa_eq_min_wk), fiber intake (fiber_day1), saturated fat intake (satfat_day1), age, and total energy intake (kcal_day1) appear among the most influential predictors, followed by income-to-poverty ratio (pir) and sleep (sleep_hours_avg).

Overall, the results suggest that our predictors contain some signal about BMI category but not enough to achieve strong discrimination across all three classes. The confusion matrices show that the primary weakness is not simply overall accuracy, but persistent overlap between UnderNormal and Overweight and a tendency to over-predict Obese: multinomial logit correctly identifies Obese relatively well but misses many non-obese cases, while GBM slightly improves sensitivity for UnderNormal and Overweight (at the cost of somewhat lower sensitivity for Obese), leaving the same basic separation challenge unresolved.

5. Discussion

The primary goal of this project was to assess whether demographic, lifestyle, and simple dietary intake measures can meaningfully predict adult BMI category in NHANES 2021–2023, and whether a flexible machine learning model improves substantially over an interpretable baseline. Using the NA-free (complete-case) dataset, both multinomial logistic regression and gradient boosting achieve only modest predictive performance on a held-out test set. Test accuracy is very similar across models (0.465 for multinomial logit vs 0.460 for GBM), with multinomial logit performing slightly better on accuracy, while GBM achieves a small advantage in macro-F1 (0.401 vs 0.407). Given the near-tie in overall performance and the multinomial model's substantially greater interpretability, logistic regression remains a reasonable preferred approach for this feature set.

The confusion matrices help explain what these headline metrics mean in practice. Both models are substantially better at identifying the Obese class than the UnderNormal or Overweight classes, indicating that class separation is limited with the current predictors. In the multinomial logit model, sensitivity is relatively high for Obese ($306/390 \approx 0.785$) but much lower for UnderNormal ($53/242 \approx 0.219$) and Overweight ($76/304 = 0.250$), reflecting a tendency to concentrate predictions in the majority class. GBM shifts this tradeoff slightly: it improves sensitivity for UnderNormal and Overweight ($61/242 \approx 0.252$ and $78/304 \approx 0.257$) but reduces sensitivity for Obese ($292/390 \approx 0.749$). This pattern is consistent with GBM's slightly higher macro-F1: it recovers somewhat more balanced performance across classes, but not enough to change the overall conclusion that discrimination remains limited.

These results also clarify why boosting does not "drastically" outperform logistic regression here. While GBM can capture non-linearities and interactions, a flexible model cannot create signal that is not present in the features. BMI is influenced by long-run behavioral patterns, genetics, and complex environmental factors, whereas our dietary measures are day-1 totals and our lifestyle measures are coarse composites. In addition, the classification boundary between UnderNormal and Overweight may be particularly subtle in these predictors, making large performance gains difficult even with non-linear learners. The fact that both models only modestly outperform a naive majority-class baseline (390/936 ≈ 0.417) reinforces that the available predictors contain some information, but not enough for strong multiclass separation.

This analysis has several limitations. First, because we restrict the analysis to complete cases, the final modeling sample may differ systematically from the full NHANES population, which can affect both generalizability and the observed class balance. Second, day-1 dietary totals may be noisy and may not represent typical intake, weakening predictive relationships. Third, the outcome is ordinal in nature (UnderNormal < Overweight < Obese), yet both methods treat it as nominal multiclass; ordinal models may better exploit the ordered structure. Finally, accuracy alone can mask poor minority-class recall, so if balanced classification performance is the goal, future work should consider alternative objectives (e.g., class-weighted loss or directly optimizing macro-F1).

Future research could extend this project by adding richer nutritional variables (e.g., macronutrient shares, added sugars, or multi-day averages), incorporating additional health biomarkers and comorbidity indicators, or using ordinal classification methods. It may also be valuable to evaluate calibration (how well predicted probabilities match observed frequencies) and to perform subgroup analyses by age or sex to see whether predictability differs across populations.

6. References

Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey (NHANES) Data, 2021–2023.* Hyattsville, MD: U.S. Department of Health and Human Services, CDC. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R* (Vol. 103). New York, NY: Springer.

