Predicting the Risk of Obesity in the United States

Caleb Kim [*1] **Daniel Song** [*1] **Christina Yang** [*1]

Abstract

This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Data Description

The purpose of this project is to predict obesity risk through looking at factors including lifestyle, demographics, and the environment. Obesity, defined as a Body Mass Index (BMI) of 30 or higher, is a major public health challenge in the United States. Its prevalence is due to both individual behaviors and structural conditions in the surrounding environment. Our dataset is constructed from multiple raw data sources that require substantial preparation and integration.

The primary source of individual health information is the Behavioral Risk Factor Surveillance System (BRFSS), which is an annual survey conducted by the Centers for Disease Control and Prevention (CDC). This survey collects nationwide self-reported health behaviors, including diet, physical activity, smoking, alcohol use, and basic demographic characteristics such as age, sex, race or ethnicity, and income. BMI is calculated directly from reported height and weight values, and we also computed an obesity indicator based on BMI ≥ 30. BRFSS provides a large, rich foundation of individual-level records that link personal health behaviors with obesity outcomes.

To complement these individual responses, we incorporate contextual measures from the United States Department of Agriculture's Food Environment Atlas. These datasets include grocery store availability, convenience and specialty food store counts, SNAP-authorized retailers, fast-food and full-service restaurant density, and per-capita food sales. While these variables were originally provided at the county level, they were aggregated to the state level so that they could be merged with the BRFSS individual-level dataset. These environmental indicators capture important structural contributors to obesity, as many health outcomes are shaped not only by personal choices but also by the accessibility and distribution of food outlets in one's community.

We merged U.S. Census Bureau's American Community Survey to include the state names. The main focus remained on merging BRFSS health behaviors with USDA state-level food environment characteristics. This streamlined approach ensured that all datasets aligned cleanly at the state level, while still capturing both individual and environmental determinants of obesity.

The outcome variable of interest is obesity status, defined as a binary classification based on BMI. Individual-level predictors include lifestyle and demographic factors from BRFSS, while environmental predictors are derived from the USDA Food Environment Atlas. Together, the combined dataset provides a comprehensive view that incorporates both personal health behaviors and structural characteristics of the food environment.

Preparing these data for analysis presents several challenges. First, BRFSS survey responses include missing values and nonstandard codes such as "Don't know" or "Refused," as well as implausible height or weight values that distort BMI calculations. These problems were addressed through data cleaning, the removal of biologically unrealistic BMI values, the conversion of encoded dietary and alcohol variables into interpretable continuous measures, and imputation strategies suited to each variable type. Second, merging datasets required geographic alignment across BRFSS and USDA sources. Because BRFSS data are at the individual level and USDA data are at the county level, we aggregated the food environment datasets to the state level to ensure consistent linkage. Third, many variables required recoding and standardization. For instance, income categories were converted to ordinal values, fruit and vegetable consumption metrics were standardized to daily frequencies, and behavioral indicators such as smoking and exercise were transformed into usable binary variables. Continuous predictors were later standardized to support model training.

The volume and complexity of the BRFSS data also in-

[*]Equal contribution [1]Department of Data Science, University of Virginia, Charlottesville Virginia, United States. Correspondence to: Caleb Kim <qch2ev@virginia.edu>, Daniel Song <yen2pn.2@virginia.edu>, Christina Yang <gca9aa@virginia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

creased the need for efficient filtering and preprocessing. With more than 130,000 observations retained after cleaning, careful handling of missingness and feature engineering was necessary to ensure a reliable analytical file. Through constructing BMI values, excluding unrealistic cases, recoding categorical responses, and merging environmental datasets at the state level, we created a structured dataset suitable for predictive modeling. Since data from BRFSS can include extreme or unrealistic responses, we applied clear rules to remove outliers. Missing values were handled through imputation, and continuous features were standardized prior to model fitting.

Through this process, we created a unified dataset that links individual health behaviors with the structural characteristics of the surrounding food environment. This allows us to examine not only how personal factors influence obesity but also how broader community contexts shape health outcomes. Through reading, cleaning, and merging these raw sources, we prepared a reliable foundation for the predictive modeling and analysis that follow.

2. Pre-Analysis Plan

2.1. Method Overview

Our main objective is to create predictive models that integrate socioeconomic, environmental, and individual data to estimate the likelihood of adult obesity in the United States. We will build a supervised learning pipeline that consists of feature engineering, data preprocessing, model training, and model evaluation in order to accomplish this. Predictors include individual-level health behaviors and demographic traits, as well as environmental conditions at the community level. The outcome variable, obesity status, is defined as a binary indicator based on Body Mass Index (BMI ≥ 30).

