Uncovering Determinants of Obesity via Feature Importance: Machine Learning Models for Population-Level BMI Prediction

Shashank Joshi
Department of Computer Science
University of Southern California
ssjoshi@usc.edu

James Simon
Department of Computer Science
University of Southern California
jcsimon@usc.edu

Phani Yalamanchili
Department of Computer Science
University of Southern California
phanirat@usc.edu

Emma Yue
Department of Computer Science
University of Southern California
emmayue@usc.edu

Abstract

Obesity represents a critical global health challenge, yet scalable population-level screening remains limited by barriers to clinical access. This research addresses this gap by uncovering key determinants of obesity through systematic feature importance analysis of survey-based data, enabling machine learning models to predict BMI at population scales without requiring clinical visits. Using the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset with 414,633 valid samples, we employ multi-method feature selection to identify 16 optimal predictors from 302 available variables. We evaluate seven machine learning models spanning linear, tree-based, and ensemble approaches on a balanced 100,000-sample dataset. Our Stacking Ensemble achieves strong performance with regression metrics (RMSE: 3.32, MAE: 2.26 BMI points, R^2: 0.8858), classification metrics (Accuracy: 0.8879, Precision: 0.8881, Recall: 0.8875, F1: 0.8878, AUC-ROC: 0.9588), and category metrics (Category Accuracy: 0.6505, Within-1-Category: 0.9631). Comprehensive fairness evaluation reveals race/ethnicity as the primary source of disparity (Δ MAE = 0.86 BMI points), followed by income (Δ MAE = 0.79). Gender disparities are moderate (Δ MAE = 0.42), while age disparities are minimal (Δ MAE = 0.12).

1 Introduction

1.1 Motivation

Obesity represents one of the most significant public health challenges of the 21st century, ranked as the fifth foremost reason for death globally Safaei et al. [2021a]. Global numbers reached 641 million obese adults in 2014, compared to only 105 million in 1975, showing an alarming 6-fold increase over four decades Safaei et al. [2021a]. Obesity contributes to numerous chronic diseases including cancers, diabetes, metabolic syndrome, and cardiovascular diseases. Early detection and intervention are crucial, yet traditional clinical BMI measurement requires in-person healthcare visits with physical measurements, creating barriers to timely assessment. This research addresses this gap by developing models that work across the entire US population using easily accessible survey features, enabling scalable screening without requiring in-person clinical visits. We distinguish this from population-level statistics (which compute prevalence or aggregate measures on populations) by noting that our models predict individual BMI values, but do so using features that are easily accessible through surveys rather than requiring clinical measurements.

1.2 Problem Statement

Current approaches to obesity screening face significant limitations. Traditional health condition diagnosis requires patients to visit physicians regularly, making the identification process tedious at early stages Butt

et al. [2021]. Clinical BMI measurement typically requires in-person healthcare visits with physical measurements, creating barriers to timely assessment. Current healthcare systems face challenges in providing scalable health screening and surveillance. Scalable tools for predictive analytics are needed to accurately stratify patients into risk subgroups using easily accessible features Topol [2019].

Survey-based data and machine learning approaches show potential, yet significant research gaps remain. Most obesity prediction studies rely on clinical measurements or specialized datasets that require in-person healthcare visits, limiting their applicability to scalable screening using easily accessible features. Limited studies have investigated how machine learning techniques can predict adult obesity using survey data, and comprehensive fairness evaluation across demographic groups remains lacking Chen et al. [2023].

1.3 High-Level Outline of Goals and Approaches

This research aims to develop and evaluate machine learning models for BMI prediction using survey-based data from the Behavioral Risk Factor Surveillance System (BRFSS), demonstrating the feasibility of scalable obesity screening using easily accessible survey features (rather than clinical measurements) while ensuring equitable performance across diverse demographic groups.

Our approach employs a comprehensive methodology organized into three major stages. Stage 1 is Data Preparation, where we load the BRFSS dataset, preprocess it by handling missing values, and create a balanced dataset of 100,000 samples with equal representation across BMI categories. Stage 2 is Model Development, where we perform feature selection on the balanced dataset to identify optimal predictors, then perform stratified train-test splitting (80,000 training, 20,000 test) and train seven different models ranging from simple linear baselines to complex ensemble methods to identify the optimal approach. Stage 3 is Evaluation and Analysis, where we comprehensively evaluate model performance using multiple metrics, conduct fairness analysis across demographic groups, and generate visualizations and reports for clinical interpretation.

We train and evaluate seven different models to identify the optimal approach. Our baseline models include Ridge and Lasso regression for linear baselines and Decision Tree for a simple non-linear baseline. Our advanced models include Random Forest, which handles non-linear relationships and feature interactions well, and Gradient Boosting, which sequentially improves predictions. Finally, we implement two ensemble methods: Voting Regressor, which averages predictions from multiple base models, and Stacking Regressor, which uses a meta-learner to optimally combine base model predictions. For evaluation, we use a comprehensive set of metrics across multiple dimensions: regression metrics (RMSE, MAE, and R-squared) for continuous prediction accuracy, classification metrics (accuracy, precision, recall, F1-score, AUC-ROC, and Brier score) for binary classification performance, category-specific metrics for exact match and within-one-category accuracy, tolerance-based metrics measuring predictions within 2 or 3 BMI points, and fairness metrics evaluating performance across four demographic dimensions.

2 Related Work / Background

2.1 Papers Using the BRFSS Dataset

The Behavioral Risk Factor Surveillance System (BRFSS) is the world's largest telephone health survey system, tracking health conditions and risk behaviors in the United States annually since 1984 Centers for Disease Control and Prevention [2024]. The 2024 dataset contains responses from approximately 457,000 individuals across all 50 states, the District of Columbia, and participating US territories, covering 302 demographic, behavioral, and health-related variables. The BRFSS has been extensively used in health research, including studies of chronic diseases, health-related quality of life, and demographic disparities.

Brown et al. [2004] used BRFSS data to examine the relationship between diabetes mellitus and health-related quality of life among older adults, demonstrating the utility of BRFSS for population-level health outcome analysis. Bhan et al. [2015] employed BRFSS data to analyze time trends in racial and ethnic

disparities in asthma prevalence, highlighting the dataset's value for fairness and equity research. However, to our knowledge, no previous studies have used BRFSS specifically for BMI prediction using machine learning methods. This represents a critical gap: while BRFSS has proven valuable for health research and contains comprehensive BMI data, its potential for ML-based individual-level BMI prediction has remained unexplored, representing a gap in the literature that this study addresses.