7. Appendix

class	term	beta	odds_ratio
Obese	**(Intercept)**	**0.5069853493**	**1.660278483**
Obese	`educSome college/AA`	0.3442794145	1.410972827
Obese	fiber_day1	-0.2953224756	0.7442915329
Obese	mvpa_eq_min_wk	-0.2941208166	0.7451864551
Obese	`educHS or less`	0.2427286076	1.274722627
Obese	alcohol_drinks_per_day	0.2065885875	1.229476646
Obese	satfat_day1	0.1667249271	1.181429241
Obese	`maritalNever married`	-0.1458061967	0.8643251961
Obese	sleep_hours_avg	-0.1068303811	0.8986780928
Obese	age	0.1026799302	1.108136671
Obese	kcal_day1	-0.0507104996	0.9505538164
Obese	sexMale	0.04507127649	1.04610242
Obese	pir	-0.04209322539	0.9587803937
Obese	`maritalWidowed/Divorced/Separated`	-0.01452665096	0.9855783518
Overweight	(Intercept)	0.2871490826	1.33262287
Overweight	`maritalNever married`	-0.1785839316	0.836453849
Overweight	sexMale	0.1757400491	1.192128124
Overweight	`educSome college/AA`	0.1735443324	1.18951342
Overweight	alcohol_drinks_per_day	0.1518279241	1.16395993
Overweight	fiber_day1	-0.1302787626	0.8778506849
Overweight	mvpa_eq_min_wk	-0.1247939582	0.8826787525
Overweight	`educHS or less`	0.116436529	1.123486199
Overweight	age	0.09200498654	1.096370289
Overweight	kcal_day1	-0.08429164671	0.9191631461
Overweight	satfat_day1	0.07775144907	1.080853978
Overweight	sleep_hours_avg	-0.02344910829	0.9768236856
Overweight	`maritalWidowed/Divorced/Separated`	-0.0108008234	0.9892572961
Overweight	pir	0.009225516598	1.009268203

Table 4: Multinomial Coefficients

best_cv_accuracy	test_accuracy	cv_minus_test
0.4486318211	0.4647435897	-0.0161117686

Table 5: Best CV Accuracy vs Test Accuracy

shrinkage	interaction.depth	n.minobsinnode	n.trees	Accuracy	Kappa	AccuracySD	KappaSD
0.01	1	10	100	0.4443569338	0.0934440347	0.01287535047	0.01908982597
0.01	1	20	100	0.4450699418	0.1001822461	0.01185124731	0.02226066673
0.05	1	10	100	0.4596550002	0.1398945512	0.00914120023	0.01274410916
0.05	1	20	100	0.4685473727	0.1552113671	0.012803426	0.02040763587
0.01	2	10	100	0.4539616821	0.125609191	0.00833262934	0.01632966619
0.01	2	20	100	0.4504067559	0.1184611485	0.00981446414	0.01697934025
0.05	2	10	100	0.4660562695	0.1575682373	0.00969729170	0.01410875221
0.05	2	20	100	0.4664108772	0.1591366894	0.01139196382	0.0159854077
0.01	3	10	100	0.4561019703	0.1314018952	0.01111839914	0.01981648014
0.01	3	20	100	0.4621442071	0.1402705159	0.00922694612	0.01435360556
0.05	3	10	100	0.4589457848	0.1482718455	0.01324206688	0.01941154304
0.05	3	20	100	0.4653489504	0.1597408408	0.01710540342	0.02320134087
0.01	1	10	300	0.4550236089	0.1254026077	0.01476695741	0.02555171433
0.01	1	20	300	0.452177898	0.1216556401	0.01127850345	0.02295110064
0.05	1	10	300	0.4550261373	0.1436926209	0.010278729	0.01881508034
0.05	1	20	300	0.4600083437	0.1506276711	0.00871807761	0.01325713059
0.01	2	10	300	0.4671200925	0.1536148047	0.00463375196	0.00608220633
0.01	2	20	300	0.4642750137	0.1488671029	0.00465658973	0.00514047177
0.05	2	10	300	0.4518359323	0.1418782209	0.02227442446	0.03566924314
0.05	2	20	300	0.4585924414	0.1522697524	0.01439223393	0.02159739071
0.01	3	10	300	0.4624962864	0.1491973498	0.00741223030	0.0116581973
0.01	3	20	300	0.4614324634	0.1487978711	0.01035558794	0.01617979583
0.05	3	10	300	0.4536159238	0.1465689535	0.01789118899	0.02632460623
0.05	3	20	300	0.455030562	0.14914254	0.00805567847	0.01206940827
0.01	1	10	600	0.461073431	0.1464830221	0.00943953871	0.01718757435
0.01	1	20	600	0.4624981827	0.1488390636	0.00933482949	0.01398592841
0.05	1	10	600	0.455387066	0.1469654259	0.00679389776	0.0120465994
0.05	1	20	600	0.4521911721	0.1421933436	0.01013086362	0.01416069933
0.01	2	10	600	0.4657003976	0.1585965864	0.0121106214	0.01863671914
0.01	2	20	600	0.4625007111	0.153528689	0.01244802071	0.01840574069