The project will go through a number of stages. The first phase is data wrangling. This involves cleaning and organizing raw survey and contextual data, dealing with inconsistent or missing entries, and transforming categorical variables into numerical formats appropriate for modeling. Because the environmental datasets in the final implementation were aggregated to the state level rather than the county level, merging across datasets relies on consistent state identifiers rather than FIPS codes. Exploratory data analysis (EDA) will then be used to identify outliers, visualize variable distributions, summarize important aspects of the data, and investigate possible connections between predictors and obesity outcomes. Following the complete preparation of the data, the modeling phase will concentrate on creating classification models that forecast the risk of obesity using the processed dataset. To find strategies that offer the best balance between interpretability and predictive accuracy, both baseline and more adaptable models will be tested.

Once the models are trained, they will be validated and evaluated to measure performance and determine generalizability to new data. Accuracy-based metrics will be used to compare various approaches, along with ROC curves and AUC scores to assess classification performance. The final step in the model interpretation process will be to summarize the environmental and individual factors that have the greatest impact on the risk of obesity. This methodical approach preserves a transparent and repeatable modeling process while relating individual health data to more general social and environmental factors.

2.2. Models to Be Used and Justification

In order to predict obesity status, we will implement models including k-Nearest Neighbor (kNN), Decision Tree classification, and Logistic Regression. These approaches complement one another when it comes to interpretability and flexibility, making them suitable for modeling a complex outcome like obesity status. First, the kNN model will serve as an intuitive baseline for classification, as it does not assume a fixed functional form between predictors and outcomes. Instead it works by assigning a label to each observation based on the majority class among its closest neighbors in the space, allowing the model to capture local patterns and nonlinear relationships across behavioral and demographic factors.

Decision trees divide the predictor space into a hierarchy of decision rules that identify important variable interactions. The visual structure of decision trees allows for clear interpretation of how different conditions, such as limited food access or low physical activity, combine to increase the likelihood of obesity.

Logistic Regression provides a more classical parametric model that directly estimates the relationship between predictors and the log-odds of obesity. Because logistic regression offers interpretable coefficients and well-understood statistical properties, it serves as an essential comparison point alongside the nonparametric and tree-based models.

Together, these three models offer a balanced combination of interpretability and predictive power and are suitable for explaining findings to a public-health audience.

Unsupervised learning techniques will also be employed in order to enrich the understanding of the data. Principal Components Analysis (PCA) will be used to identify key latent factors, particularly among socioeconomic and environmental indicators. Reducing dimensionality through PCA helps simplify high-dimensional features while retaining substantial explanatory variance. Clustering analysis, specifically k-Means clustering, will also be used to explore the possible existence of distinct population segments with similar behavioral and environmental characteristics. Al-

though clustering does not directly predict obesity, it can reveal natural groupings—such as communities with comparable food environments or activity patterns—that provide valuable context for interpreting the supervised models.

2.3. Model Training Procedure

To train the models, we will prepare the data for analysis, select the appropriate model hyperparameters, and then tune the models to optimize performance. Prior to training, all predictors including demographic variables, behavioral measures, and food environment indicators will be standardized to have a mean of zero and a standard deviation of one. This ensures that all variables contribute equally to distance-based algorithms such as kNN. Because the final dataset contains no nominal categorical variables after preprocessing, one-hot encoding is not required; instead, the dataset consists entirely of numeric features. Missing values in continuous variables will be replaced using mean imputation, while binary or categorical indicators will be imputed using their most frequent category. Implausible BMI values and extreme survey responses will be removed to ensure that the model is trained on realistic and representative cases.

Once the data is cleaned, the dataset will be randomly divided into training and testing subsets, with 80% allocated for training and 20% for testing. Stratified sampling will be used to ensure balanced representation of obese and non-obese individuals. Within the training set, 5-fold stratified cross-validation will be used to identify optimal hyperparameters for each model. For kNN, the number of neighbors and weighting schemes will be evaluated. For the Decision Tree classifier, parameters such as maximum depth, splitting criteria, and minimum leaf sizes will be optimized through grid search. For Logistic Regression, hyperparameters such as regularization strength (C), solver choice, and maximum iterations will be tuned to achieve stable convergence and maximize classification accuracy.

Unsupervised models will also be trained on the standardized data. Principal Components Analysis (PCA) will be used to identify latent components, retaining those that capture the greatest proportion of variance. Additionally, k-Means clustering will be used to examine population segments, with the optimal number of clusters selected using silhouette scores.