2.2 Papers on BMI Prediction

Several studies have investigated BMI prediction using machine learning methods, though most rely on clinical measurements or specialized datasets rather than survey data. Yamada et al. [2020] developed models to predict BMI distributions using national survey data from Mexico, Colombia, and Peru, demonstrating the feasibility of survey-based BMI prediction, though their approach focused on population-level distributions rather than individual-level prediction. Safaei et al. [2021b] conducted a systematic review of 93 papers on machine learning approaches for obesity prediction from 2010 to 2020, finding that ensemble methods show particular promise for obesity prediction tasks. However, most studies reviewed by Safaei et al. primarily used clinical features from electronic health records (EHR) or specialized datasets, which require in-person healthcare visits. This creates a fundamental scalability barrier: clinical data collection requires patients to visit healthcare facilities, making it impractical for population-wide screening. Survey-based approaches, in contrast, can reach individuals without clinical barriers, enabling scalable screening at population scales. This distinction motivates our use of survey data from BRFSS rather than clinical measurements.

2.3 Papers on Ensemble Methods in Healthcare

Ensemble learning approaches have shown particular promise in health prediction tasks. Safaei et al. [2021b] identified ensemble methods as particularly effective for obesity prediction in their systematic review of 93 papers. Ooba et al. [2023] used Light Gradient Boosting Machine (LightGBM), an ensemble method, to predict subjective well-being in pregnant women using survey data, achieving 84% accuracy on 2020 data and 88% accuracy when applied to 2021 data. While this study used gradient boosting rather than stacking, it demonstrates the effectiveness of ensemble methods for health prediction from survey data, supporting our methodological choice. Our study extends this work by directly comparing voting and stacking ensemble methods for BMI prediction, finding that stacking ensembles achieve superior performance by using a meta-learner to optimally combine base model predictions.

2.4 Related Health Prediction Studies

Machine learning techniques have shown remarkable success in healthcare applications, though most studies use clinical features rather than survey data. Butt et al. [2021] demonstrated the effectiveness of ML for diabetes classification and prediction, achieving 86.08% accuracy using a Multilayer Perceptron (MLP) for classification and 87.26% accuracy using Long Short-Term Memory (LSTM) networks for prediction. While this study focuses on diabetes prediction rather than BMI prediction, diabetes and BMI are closely correlated, making this methodology relevant for BMI prediction. However, their study employed clinical measurements and EHR data, which require in-person healthcare visits, whereas our approach uses easily accessible survey features, enabling scalable population-level screening without clinical barriers.

2.5 Fairness in Healthcare Machine Learning

AI and ML models are increasingly integrated into healthcare systems, making ensuring algorithmic fairness and identifying potential biases critical concerns. Chen et al. [2023] highlight that insufficiently fair AI systems can undermine the delivery of equitable care, with audit studies revealing inequalities in how patients are diagnosed, treated, and billed. For obesity prediction models, fairness evaluation requires systematic subgroup analysis across demographic dimensions including race/ethnicity, sex, age, income, and geographic region. However, few obesity prediction studies have conducted systematic fairness anal-

ysis across multiple demographic dimensions, leaving potential biases unexamined. This study addresses this gap by performing systematic subgroup analysis across multiple demographic dimensions, identifying disparities and investigating their sources through feature importance analysis.

2.6 Research Gaps and Contributions

ML methods have been extensively applied to health prediction tasks, yet several critical gaps remain in the literature on obesity prediction from survey data. First, while BRFSS has been used extensively for health research, no previous studies have applied machine learning methods to BRFSS data for individual-level BMI prediction. Second, most existing BMI prediction studies use clinical features from EHR systems, which require in-person healthcare visits and are not easily accessible for population-level screening, creating scalability barriers. Third, limited studies have investigated advanced ensemble methods, particularly stacking methods, for BMI prediction from survey data. Fourth, fairness analysis in obesity prediction models is insufficient, with few studies systematically evaluating model performance across demographic groups.

This study addresses these gaps through four key contributions. First, we present the first ML-based BMI prediction study using BRFSS survey data, enabling scalable screening without clinical barriers. Second, we comprehensively evaluate seven ML models including stacking ensembles, demonstrating that stacking achieves superior performance by using a meta-learner to optimally combine base model predictions. Third, we perform systematic feature selection from 302 BRFSS variables to identify 16 optimal predictors, enabling interpretable and efficient models. Fourth, we conduct systematic subgroup analysis across multiple demographic dimensions (race/ethnicity, sex, age, income, geographic region), identifying disparities and investigating their sources through permutation feature importance analysis. Together, these contributions demonstrate the feasibility of survey-based obesity screening at population scales while ensuring equitable performance across diverse demographic groups.

The following sections detail our methodology for addressing these gaps. We begin by describing the BRFSS dataset and our preprocessing approach, which enables us to leverage survey data for ML-based BMI prediction. We then present our feature selection methodology, which identifies optimal predictors from the 302 available BRFSS variables. Finally, we describe our model development and evaluation framework, which comprehensively assesses both predictive performance and fairness across demographic groups.

3 Data

3.1 Description

This study utilizes the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset, which represents a nationally representative health survey. This dataset provides comprehensive population-level data on health behaviors, chronic conditions, and social determinants of health. The BRFSS represents the world's largest telephone health survey system, tracking health conditions and risk behaviors in the United States annually since 1984 Centers for Disease Control and Prevention [2024].

3.2 Data Sources

The Behavioral Risk Factor Surveillance System is a collaborative project. It involves US states, territories, and the Centers for Disease Control and Prevention (CDC). It is administered by CDC's National Center for Chronic Disease Prevention and Health Promotion Centers for Disease Control and Prevention [2024]. The survey employs a dual-frame methodology. It conducts interviews using both landline phones and cell phones. All responses are self-reported Centers for Disease Control and Prevention [2024].

3.3 Data Statistics on Data Source Size

The 2024 BRFSS dataset contains responses from approximately 457,000 individuals. These come from all 50 states, the District of Columbia, and participating US territories. The dataset covers 302 demographic, behavioral, and health-related variables. This provides comprehensive population-level data suitable for machine learning model development.

3.4 Preprocessing

BMI categories are defined according to WHO/CDC standards: Underweight (BMI < 18.5), Normal ($18.5 \leq$ BMI < 25), Overweight ($25 \leq$ BMI < 30), Obese Class I ($30 \leq$ BMI < 35), Obese Class II ($35 \leq$ BMI < 40), and Obese Class III (BMI ≥ 40), as shown in Table 1.

The BRFSS 2024 dataset is processed through a standardized pipeline. The _BMI5 variable is converted to standard BMI values, with invalid values (BMI < 12 or ≥ 100) removed, resulting in approximately 414,633 valid samples. We create a balanced dataset of 100,000 samples with equal representation from each of the six BMI categories (16.7% each), ensuring balanced representation across all categories for model training and evaluation. Feature selection is performed on the training set only (after train-test splitting) to identify the optimal 16 features (see Section 4.1 for details), following best practices to prevent data leakage by ensuring the test set is never used for feature selection.