0.05	2	10	600	0.4525451477	0.1449993933	0.01280893989	0.01884663645
0.05	2	20	600	0.4472165509	0.1382753441	0.01973070379	0.02979362919
0.01	3	10	600	0.4639223024	0.1575232379	0.01220609589	0.01841668309
0.01	3	20	600	0.466409613	0.1617201436	0.01477629504	0.02267007309
0.05	3	10	600	0.4440238175	0.1352403161	0.02062413502	0.02918428396
0.05	3	20	600	0.4521924363	0.1476512048	0.02432343244	0.0365710638
0.01	1	10	1000	0.4589426243	0.1476805446	0.01327705555	0.02229436423
0.01	1	20	1000	0.460366744	0.1493993219	0.01087169814	0.01593690826
0.05	1	10	1000	0.4500483556	0.1407018362	0.01562631509	0.02698707287
0.05	1	20	1000	0.4415289217	0.1275296405	0.0117684894	0.01494189283
0.01	2	10	1000	0.4593041851	0.150635568	0.01384676915	0.01954789913
0.01	2	20	1000	0.4625026074	0.1562999999	0.01367456616	0.01904743443
0.05	2	10	1000	0.4468587827	0.1387444921	0.01954139636	0.02948676955
0.05	2	20	1000	0.4472178151	0.1397652904	0.01541320133	0.02286939547
0.01	3	10	1000	0.4625013432	0.1583270414	0.01626608907	0.025268087
0.01	3	20	1000	0.4589445206	0.1518242587	0.01513966851	0.02329527919
0.05	3	10	1000	0.4369101724	0.125741198	0.01719852951	0.02361655793
0.05	3	20	1000	0.4330012705	0.1209655641	0.01924378879	0.0274590991

Table 6: GBM CV Result

decay	Accuracy	Kappa	AccuracySD	KappaSD
0	0.4486318211	0.1135298059	0.02036182722	0.03401362108
1.00E-04	0.4486318211	0.1135298059	0.02036182722	0.03401362108
0.001	0.4486318211	0.1135298059	0.02036182722	0.03401362108
0.01	0.4486318211	0.1135298059	0.02036182722	0.03401362108
0.1	0.4486318211	0.1135298059	0.02036182722	0.03401362108

Table 7: Multinomial CV Result

feature	Overall
mvpa_eq_min_wk	227.7883936
fiber_day1	123.8623197
age	103.1545435
satfat_day1	54.36258758
sleep_hours_avg	53.88217032

alcohol_drinks_per_day	38.84051625
educSome college/AA	35.88822358
kcal_day1	30.28654074
pir	29.84466729
maritalNever married	16.09206927
sexMale	14.75835097
educHS or less	1.348236115
maritalWidowed/Divorced/Separated	0

Table 8: GBM Variable Importance