World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

MACHINE LEARNING-BASED PREDICTION OF POOR SELF-RATED HEALTH AMONG U.S. ADULTS USING BEHAVIORAL AND SOCIOECONOMIC FACTORS

Mahfuz Islam Khan Jabed [1,*], Muhammad Imran [1], Clapher Ankur Gomes [1] and Prudvi Saisaran Ponduru [2]

[1] Department of Information Technology, Washington University of Science and Technology, Virginia, USA.
[2] Department of Computer Science, University of the Potomac, Washington, D.C, USA.
* Corresponding Author

ORCID Details
Mahfuz Islam Khan Jabed: https://orcid.org/0009-0001-2141-2894
Muhammad Imran: https://orcid.org/0009-0001-2069-0868
Clapher Ankur Gomes: https://orcid.org/0009-0004-9566-6784
Prudvi Saisaran Ponduru: https://orcid.org/0009-0009-5930-8265

World Journal of Advanced Research and Reviews, 2025, 27(01), 2817–2829

Publication history: Received on 07 June 2025; revised on 23 July 2025; accepted on 28 July 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.1.2649

ABSTRACT

Fair or poor self-rated health (SRH) is a concise population-health indicator, and scalable classification may support public-health surveillance. However, temporal robustness, calibration, complex survey design, and subgroup performance require careful evaluation. This study developed and temporally validated machine-learning models for identifying U.S. adults reporting fair or poor SRH using behavioral and socioeconomic factors. Official 2021 Behavioral Risk Factor Surveillance System data formed the development cohort, while 2022 data were reserved for temporal testing. Seventeen harmonized predictors were evaluated using logistic regression, decision tree, random forest, extra-trees, and histogram gradient-boosting models. State-grouped three-fold cross-validation guided model configuration and threshold selection. Survey weights informed model fitting and evaluation, and 200 stratified primary-sampling-unit bootstrap replicates produced 95% confidence intervals. The analysis also examined calibration, subgroup and state-level performance, predictor domains, parsimonious modeling, and complete cases. The analytical cohorts included 430,494 adults in 2021 and 434,659 in 2022, with weighted fair or poor SRH prevalence of 16.13% and 17.89%, respectively. Histogram gradient boosting performed best in the 2022 temporal test, achieving a weighted ROC-AUC of 0.801 (95% CI, 0.798–0.805), PR-AUC of 0.504 (0.497–0.513), balanced accuracy of 0.727 (0.724–0.731), and Brier score of 0.1167 (0.1154–0.1180). Its improvement over logistic regression was modest. Employment, physical activity, healthcare cost barriers, body mass index, income, and age were influential, with temporally stable importance rankings (Spearman ρ=0.978). Although behavioral and structural factors supported useful temporal classification, performance varied across population and geographic groups. Residual miscalibration, subgroup threshold disparities, self-reported data, and cross-sectional measurement preclude immediate clinical or individual-level deployment.

Keywords: Self-Rated Health, Machine Learning, Health Informatics, Healthcare Analytics, Predictive Analytics, Population Health Management, Health Equity

1 INTRODUCTION

Self-rated health (SRH) condenses perceived physical, functional, psychological, and social circumstances into a single response. Although subjective, it consistently predicts mortality across communities [[7]], improves mortality risk stratification in meta-analysis [8], and is understood as an integrative appraisal rather than a direct biological measurement [[9]]. SRH also predicts healthcare expenditure [[11]]. In U.S. adults, its association with mortality persists after adjustment for sociodemographic characteristics [[11]], but response reliability and meaning vary across people and contexts [[12]], [[13]]. Fair or poor SRH is therefore useful for population surveillance when interpreted as a multidimensional reported-health indicator—not as a diagnosis.

Behavior and social position jointly shape this appraisal. Smoking, physical inactivity, alcohol use, and body weight contribute to self-ratings [[14]], while income, education, employment, housing, and access to care represent material conditions and opportunities rather than merely individual choices [[16]]. U.S. income gradients in life expectancy illustrate the national scale of socioeconomic health inequality [[17]]. State context also contributes to SRH beyond individual characteristics [[15]]. Healthy People 2030 consequently emphasizes equitable well-being, upstream determinants, national measurement, and scalable action [[18]]. A nationwide surveillance model should therefore assess modifiable behaviors and structural conditions together and must examine whether its performance changes across demographic and geographic groups. Group differences require uncertainty-aware interpretation; numerical gaps alone do not establish unlawful discrimination or algorithmic bias [[31]].