All training and analysis will be conducted in Python using libraries such as pandas and numpy for data management and scikit-learn for modeling and hyperparameter optimization. The configurations, random seeds, cross-validation results, and final model parameters will be documented to ensure reproducibility.

2.4. Validation Plan

We will use a validation framework that divides the dataset into training and testing subsets by random allocation in order to evaluate model generalization. While the training phase fits model parameters, the testing phase provides an impartial evaluation of predictive performance on unseen observations. This partitioning approach ensures that we can distinguish true signal from overfitting.

Classification accuracy will serve as a core evaluation metric, supplemented by ROC curves and AUC to assess the models' ability to distinguish between obese and non-obese individuals. Confusion matrices will be used to visualize classification errors across categories and identify systematic patterns of misclassification. We will also compare performance across different model configurations, such as varying values of k in kNN or different regularization levels in logistic regression, to diagnose potential overfitting or underfitting.

After establishing adequate performance on the training data, the models will be evaluated on the held-out test set to estimate real-world generalization accuracy. Consistency between training and testing metrics indicates successful generalization, whereas substantial discrepancies may require adjustments to preprocessing or hyperparameter selection. Throughout the analysis, data handling, model fitting, and validation procedures will be conducted transparently to support replicability and ensure adherence to accepted modeling practices.

3. Results

3.1. Model Implementation and Training

The cleaned dataset contained 130,518 individuals after preprocessing, and the modeling process followed the pipeline described in the pre-analysis plan. The data was divided into an 80–20 stratified training and testing split in order to preserve the proportion of obese and non-obese individuals. This resulted in 104,414 training observations and 26,104 testing observations, with both subsets maintaining an identical obesity prevalence of 30.02%. Before fitting any models, all predictor variables were standardized using z-score normalization, so that features with larger numerical ranges such as age, income, or county-level densities would not dominate the distance calculations for methods like kNN. For example, the AGE variable had a training-set mean of 55.58 and a standard deviation of 17.36 before standardization, which transformed to approximately zero mean and unit variance afterward.

Caleb

Three supervised learning models were trained: k-Nearest Neighbors, a Decision Tree classifier, and Logistic Regression. Each model underwent hyperparameter tuning through

five-fold stratified cross-validation. For kNN, the search grids included multiple values of k, weighting schemes, and a fixed Euclidean distance metric. For the Decision Tree, they included a wide range of tree depths, leaf sizes, and splitting criteria. And for Logistic Regression, they included various regularization strengths, solvers, and iteration limits. Cross-validation ensured that each selected model represented its most stable and accurate configuration on the training data before final evaluation.

3.2. Evaluation Benchmarks

The three optimally tuned models were evaluated on the held-out test set using accuracy, ROC AUC, confusion matrices, and classification reports. These metrics allowed for a comprehensive assessment of how well each classifier generalized to unseen data. Accuracy reflected the overall proportion of correct predictions, while ROC AUC provided a measure of the model's ability to rank obese against non-obese individuals. The confusion matrices revealed how well each classifier distinguished between the two classes, particularly given the imbalance toward non-obese cases.

The best-performing configurations were a kNN model with fifteen neighbors, uniform weighting, and Euclidean distance, a Decision Tree with a maximum depth of seven and a minimum leaf size of two, and a Logistic Regression model using an L2 penalty with a regularization parameter of 0.01 and a maximum of 300 iterations. Cross-validation accuracies ranged from approximately 0.689 to 0.701, which indicated consistent performance across models and folds.

3.3. Overall Model Performance

Across all three classifiers, performance on the test set was relatively similar, with accuracies clustering in the 0.69 to 0.70 range. The Decision Tree achieved the highest accuracy at 69.99 percent and the highest ROC AUC at 0.6346. Logistic Regression followed with a test accuracy of 69.93 percent and an AUC of 0.6152, while kNN achieved 68.94 percent accuracy and an AUC of 0.6139. These results indicate that the Decision Tree captured slightly more predictive signal than the other models, although the differences were modest.

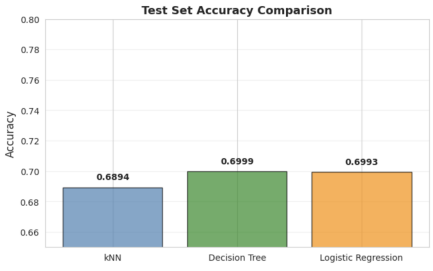

Figure 1. Test Set Accuracy Comparison

The accuracy and ROC AUC bar charts visually reinforced this observation, as the Decision Tree bars were consistently slightly higher than those of kNN and Logistic Regression. The ROC curves further illustrated this pattern, as the Decision Tree curve remained above the others across most thresholds, demonstrating its superior ranking capability. Although none of the models achieved exceptionally high discriminative performance, each performed noticeably better than random guessing, and their curves showed consistent separation between obese and non-obese individuals.