Missing values are handled through median imputation for numerical features, ensuring that all samples can be included in the analysis while preserving the distributional characteristics of the data. All features are standardized using StandardScaler, which transforms features to have zero mean and unit variance.

Multiple derived variables are created from the continuous BMI values to support multi-task evaluation. A binary obesity classification is created using the threshold BMI ≥ 30.

Category	**BMI Range**
Underweight	BMI < 18.5
Normal	$18.5 \leq$ BMI < 25
Overweight	$25 \leq$ BMI < 30
Obese Class I	$30 \leq$ BMI < 35
Obese Class II	$35 \leq$ BMI < 40
Obese Class III	BMI ≥ 40

Table 1: BMI categories based on WHO/CDC standards

Data splitting uses an 80-20 stratified split by BMI category with random_state=42. Stratification ensures that each BMI category maintains proportional representation in both training and test sets. Within each BMI category, samples are randomly selected to achieve balanced representation. After splitting, the training set contains 80,000 samples and the test set contains 20,000 samples, with each BMI category maintaining 16.7% representation in both sets. Survey weights (_LLCPWT) are preserved and applied during training to reflect population-level distributions.

3.5 Data Statistics of Processed Dataset

After preprocessing and filtering, the final processed dataset contains 100,000 samples with valid BMI values. The dataset uses an 80-20 stratified split, resulting in 80,000 training samples and 20,000 test samples. The target variable is continuous BMI (range: 12-100 kg/m^2). Table 2 presents the comprehensive dataset statistics, including original BRFSS percentages, subsampled percentages, and train-test distribution across BMI categories.

BMI Category	Original BRFSS	100K Balanced	Training Set	Test Set
Underweight	1.78%	16.7%	13,333 (16.7%)	3,333 (16.7%)
Normal	29.20%	16.7%	13,333 (16.7%)	3,333 (16.7%)
Overweight	35.35%	16.7%	13,333 (16.7%)	3,333 (16.7%)
Obese Class I	19.81%	16.7%	13,333 (16.7%)	3,333 (16.7%)
Obese Class II	8.21%	16.7%	13,333 (16.7%)	3,333 (16.7%)
Obese Class III	5.66%	16.7%	13,333 (16.7%)	3,333 (16.7%)
Total	**100%**	**100%**	**80,000 (100%)**	**20,000 (100%)**

Table 2: Comprehensive dataset statistics for the Balanced Dataset (100K) showing original BRFSS percentages (from 414,633 valid samples), subsampled percentages (balanced to 16.7% each via oversampling Underweight), and train-test distribution. Original percentages show that Normal (29.20%) and Overweight (35.35%) categories were under-sampled, while Underweight (1.78%) and Obese Class II/III (8.21%/5.66%) were heavily over-sampled to achieve balanced representation. This is the dataset used for the main analysis (Sections 4-7).

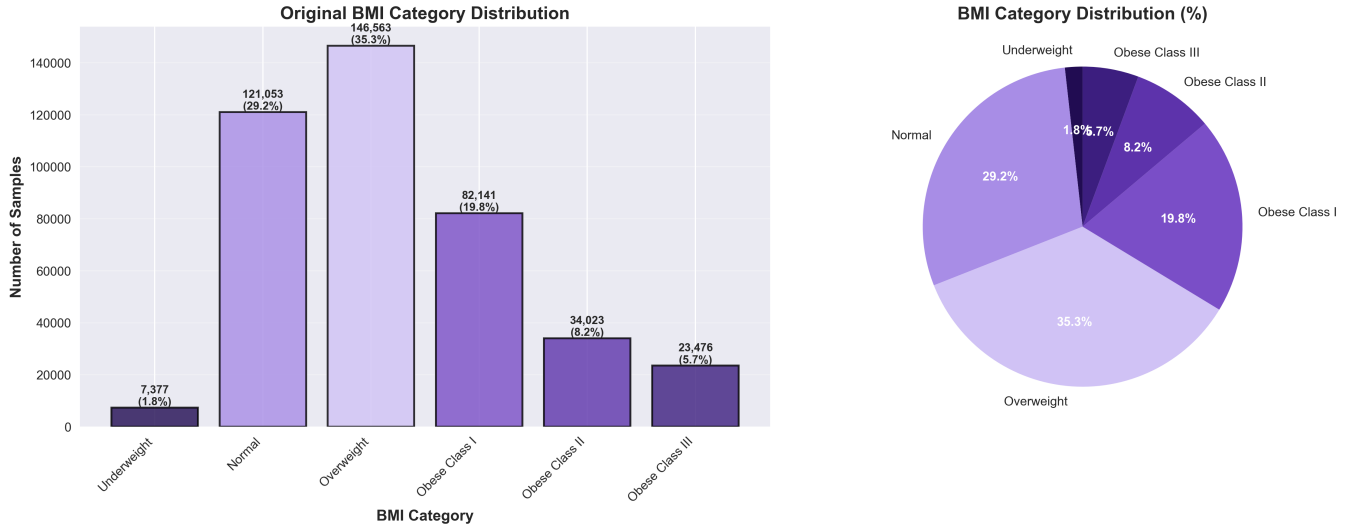

Figure 1: Original BMI category distribution in the BRFSS 2024 dataset (414,633 valid samples). The distribution is highly imbalanced, with Overweight (35.35%) and Normal (29.20%) categories dominating, while Underweight (1.78%) and Obese Class III (5.66%) are underrepresented.

To evaluate the impact of dataset configuration on model performance, we train the same Stacking Ensemble model on all three datasets using identical hyperparameters and an 80-20 stratified train-test split. All metrics (RMSE, MAE, R^2) are calculated on the test set only to ensure fair comparison and prevent data leakage.

The Original (414K) dataset achieves $R^2 = 0.1499$, which is low but mathematically expected given the severe class imbalance (Overweight = 35.35%, Normal = 29.20%, Underweight = 1.78%). This imbalance causes regression models to predict near the population mean BMI (approximately 27.5), resulting in large errors for minority classes (Underweight, Obese Class II/III) and low explained variance. The Undersampled Balanced Dataset (44K) achieves $R^2 = 0.1905$, slightly better than Original due to class balance, but still low because 7,377 samples per category is insufficient for robust learning with complex ensemble models.

The Balanced Dataset (100K) achieves $R^2 = 0.8858$ under the balanced test distribution, with MAE improving from 4.48 to 2.26 BMI points (50% reduction) and RMSE from 6.06 to 3.32 (45% reduction). These improvements demonstrate substantial real performance gains beyond variance effects. Note that

R^2 values are not directly comparable across different test distributions due to variance differences; the balanced 100K test set has higher BMI variance than the imbalanced Original test set, which contributes to the higher R^2 value. However, the dramatic improvements in absolute error metrics (MAE, RMSE) confirm that the model genuinely performs better on the balanced dataset, not merely due to distributional differences.