Machine learning can capture nonlinearities and interactions across these domains, but apparent gains may reflect leakage, weak validation, ignored sampling weights, or absent calibration. This study asked six questions: how accurately behavioral and socioeconomic factors classify fair/poor SRH; whether ensembles outperform an interpretable weighted-logistic baseline; which factors contribute most; how performance varies by demographic, socioeconomic, and geographic group; how much combined domains improve on either alone; and whether a limited predictor set retains useful discrimination and calibration. We developed models exclusively in 2021 BRFSS and evaluated them once in the official 2022 release, preserving a genuine temporal test. The intended contribution is a reproducible, survey-aware, equity-audited framework for population-health surveillance and planning—not individual diagnosis or prediction of a future health event.

2 RELATED WORKS

Clark et al. applied regularized logistic regression, random forest, and support-vector machines to 2017 BRFSS and replicated analyses in 2016, reporting ROC-AUCs near 0.80–0.81 and emphasizing life-course and equity differences [[19]]. That study is the closest pre-2025 antecedent and established that socioeconomic factors are prominent predictors. Our design differs by restricting predictors to behavioral, demographic, socioeconomic, and justified access factors; excluding chronic disease, disability, mental-health-day, and alternative SRH encodings; locking all decisions in 2021; applying final survey weights to fitting and evaluation; reporting calibration and PR-AUC; and evaluating a single model across multiple subgroups and states in 2022.

Other work has examined different populations or predictor scopes. Gumà-Lao and Arpino used classification trees to study how health conditions define SRH across education groups among older Europeans [[20]]. Hoekstra et al. used random forests and external-exposome data to predict poor self-perceived general health in Dutch population surveys [[21]]. Chen et al. combined XGBoost and SHAP to study built and natural environmental contributors among older Chinese adults [[22]]. These studies support nonlinear, subgroup-aware interpretation but do not provide survey-weighted U.S. temporal validation using recent BRFSS core factors. More broadly, machine learning does not reliably outperform logistic regression when validation is rigorous [[29]], reinforcing the need to compare discrimination, calibration, stability, and complexity rather than announce a winner from accuracy alone.

3 MATERIAL AND METHODS

3.1 Study design and data sources

This prediction-model study used repeated cross-sectional, publicly available, deidentified BRFSS data. Official CDC SAS transport archives, annual pages, overviews, questionnaires, calculated-variable documentation, codebooks, complex-weight guidance, weighting formulae, and data-quality reports were retrieved for 2021 and 2022 [[1]]–[[3]]. BRFSS is a state-based landline and cellular telephone survey; its post-2011 design incorporates cell-phone respondents and raking [[6]]. Published validation supports many BRFSS measures while noting limitations of self-report [[4]]. Because

stacking state weights is not identical to constructing a separately calibrated national sample, estimates are interpreted as national aggregates of participating state and District of Columbia surveys; final weights and geography remain explicit [[5]]. Reporting followed applicable TRIPOD+AI principles [[24]], and design choices addressed PROBAST participant, predictor, outcome, and analysis domains [[25]].

3.2 Study population and outcome

The 2021 file was the development cohort and the 2022 file was the untouched temporal test. To maintain a consistent state/DC target population, Guam, Puerto Rico, the U.S. Virgin Islands, and nonstate codes were excluded. GENHLTH responses 4 (fair) or 5 (poor) formed the positive class; 1 (excellent), 2 (very good), or 3 (good) formed the negative class. Refused, "do not know," missing, and invalid responses were excluded. The derived outcome was cross-checked against CDC's _RFHLTH; the latter was never used as a predictor.

3.3 Predictor selection and harmonization

Seventeen core predictors were verified in both codebooks. Twelve socioeconomic/demographic/access variables were age group, sex, imputed race/ethnicity, education, household income, employment, marital status, home ownership, Census region, health coverage, personal-doctor status, and inability to see a doctor because of cost. Five behavioral variables were smoking status, any leisure-time physical activity in the past 30 days, BMI category, alcohol-use days per month, and binge drinking. The access variables were retained because they are actionable planning indicators and plausibly shape reported health. ALCDAY5 (2021) and ALCDAY4 (2022) were harmonized from official weekly/monthly codes to days per 30-day month. Sleep, fruit/vegetable intake, and other rotating or inconsistently administered items were excluded rather than forcing uncertain harmonization.