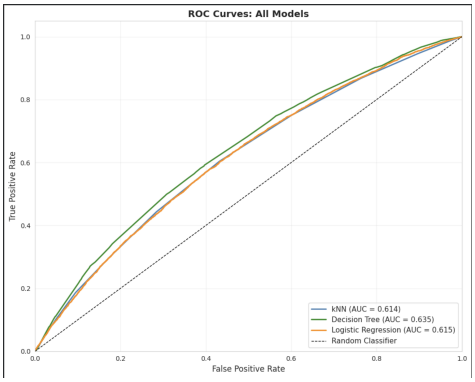

Figure 2. ROC Curves: All Models

3.4. Model-Specific Performance Diagnostics

The confusion matrices highlighted a critical aspect of model behavior. All three models performed strongly on the non-obese class, with recall values exceeding 95 percent, but they struggled to correctly identify obese individuals. Logistic Regression performed the poorest in this respect, correctly identifying fewer than three percent of obese individuals, while kNN achieved approximately nineteen percent recall for the obese class, and the Decision Tree achieved slightly above eleven percent. The classification reports confirmed these findings, as the positive class (obese) exhibited notably low F1-scores across all models.

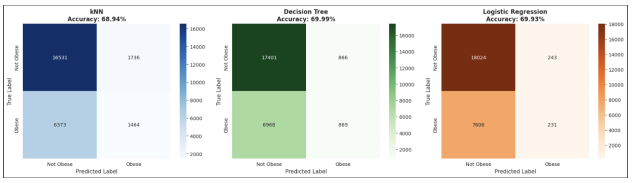

Figure 3. Confusion Matrices

These patterns suggest that class imbalance, combined with substantial overlap in the predictor distributions between obese and non-obese individuals, significantly affected model sensitivity. Logistic Regression, in particular, defaulted heavily toward the majority class despite tuning

the regularization parameter. The Decision Tree provided somewhat better sensitivity, indicating that nonlinear splits captured additional distinctions that linear decision boundaries could not.

3.5. Feature Importance Analysis

An examination of feature importance helped clarify the variables most strongly associated with obesity likelihood. The Decision Tree importance scores and the Logistic Regression standardized coefficients jointly revealed that both behavioral and environmental characteristics contributed meaningfully to prediction. Exercise frequency emerged as the most influential predictor for both models, with reduced physical activity associated with higher obesity risk. Features capturing limited food access, such as the percentage of the population living in low-access areas, also showed strong contributions and were particularly elevated in the Decision Tree analysis.

Race-specific indicators, especially the proportion of individuals identifying as Black, appeared as notable predictors, consistent with documented disparities in obesity prevalence. Measures of grocery store availability, full-service restaurant density, income levels, and physical activity minutes additionally ranked near the top of both importance lists. Logistic Regression coefficients clarified the directionality of associations, whereas the Decision Tree revealed nonlinear interactions among environmental and behavioral factors. Together, the two models suggested that obesity risk is shaped jointly by lifestyle, socioeconomic status, and broader structural food-environment conditions.

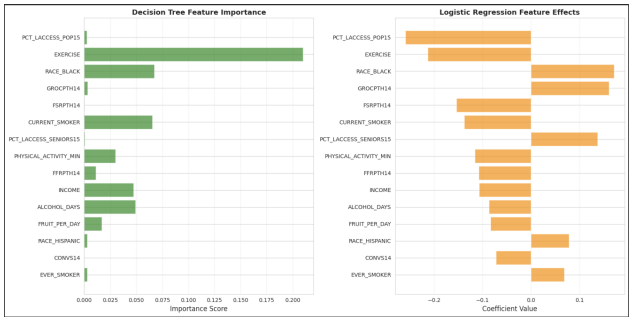

Figure 4. Feature Importance Comparison

3.6. Interpretation and Takeaways

The early findings demonstrate that obesity risk can be predicted with modest accuracy using a combination of demographic, behavioral, and environmental features. While overall accuracy was reasonably high, all models struggled with sensitivity to the obese class, suggesting that the distributions of predictors for obese and non-obese individuals substantially overlap. The Decision Tree consistently produced the strongest results and therefore appears to capture complex interactions that the other models could not fully detect. The analysis further confirms that environmental indicators such as food access and store density meaningfully contribute to prediction, reinforcing the importance of structural determinants of health. Behavioral factors remained the strongest predictors across all models, which aligns with public-health literature emphasizing the roles of physical activity and diet-related behaviors.