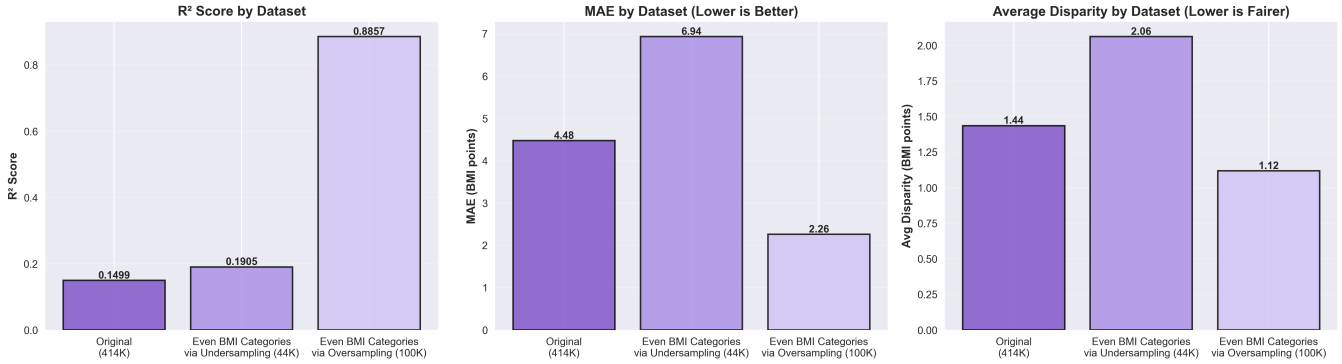

Figure 2: Dataset configuration comparison: R^2 score, MAE, and average disparity across three dataset configurations. The Balanced Dataset (100K, via Oversampling Underweight) achieves the best balance of accuracy and fairness. The Original (414K) dataset achieves $R^2 = 0.1499$ due to severe class imbalance causing models to predict near the population mean. The Undersampled Balanced Dataset (44K) achieves $R^2 = 0.1905$ due to insufficient sample size (7,377 per category) for robust model learning.

4 Task + Approach + Results

This section presents our comprehensive evaluation framework, organized into four main tasks: (1) Feature Selection, which identifies optimal predictors from 302 available variables; (2) BMI Prediction (Regression), which develops and compares seven machine learning models for continuous BMI prediction; (3) Obesity Classification, which evaluates binary and multi-class classification performance; and (4) Fairness Analysis, which systematically assesses model performance across demographic groups. Each task builds upon previous tasks: feature selection informs model development, regression predictions enable classification, and all predictions are evaluated for fairness across demographics.

4.1 Task 1: Feature Selection

The task is to identify an optimal feature set from 302 available BRFSS variables for BMI prediction. The input consists of all 302 BRFSS variables, and the goal is to select a compact set of predictive features that balances model performance with interpretability.

We employ systematic feature selection using multiple complementary methods, with all feature selection performed on the training set only (after train-test splitting) to prevent data leakage. The balanced dataset of 100,000 samples is first split into training (80,000 samples) and test (20,000 samples) sets using stratified sampling. Feature selection is then performed exclusively on the training set. The process involves four sequential steps applied to the training data. First, we compute Pearson correlation coefficients between each feature and BMI to capture linear relationships. Second, we calculate mutual information scores to capture non-linear dependencies between features and BMI. Third, we train a Random Forest model on the training set and compute feature importance scores using the mean decrease in impurity method, which captures complex feature interactions. Fourth, permutation importance analysis is applied as a model-agnostic validation step, where each feature is shuffled and the impact on model accuracy is measured (on training data only). These four methods (Correlation, Mutual Information, Random Forest importance, and Permutation importance) are normalized to a common [0,1] scale. The Combined score

is computed as the mean of the normalized Correlation, Mutual Information, and Random Forest scores. We select the top 16 features based on their combined importance scores, as this number provides an optimal balance between model performance and interpretability: preliminary experiments showed diminishing returns beyond 16 features, with additional features contributing minimal predictive power while increasing model complexity. Permutation importance is used separately to validate the selection and assess discrimination risk for demographic features.

The feature selection process successfully identified 16 optimal features from 302 available variables, representing a 94.7% reduction in dimensionality while maintaining predictive power. The top features by combined importance score are: Age Group (_AGEG5YR, 0.882), Mental Health Days (MENTHLTH, 0.607), General Health (GENHLTH, 0.576), Diabetes Status (DIABETE4, 0.565), and Employment Status (EMPLOY1, 0.527). The selected features span seven categories: Health Status and Medical (4 features), Demographics Age (3 features), Demographics Sex and Race (2 features), Socioeconomic Status (3 features), Family (1 feature), Physical Activity (2 features), and Behavioral Alcohol (1 feature). Table 3 presents all 16 selected features with their importance scores.

Rank	**Feature**	**Description**	**Corr**	**MI**	**RF**	**Combined**
1	_AGEG5YR	Age Group (5-year)	0.1485	1.0144	0.1367	0.882
2	MENTHLTH	Mental Health Days	0.0727	0.8215	0.1086	0.607
3	GENHLTH	General Health	0.1671	0.5008	0.0916	0.576
4	DIABETE4	Diabetes Status	0.2238	0.2385	0.0940	0.565
5	EMPLOY1	Employment Status	0.1605	0.5821	0.0659	0.527
6	_INCOMG1	Income Group	0.0710	0.6840	0.1008	0.526
7	ALCDAY4	Alcohol Consumption	0.0106	0.8024	0.0916	0.454
8	_AGE65YR	Age 65+ Indicator	0.2134	0.2208	0.0235	0.345
9	_AGE_G	Age Group (Detailed)	0.0943	0.5559	0.0280	0.307
10	_RACE	Race/Ethnicity	0.0394	0.4921	0.0639	0.296
11	SDHFOOD1	Food Security	0.0746	0.3379	0.0564	0.267
12	CHILDREN	Number of Children	0.1028	0.3342	0.0383	0.259
13	HAVARTH4	Arthritis Status	0.0975	0.1940	0.0313	0.174
14	_SEX	Biological Sex	0.0153	0.3742	0.0335	0.125
15	EXERANY2	Any Exercise	0.0824	0.1960	0.0183	0.115
16	_TOTINDA	Physical Activity Index	0.0818	0.1961	0.0175	0.112

Table 3: All 16 selected features sorted by Combined importance score. The Combined score integrates normalized scores from Correlation (Corr), Mutual Information (MI), and Random Forest (RF) importance methods. Permutation importance is shown separately in Table 4 for validation. Age group (_AGEG5YR) emerges as the most important predictor.