Variables that reproduce or closely encode health status were excluded: _RFHLTH, physical-, mental-, and poor-health days; diagnosed chronic diseases; and disability indicators. Predictors were organized a priori as potentially modifiable behaviors, structural socioeconomic conditions, demographic controls, and healthcare access. Feature importance was not interpreted causally.

3.4 Complex survey design and descriptive analysis

Original _LLCPWT, _STSTR, and _PSU values were used for weighted prevalence, characteristics, test metrics, and uncertainty. Fitting weights were divided by their mean to preserve relative survey influence without changing the effective loss scale. Unweighted sample counts were always reported separately. BRFSS raking adjusts for selection and demographic margins, but cannot remove all nonresponse, under coverage, mode, recall, or reporting error; the 2022 median combined response rate was approximately 45% [[3]]. Survey-weighted prediction metrics were computed because metrics from complex survey samples can differ from naïve sample metrics [[30]].

3.5 Preprocessing and missing data

The calendar-year split preceded imputation, encoding, scaling, selection, or tuning. Within every training fold, categorical missingness was assigned an explicit "Missing/unknown" level; numeric alcohol frequency used training-median imputation plus a missingness indicator and training-fitted scaling. One-hot encoders ignored unseen test categories. All transformations and models were joined in scikit-learn pipelines [[36]]. Missingness was tabulated overall and by outcome, sex, and race/ethnicity. Income was the most incomplete predictor (21.57% in 2022 overall), followed by BMI (11.04%), binge drinking (11.55%), alcohol frequency, and smoking. A complete-case sensitivity analysis used the same locked algorithm class but represented a different observed-complete population. No SMOTE or synthetic data were used because oversampling would alter the relationship between the survey sample and target population. Survey weights, not outcome-dependent resampling, defined the fitting objective. The operating threshold was selected within 2021 by the weighted Youden criterion; it was not re-estimated in 2022.

3.6 Model development and internal validation

We evaluated ridge logistic regression, decision tree, random forest [[33]], histogram gradient boosting [[35]], and extremely randomized trees [[34]]. These span interpretable linear, single-tree, bagged/randomized, and boosted nonlinear learners. Candidate grids covered regularization, depth, leaf size, learning rate, iterations, and L2 penalty. A fixed seed (20240816) controlled splitting and estimators.

Three-fold Stratified Group KFold cross-validation in 2021 kept every state wholly within a fold while approximately preserving outcome balance. Each candidate was trained with normalized survey weights. Mean weighted ROC-AUC was primary, with mean Brier score as a tie-breaker; PR-AUC was also retained. Weighted out-of-fold predictions from the selected configuration determined its threshold. After every model and threshold were locked, the pipeline was refit on all 2021 data and applied once to 2022. This was temporal validation; grouped cross-validation was internal geographic validation. General machine-learning-in-epidemiology guidance informed the workflow [[23]].

3.7 Evaluation and uncertainty

On the 2022 temporal test we calculated survey-weighted and unweighted ROC-AUC, average precision (PR-AUC), balanced accuracy, sensitivity, specificity, positive predictive value (precision), negative predictive value, F1, Brier score, and confusion matrices. PR-AUC was emphasized because fair/poor SRH was the minority class [[28]]. Calibration intercept and slope came from weighted logistic recalibration of the observed outcome on the prediction logit; ideal values are 0 and 1, respectively [[26]], [[27]]. A positive intercept indicates average underprediction conditional on slope.

Ninety-five percent percentile intervals used 200 stratified PSU perturbation-bootstrap replicates (seed 20240817). Independent exponential PSU multipliers were normalized within each _STSTR before multiplying _LLCPWT, maintaining stratum weight totals. One hundred analogous replicates supported subgroup ROC-AUC and PR-AUC intervals. These replication counts balance computational burden and uncertainty estimation in more than 430,000 observations.

