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 datasets to build a unified analytical file. Missing values

[*] Equal contribution [1] Department of Data Science, University of Virginia, Charlottesville Virginia, United States. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

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.