Rank	Feature	Description	Permutation
1	_AGEG5YR	Age Group (5-year)	0.5365
2	GENHLTH	General Health	0.4121
3	DIABETE4	Diabetes Status	0.3218
4	_INCOMG1	Income Group	0.2543
5	EMPLOY1	Employment Status	0.2482
6	MENTHLTH	Mental Health Days	0.2229
7	ALCDAY4	Alcohol Consumption	0.1992
8	HAVARTH4	Arthritis Status	0.1699
9	_RACE	Race/Ethnicity	0.1433
10	_SEX	Biological Sex	0.1048
11	_AGE_G	Age Group (Detailed)	0.1005
12	SDHFOOD1	Food Security	0.0882
13	CHILDREN	Number of Children	0.0699
14	_AGE65YR	Age 65+ Indicator	0.0521
15	EXERANY2	Any Exercise	0.0455
16	_TOTINDA	Physical Activity Index	0.0426

Table 4: Features sorted by Permutation importance for feature selection validation. Permutation importance measures the decrease in model accuracy when each feature is randomly shuffled, providing a model-agnostic assessment of feature importance. This table validates the feature selection process by showing that age-related features (_AGEG5YR) and health status indicators (GENHLTH, DIABETE4) are consistently important across different evaluation methods.

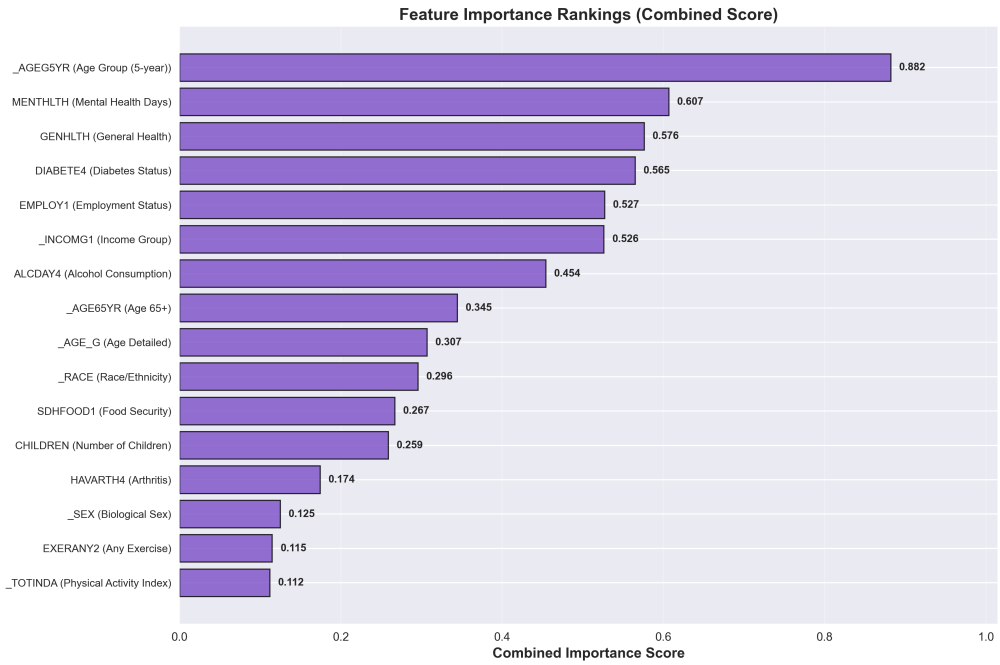

Figure 3: Combined feature importance rankings used for feature selection. Age group (_AGEG5YR) shows the highest combined importance score across correlation, mutual information, and random forest methods.

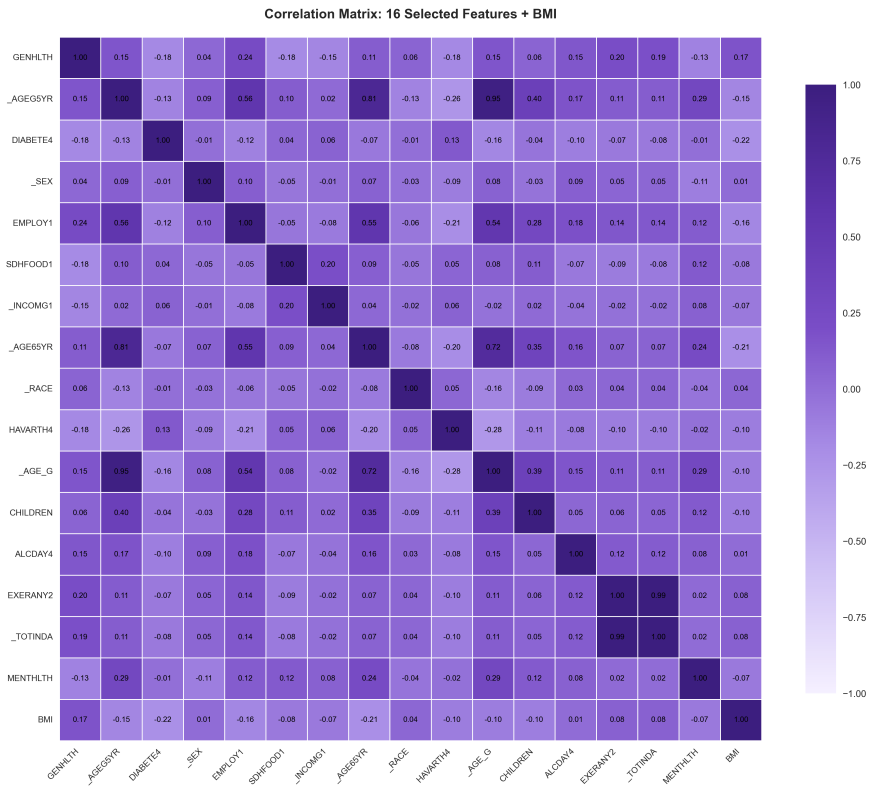

Figure 4: Correlation matrix heatmap for the 16 selected features and BMI. Age-related features (_AGEG5YR, _AGE65YR, _AGE_G) show strong inter-correlations, while most features show low to moderate correlation with BMI.

4.2 Task 2: BMI Prediction (Regression)

The primary task is continuous BMI prediction using the selected 16 features. The input consists of standardized feature vectors (16 features), and the output is predicted BMI values (continuous regression). This task enables scalable BMI estimation across the US population using easily accessible survey features without requiring clinical measurements.

We evaluate seven machine learning algorithms spanning different methodological approaches. Our baseline models include Ridge Regression (alpha=0.1) and Lasso Regression (alpha=0.01) for linear baselines, and Decision Tree (max_depth=15) for a simple non-linear baseline. Our advanced models include Random Forest (n_estimators=100, max_depth=15), which handles non-linear relationships and feature interactions well, and Gradient Boosting (n_estimators=100, learning_rate=0.08, max_depth=10), which sequentially improves predictions. Finally, we implement two ensemble methods: Voting Regressor, which averages predictions from multiple base models (Ridge, Random Forest, Gradient Boosting), and Stacking Regressor, which uses a meta-learner (Ridge with cv=5) to optimally combine base model predictions (Ridge, Random Forest, Gradient Boosting). All models use standardized features and survey weights during training. Models are trained on the training set (80,000 samples) and evaluated on the test set (20,000 samples).