3.8 Domain, parsimonious, interpretability, and subgroup analyses

Using the best algorithm class and 2021-only selection, separate models included socioeconomic, demographic, access factors, behavioral factors, or all predictors. A parsimonious model used the eight most important raw predictors from a 2021 state holdout, avoiding selection on 2022. For the logistic model, all penalized coefficients were exported under its full one-hot parameterization; exponentiated coefficients are predictive parameter contrasts, not conventional unpenalized reference-category causal odds ratios. For the best nonlinear model, raw-feature permutation importance measured the decrease in weighted ROC-AUC after shuffling each variable in fixed 30,000-person samples with five repeats. Rank stability compared the 2021 state holdout with 2022.

The locked best model and one population threshold were evaluated without retuning across age (18–34, 35–49, 50–64, ≥65 years), sex, imputed race/ethnicity, income, education, Census region, and state. We reported prevalence, discrimination, sensitivity, specificity, error rates, Brier score, and calibration. Subgroup gaps were treated as descriptive evidence with sampling uncertainty, not automatic proof of bias.

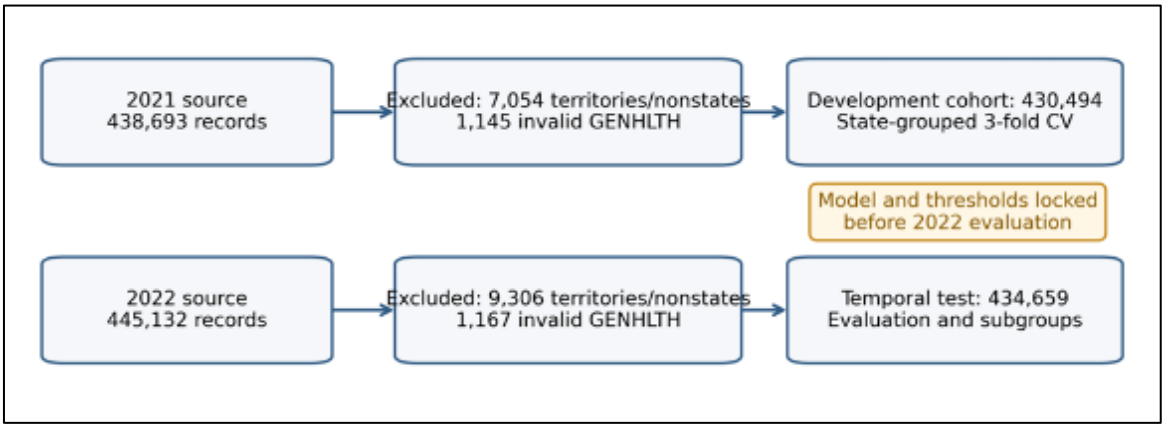

Figure 1: Study flow and temporal-validation design. All preprocessing, hyperparameters, feature decisions, and operating thresholds were determined from 2021 before the 2022 temporal test.

4 RESULTS

4.1 Cohorts, outcome, and participant characteristics

The official 2021 file contained 438,693 records. We excluded 7,054 territory/nonstate records and 1,145 state/DC records with invalid or missing GENHLTH; no records were removed for invalid final weight, leaving 430,494 adults. Fair/poor SRH occurred in 71,103 (16.52% unweighted), corresponding to 16.13% weighted prevalence. The 2022 file contained 445,132 records; exclusions were 9,306 territory/nonstate records and 1,167 invalid/missing outcomes, leaving 434,659. Fair/poor SRH occurred in 77,507 (17.83% unweighted), corresponding to 17.89% weighted prevalence. _RFHLTH cross-checks produced zero discrepancies in either year. The annual weight sums were 242.49 million and 260.88 million population units, respectively; these are weighted totals, not additional respondents.

In 2022, fair/poor SRH represented larger weighted shares among adults with less than high-school education (24.31% of positive cases versus 8.71% of negative cases), household income below $25,000 (26.90% versus 9.42%), inability to work (21.45% versus 2.94%), physical inactivity (44.03% versus 19.13%), obesity (40.42% versus 26.45%), and

current daily smoking (14.14% versus 6.79%). These are descriptive compositions, not adjusted or causal effects (Table 1).

