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 need 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 specific survey collects nationwide self-reported health behaviors, including diet, physical activity, smoking, alcohol use, and even basic demographic characteristics such as age, sex, race or ethnicity, and income. BMI is calculated directly from reported height and weight values. BRFSS is a valuable foundation for this project because it provides large-scale data that links personal health behaviors with obesity outcomes.

To complement these survey responses, we also incorporate measures of the food environment from the United States Department of Agriculture's Food Environment Atlas. This dataset provides contextual variables such as grocery store availability, fast-food restaurant density, and indicators of food access. These measures are important because obesity is often associated not only with individual choices but also with the accessibility of healthy food in one's environment. Finally, we integrate demographic and socioeconomic indicators from the U.S. Census Bureau's American Community Survey. These data include measures of household income, educational attainment, employment levels, and population density, which allow us to account for the broader socioeconomic context in which health behaviors occur.

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 obtained from the USDA and Census datasets. Together, the combined dataset provides a comprehensive view that takes into account both personal health decisions and structural determinants of health.

Preparing these data for analysis presents several challenges. First, survey responses from BRFSS include missing values, non-standard entries such as “Don't know” or “Refused,” and implausible measures of height or weight that can distort BMI calculations. These issues must be addressed through data cleaning, imputation, and the removal of extreme outliers. Second, merging across datasets is complex because BRFSS data are reported at the individual level, whereas the USDA and Census data are aggregated at the county level. This requires a careful geographic linkage using county FIPS codes, ensuring that respondents are matched to the appropriate community-level measures. Third, many variables require recoding and standardization. For example, income categories, education levels, and frequency of physical activity must be transformed into usable numerical or binary indicators, while continuous predictors such as median household income or food outlet density must be normalized to allow comparability across regions.

The volume and complexity of the data further complicate preparation. BRFSS contains hundreds of thousands of responses each year, making it necessary to implement efficient filtering and sampling strategies to manage computational resources. Integrating multiple years of BRFSS data with county-level measures also raises the risk of inconsistencies or duplicate entries that must be resolved before modeling. Our cleaning plan involves constructing BMI values from reported height and weight, excluding unrealistic cases, recoding categorical responses, and merging external

[*]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. Copyright 2025 by the author(s).

datasets to build a unified analytical file. Missing values will be handled through imputation methods appropriate to the data type, while continuous features will be standardized to support model training.

Through this process, we will create a structured dataset that links individual health behaviors to environmental and socioeconomic conditions. This approach enables 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 will prepare a reliable foundation for the predictive modeling and analysis that will 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 and economic conditions at the community level. The outcome variable, obesity status, is defined as a binary indicator based on Body Mass Index (BMI $\geq$ 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, combining datasets using county-level FIPS codes, and transforming categorical variables into numerical formats appropriate for modeling. 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. 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) and Decision Tree classification. 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. This is possible 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. Through utilizing this method, we will be able to capture local structure in the data and find subtle nonlinear relationships between lifestyle, environmental, and demographic variables and obesity. Differently, decision trees divide the predictor space into a hierarchy of decision rules that identify important variable interactions. The visual structure of decision trees allow for clear interpretation of how different conditions like limited food access or low physical activity combine to increase the chances of obesity. Both models are interpretable and suitable for explaining findings to a public-health audience.

In addition to these supervised models, unsupervised learning techniques will be employed in order to enrich the understanding of the data. First, Principal Components Analysis (PCA) will be used to identify key latent factors, underlying groups of corrected predictors, particularly socioeconomic and environmental indicators. Reducing dimensionality through PCA will help to simplify the dataset and improve computational efficiency. This will happen while retaining the majority of the explanatory variance. Next, Clustering analysis, specifically k-Means clustering, will also be used in order to explore the possible existence of distinct population segments with similar behavioral and environmental characteristics. Clustering may not be able to directly predict obesity. But, it can possibly reveal natural groupings like communities that share comparable food environments or activity patterns that provide valuable context for interpreting the supervised models.

2.3. Model Training Procedure

To train the model, we will prepare the data for analysis, select the appropriate model hyperparameters, and then tune the model to optimize performance. Prior to training, all predictors, such as age, income, and food outlet density, 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 regardless of their size. one-hot encoding will be applied to nominal variables such as gender, race, and region to prevent introducing false ordinal relationships, while ordinal encoding will be used for naturally ordered variables like education level or income category. Missing values will be handled depending on their variable type. Continuous variables like income or BMI will be replaced with their mean values, whereas categorical variables like smoking status will be replaced with their most frequent category. This will help to retain valuable data while minimizing the bias from missing values. Outlier values in BMI, height, or weight will be excluded or capped 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. 80% of the data will be allocated for training and the remaining 20% will be for testing. Stratified sampling will be used to ensure that both obese and non-obese cases are included in both subsets. Within the training set, 5-fold cross-validation will be used to identify optimal hyperparameters for each model. For kNN, the number of neighbors k will vary from 3 to 25, and the distance metrics will be tested to determine the optimal setup. For the Decision Tree classifier, parameters such as maximum depth, minimum samples per leaf, and splitting criteria will be optimized using grid search, with pruning applied to prevent overfitting. The optimal model will be chosen based on cross-validation accuracy and stability across folds.

Unsupervised models will also be trained on the standardized data. Principle Components Analysis (PCA) will be used to identify the main latent factors. Components explaining the majority of the variance will be retained to avoid the loss of significant information. Additionally, k-Means clustering with be used to detect population subgroups with shared traits. The number of clusters will vary between 3 and 10, and the optimal value will be selected using silhouette scores and within-cluster sum of squares (WCSS).

All training and analysis will be completed through Python, using libraries such as *pandas* and *numpy* for data management and *scikit-learn* for modeling and parameter tuning. 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 testing phase provides an impartial evaluation of predictive performance on observations that haven't been seen before, the training phase fits model parameters. We will be able to differentiate between true pattern recognition and simple training example memorization thanks to this partitioning technique.

Classification accuracy, or the percentage of cases that are correctly classified, is the main metric we would use to assess model performance. This measure provides a clear interpretation of how well the model differentiates between obese and non-obese individuals. By showing the distribution of classification errors across categories, confusion matrices supplement this one-number summary and enable the detection of systematic biases toward specific groups. Through visual comparison of training and testing trajectories, we will also analyze performance across different levels of model complexity, such as different values of k in k-nearest neighbors classification, to identify possible overfitting or underfitting.

After establishing adequate performance on training data, we would apply the model to the test set to obtain an estimate of generalization accuracy. While significant differences call for reevaluating the model specification or data preprocessing techniques, such as feature scaling and categorical encoding, consistency between training and testing performance indicates successful generalization. To guarantee transparency and conformity to accepted modeling practices, we handle and visualize the data consistently throughout the analysis. This validation method supports the choice of a trustworthy and understandable predictive model by striking a balance between thorough performance evaluation and useful diagnostic tools.