Regression-Based BMI Prediction Our Stacking Ensemble achieves the best regression performance with RMSE of 3.32 BMI points, MAE of 2.26 BMI points, and R^2 of 0.8858, indicating that the model explains 88.58% of the variance in BMI values. This represents clinically acceptable accuracy, as the mean absolute error of 2.26 BMI points falls well within acceptable clinical tolerances for population-level screening. Comparison to baselines reveals substantial improvements. Ridge and Lasso Regression achieve RMSE of 8.96 BMI points, MAE of 7.15 BMI points, and R^2 of only 0.17, demonstrating the limitations

of linear models for this task. Gradient Boosting achieves RMSE of 3.54 BMI points, MAE of 2.44 BMI points, and R^2 of 0.8699, performing nearly as well as the Stacking Ensemble. Table 5 presents regression performance metrics for all seven models plus a mean baseline.

Model	RMSE	MAE	R^2
Stacking Ensemble	3.32	2.26	0.8858
Gradient Boosting	3.54	2.44	0.8699
Random Forest	3.67	2.34	0.8606
Decision Tree	4.49	2.65	0.7909
Voting Ensemble	4.80	3.87	0.7610
Ridge Regression	8.96	7.15	0.1666
Lasso Regression	8.96	7.15	0.1666
Mean Baseline	9.82	8.00	0.0000

Table 5: Regression-Based BMI Prediction: RMSE and MAE are in BMI points. R^2 indicates the proportion of variance explained. The Stacking Ensemble achieves the best R^2 (0.8858) and lowest MAE (2.26). Mean Baseline shows the performance of predicting the training mean for all samples.

Figure 5 shows the regression performance comparison across all seven models using three key metrics (R^2, MAE, and RMSE), demonstrating that ensemble methods substantially outperform baseline models.

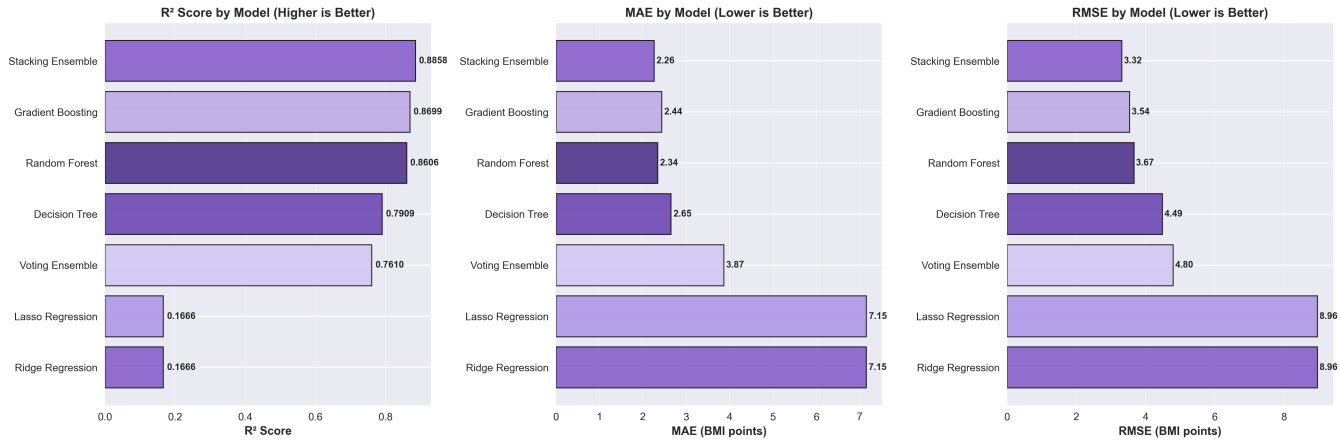

Figure 5: Regression performance comparison across all seven models. R^2 Score (higher is better), MAE (lower is better), and RMSE (lower is better) are shown. Stacking Ensemble and Random Forest achieve the best performance.

4.3 Task 3: Obesity Classification

The task encompasses both binary classification (obese vs. non-obese, BMI ≥ 30 threshold) and multi-class classification (six BMI categories). The input consists of predicted BMI values from Task 2, and the output is binary classification (obese/non-obese) or multi-class classification (six BMI categories).

Classification predictions are derived from continuous BMI predictions through post-processing. Binary classification uses the threshold BMI ≥ 30, while multi-class classification maps continuous BMI to six WHO/CDC categories. All seven models are evaluated on both classification tasks using standard metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC for binary classification, and per-category metrics for multi-class classification.

Multi-Category Classification and Obesity Prediction (Binary Classification) Our Stacking Ensemble achieves superior classification performance for binary obesity classification (obese vs. non-obese,

BMI ≥ 30), with Accuracy of 0.8879, Precision of 0.8881, Recall of 0.8875, F1-Score of 0.8878, and AUC-ROC of 0.9588. Gradient Boosting achieves comparable performance with Accuracy of 0.8801 and AUC-ROC of 0.9525. Table 6 presents comprehensive classification metrics for all seven models.

Model	Accuracy	Precision	Recall	F1	AUC-ROC
Stacking Ensemble	0.8879	0.8881	0.8875	0.8878	0.9588
Random Forest	0.8832	0.8779	0.8902	0.8840	0.9531
Gradient Boosting	0.8801	0.8761	0.8854	0.8807	0.9525
Voting Ensemble	0.8646	0.8564	0.8760	0.8661	0.9412
Decision Tree	0.8577	0.8466	0.8736	0.8599	0.9235
Ridge Regression	0.6562	0.6378	0.7231	0.6778	0.7196
Lasso Regression	0.6564	0.6377	0.7247	0.6784	0.7193

Table 6: Binary Obesity Classification (BMI ≥ 30): Performance metrics for all seven models. Classification is derived from predicted BMI values using the threshold of 30. Stacking Ensemble and Random Forest achieve the highest performance across all metrics.

Category & Tolerance Metrics Our Stacking Ensemble achieves category accuracy of 0.6505 (65.1%) and within-1-category accuracy of 0.9631 (96.3%). Tolerance-based accuracy shows ±2 BMI accuracy of 0.6122 and ±3 BMI accuracy of 0.7428, indicating that the majority of predictions fall within clinically acceptable ranges. Table 7 presents category and tolerance metrics for all seven models.

Model	Cat-Acc	±1-Cat	±2 BMI	±3 BMI
Stacking Ensemble	0.6505	0.9631	0.6122	0.7428
Random Forest	0.6138	0.9502	0.6072	0.7155
Decision Tree	0.6212	0.9073	0.5975	0.6794
Gradient Boosting	0.5858	0.9554	0.5734	0.7015
Voting Ensemble	0.3828	0.9363	0.2799	0.4315
Ridge Regression	0.2084	0.5970	0.1709	0.2510
Lasso Regression	0.2079	0.5966	0.1701	0.2507

Table 7: Category & Tolerance Metrics: Category accuracy (Cat-Acc) measures exact BMI category match. Within-1-Category (±1-Cat) measures predictions within one category of the true category. ±2 BMI and ±3 BMI measure predictions within 2 or 3 BMI points of the true value. The Stacking Ensemble achieves the highest within-1-category accuracy (96.3%).

Risk Probability Estimation Our Stacking Ensemble achieves a Brier score of 0.0983 for risk probability estimation, indicating excellent calibration of predicted obesity risk probabilities. Random Forest and Gradient Boosting achieve comparable Brier scores of 0.1041. Table 8 presents Brier scores for all seven models.

Model	Brier Score
Stacking Ensemble	0.0983
Random Forest	0.1041
Gradient Boosting	0.1041
Decision Tree	0.1159
Voting Ensemble	0.1176
Ridge Regression	0.2203
Lasso Regression	0.2203

Table 8: Risk Probability Estimation: Brier score measures the calibration quality of predicted probabilities (lower is better). Risk probabilities are derived from predicted BMI using a piecewise interpolation function. Stacking Ensemble achieves the best calibration.

Summary Across all evaluation dimensions, the Stacking Ensemble consistently achieves the best or near-best performance, with Gradient Boosting as a close second. These dimensions include regression-based

BMI prediction (R^2: 0.8858), binary obesity classification (AUC-ROC: 0.9588), category accuracy (65.1%), and risk probability estimation (Brier: 0.0983). Figure 6 shows a radar chart comparing all seven models across five classification metrics.

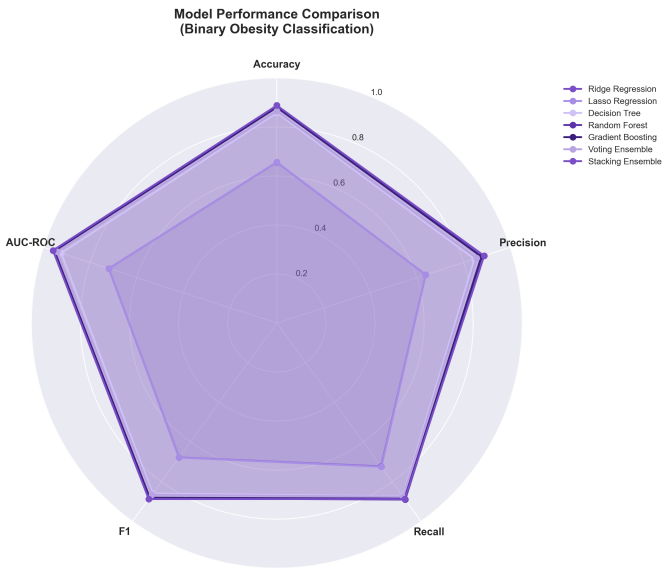

Figure 6: Model performance radar chart comparing all seven models across five classification metrics (Accuracy, Precision, Recall, F1, AUC-ROC). Stacking Ensemble and Random Forest achieve the highest scores across most dimensions.

4.4 Task 4: Fairness Analysis

The task is to evaluate model performance across demographic groups to assess equity and identify potential biases. The input consists of model predictions from Task 2 and demographic features (_SEX, _RACE, _AGEG5YR, _INCOMG1), and the output is fairness metrics across demographic subgroups.

Systematic subgroup analysis is implemented across demographic dimensions (gender, age, race/ethnicity, income). For each demographic group, performance metrics (MAE, Accuracy, AUC-ROC) are computed separately. Disparities are measured as the range (maximum minus minimum) across groups within each demographic dimension. Permutation feature importance analysis is employed by shuffling each variable and measuring its impact on model accuracies. This approach enables us to determine which demographic features are best to include and assess discrimination risk. By comparing the importance of race/ethnicity features with income and geographic features, we investigate whether observed patterns are more likely due to cultural practices specific to racial and ethnic groups or socioeconomic and regional environmental factors, providing insights into the sources of potential biases.

Subgroup analysis reveals disparities across demographic dimensions, as shown in Table 9. Race/ethnicity shows the highest MAE disparity (0.86 BMI points), with Hispanic individuals experiencing the lowest prediction error and White individuals the highest. Income disparity is also substantial (0.79 BMI points), with low-income individuals experiencing lower prediction error than high-income individuals. Gender disparity is moderate (0.42 BMI points), while age disparity is minimal (0.12 BMI points).

Demographic	MAE Δ	Acc Δ	AUC Δ	Best Group	Worst Group
Race/Ethnicity	0.86	0.0407	0.0282	Hispanic	White
Income	0.79	0.0566	0.0347	Low (<$25k)	High ($75k+)
Sex/Gender	0.42	0.0129	0.0074	Male	Female
Age	0.12	0.0129	0.0173	Young Adult	Senior

Table 9: Demographic Disparity Analysis: MAE Δ is the difference in Mean Absolute Error (BMI points) between best and worst performing demographic groups. Acc Δ and AUC Δ are accuracy and AUC-ROC disparities. Race/ethnicity and income show the largest disparities.

Permutation importance analysis reveals that income importance (permutation score: 0.2543) exceeds race importance (0.1433), suggesting that socioeconomic factors may be more influential than demographic factors in prediction quality.

5 Conclusion

5.1 Contributions to Problem Statement

This research addresses critical gaps in obesity prediction by developing and evaluating machine learning models for BMI prediction using survey-based data from the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset. Through comprehensive feature selection, we identified 16 optimal predictors from 302 available variables. Our evaluation of seven machine learning models demonstrates that ensemble methods, particularly the Stacking Ensemble, achieve substantially better performance than baseline linear models across all evaluation dimensions.

Regression-Based BMI Prediction The Stacking Ensemble achieves clinically acceptable accuracy with RMSE of 3.32 BMI points, MAE of 2.26 BMI points, and R^2 of 0.8858, explaining 88.58% of BMI variance. Gradient Boosting performs comparably with R^2 of 0.8699. This performance supports the feasibility of survey-based prediction for population-level obesity screening.

Binary Obesity Classification For binary obesity classification (BMI $\geq$ 30), the Stacking Ensemble achieves Accuracy of 0.8879, Precision of 0.8881, Recall of 0.8875, F1-Score of 0.8878, and AUC-ROC of 0.9588, demonstrating excellent discrimination between obese and non-obese individuals.

Risk Probability Estimation The Stacking Ensemble achieves a Brier score of 0.0983, indicating excellent calibration of predicted obesity risk probabilities and enabling reliable risk stratification for clinical decision-making.

Category & Tolerance Metrics The Stacking Ensemble achieves category accuracy of 0.6505 (65.1%) and within-1-category accuracy of 0.9631 (96.3%), demonstrating that the vast majority of predictions fall within clinically acceptable ranges. Tolerance-based metrics show ±2 BMI accuracy of 0.6122 and ±3 BMI accuracy of 0.7428, further supporting clinical utility.

Fairness Analysis Subgroup analysis reveals race/ethnicity as the primary source of disparity with MAE difference of 0.86 BMI points between best (Hispanic) and worst (White) performing groups. Income disparity is also substantial at 0.79 BMI points, with low-income individuals experiencing lower prediction error than high-income individuals. Gender disparities are moderate (Δ MAE = 0.42), while age disparities are minimal (Δ MAE = 0.12). Permutation feature importance analysis reveals that income importance (0.2543) exceeds race importance (0.1433), suggesting that socioeconomic factors may be more influential than demographic factors in prediction quality.

5.2 Most Important Contributions

This work makes several novel contributions. First, we present a novel application of ensemble methods to BRFSS-based BMI prediction, demonstrating the feasibility of survey-based obesity prediction at pop-

ulation scales. Second, we conduct systematic subgroup analysis across multiple demographic dimensions, identifying specific disparities and investigating their sources. Third, we develop a multi-metric evaluation framework combining regression, classification, and fairness metrics. Fourth, we employ permutation feature importance analysis to identify critical predictors and assess discrimination risks. Fifth, we investigate ethnic versus regional cultural factors affecting obesity, providing insights into whether biases stem from cultural practices specific to racial and ethnic groups or geographic and environmental factors.

5.3 Limitations, Assumptions, and Caveats

Key limitations include the following. First, self-reported data may introduce measurement error, though research suggests such errors are generally acceptable for population-level analyses Pierannunzi et al. [2013]. Second, ensemble methods limit interpretability compared to simple linear models. Third, small sample sizes exist for some demographic groups, potentially affecting fairness evaluation accuracy. Additionally, we do not report confidence intervals or statistical significance tests for our metrics, which limits our ability to quantify uncertainty in performance estimates. Fourth, we have no clinical validation against measured BMI. Fifth, we have no comparison to existing methods. Sixth, there are temporal and geographic limitations, with data from US only and from 2024 only. Seventh, R^2 values are not directly comparable across different test distributions: the balanced 100K test set has higher BMI variance than the imbalanced Original test set, which contributes to the higher R^2 value. However, the substantial improvements in absolute error metrics (MAE: 4.48 to 2.26, RMSE: 6.06 to 3.32) confirm real performance gains beyond variance effects. Eighth, the balanced 100K dataset was created via oversampling before train-test splitting; specifically, only the Underweight category was oversampled from approximately 7,377 samples to 16,667 samples, while other categories maintained their original sizes or were undersampled. Despite optimal demographic representation, biases persist (race/ethnicity disparity: Δ Accuracy = 0.0407) due to feature limitations in capturing cultural determinants.

5.4 Possible Extensions

Future work should address several areas. First, hyperparameter optimization and additional ensemble methods such as XGBoost and LightGBM should be explored. Second, multi-year data integration and external validation should be conducted. Third, cross-validation and bootstrap confidence intervals should be implemented. Fourth, bias mitigation techniques and intersectional fairness analysis should be developed. Fifth, clinical validation studies should compare survey-based predictions to clinically measured BMI. Sixth, uncertainty quantification and interpretability improvements should use SHAP values. Seventh, expansion to other health outcomes and longitudinal analysis should be pursued.

References

Nandita Bhan, Ichiro Kawachi, Maria M. Glymour, and S. V. Subramanian. Time trends in racial and ethnic disparities in asthma prevalence in the united states from the behavioral risk factor surveillance system (brfss) study (1999–2011). *American Journal of Public Health*, 105:1269–1275, 2015. doi: 10.2105/AJPH.2014.302172. URL `https://doi.org/10.2105/AJPH.2014.302172`.

David W Brown, Lina S Balluz, Wayne H Giles, Gloria L Beckles, David G Moriarty, Earl S Ford, and Ali H Mokdad. Diabetes mellitus and health-related quality of life among older adults: Findings from the behavioral risk factor surveillance system (brfss). *Diabetes Research and Clinical Practice*, 65(2):105–115, 2004. doi: 10.1016/j.diabres.2003.11.014. URL `https://doi.org/10.1016/j.diabres.2003.11.014`.

Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, and Hafiz Husnain Raza Sherazi. Machine learning based diabetes classification and prediction for healthcare applications. *Journal of Healthcare Engineering*, 2021:9930985, 2021. doi: 10.1155/2021/9930985.

Centers for Disease Control and Prevention. Behavioral risk factor surveillance system survey data and

documentation. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2024.

Ronald J. Chen, Jingxuan J. Wang, Drew F. K. Williamson, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7:719–742, 2023. doi: 10.1038/s41551-023-01056-8. URL https://doi.org/10.1038/s41551-023-01056-8.

Hikaru Ooba, Jota Maki, Takahiro Tabuchi, and Hisashi Masuyama. Partner relationships, hopelessness, and health status strongly predict maternal well-being: an approach using light gradient boosting machine. *Scientific Reports*, 13:17032, 2023. doi: 10.1038/s41598-023-44410-1. URL https://doi.org/10.1038/s41598-023-44410-1.

Catherine Pierannunzi, Shangwei S. Hu, and Lina Balluz. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (brfss), 2004–2011. *BMC Medical Research Methodology*, 13:49, 2013. doi: 10.1186/1471-2288-13-49. URL https://doi.org/10.1186/1471-2288-13-49.

Mahmood Safaei, Elankovan A. Sundararajan, Maha Driss, Wadii Boulila, and Azrulhizam Shapi'i. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, 136:104754, 2021a. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.104754. URL https://www.sciencedirect.com/science/article/pii/S0010482521005485.

Mahmood Safaei, Elankovan A. Sundararajan, Maha Driss, Wadii Boulila, and Azrulhizam Shapi'i. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, 136:104754, 2021b. doi: 10.1016/j.compbiomed.2021.104754. Systematic review of 93 papers on ML for obesity prediction from 2010-2020.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.

Gustavo Yamada, Carlos Castillo-Salgado, Jessica C. Jones-Smith, and Lawrence H. Moulton. Obesity prediction by modelling bmi distributions: application to national survey data from mexico, colombia and peru, 1988–2014. *International Journal of Epidemiology*, 49(3):824–833, 2020. doi: 10.1093/ije/dyz195.