Table 1: Unweighted and survey-weighted characteristics of the 2022 temporal-test cohort (N=434,659)

Characteristic	Category	Unweighted n (%)	Overall weighted %	Within fair/poor weighted %
Age	18-34	72,843 (16.8)	29.0	19.5
	35-49	84,853 (19.5)	22.6	19.5
	50-64	112,066 (25.8)	23.7	28.2
	65+	156,051 (35.9)	22.6	31.0
	Missing/unknown	8,846 (2.0)	2.1	1.8
Sex	Female	229,974 (52.9)	51.3	53.5
	Male	204,685 (47.1)	48.7	46.5
Race/ethnicity	White, non-Hispanic	332,190 (76.4)	59.9	53.8
	Black, non-Hispanic	34,660 (8.0)	11.7	13.8
	Hispanic	37,070 (8.5)	17.3	22.6
	Asian, non-Hispanic	12,750 (2.9)	6.2	3.4
	AIAN, non-Hispanic	7,081 (1.6)	1.2	2.0
	Other/multiracial, non-Hispanic	10,908 (2.5)	3.6	4.4
Education	Less than high school	24,789 (5.7)	11.5	24.3
	High school graduate	106,057 (24.4)	27.4	30.7
	Some college	117,683 (27.1)	30.1	29.0
	College graduate	183,864 (42.3)	30.4	15.5
	Missing/unknown	2,266 (0.5)	0.6	0.5
Income	<$25,000	52,408 (12.1)	12.5	26.9
	$25,000-$49,999	86,820 (20.0)	19.4	23.6
	$50,000-$99,999	106,165 (24.4)	22.2	15.6
	>= $100,000	95,524 (22.0)	22.1	8.5
	Missing/unknown	93,742 (21.6)	23.8	25.3
Employment	Employed	219,821 (50.6)	55.9	35.2
	Unemployed	16,098 (3.7)	5.0	7.3
	Homemaker/student	27,276 (6.3)	9.4	7.1
	Retired	134,442 (30.9)	20.1	25.9
	Unable to work	26,058 (6.0)	6.3	21.4
	Missing/unknown	10,964 (2.5)	3.4	3.1
Physical activity	Active in past 30 days	330,961 (76.1)	76.1	55.5
	Inactive in past 30 days	102,680 (23.6)	23.6	44.0
	Missing/unknown	1,018 (0.2)	0.3	0.5
Smoking	Never	238,877 (55.0)	56.9	45.3

	Former	111,974 (25.8)	22.0	26.5
	Current some days	13,679 (3.1)	3.5	5.1
	Current daily	35,232 (8.1)	8.1	14.1
	Missing/unknown	34,897 (8.0)	9.5	8.9
BMI	Underweight	6,558 (1.5)	1.8	2.4
	Normal	114,353 (26.3)	26.5	19.7
	Overweight	136,448 (31.4)	29.6	24.3
	Obesity	129,326 (29.8)	28.9	40.4
	Missing/unknown	47,974 (11.0)	13.1	13.1

Values are unweighted n (column %) and survey-weighted column percentages. The final column is the weighted composition among fair/poor-SRH respondents, it is not category-specific prevalence. Missing categories are retained for descriptive transparency.

4.2 Temporal model performance

Histogram gradient boosting (HGB) ranked first in 2021 grouped cross-validation (mean weighted ROC-AUC 0.808) and in the untouched 2022 test (Table 2). Its weighted ROC-AUC was 0.801 (95% CI 0.798–0.805), PR-AUC 0.504 (0.497–0.513), balanced accuracy 0.727 (0.724–0.731), sensitivity 0.717 (0.710–0.724), specificity 0.736 (0.733–0.739), precision 0.372 (0.367–0.379), NPV 0.923 (0.921–0.925), F1 0.490 (0.485–0.496), and Brier score 0.1167 (0.1154–0.1180). Calibration intercept was 0.132 (0.102–0.165) and slope 0.999 (0.984–1.017), indicating an approximately correct spread but modest average underprediction in 2022.

The ensemble advantage was small. Weighted logistic regression achieved ROC-AUC 0.793 (0.789–0.797), PR-AUC 0.497 (0.489–0.506), balanced accuracy 0.721 (0.717–0.724), and Brier score 0.1179 (0.1166–0.1193). Random forest, extra-trees, and decision tree achieved ROC-AUCs of 0.800, 0.798, and 0.779. At the locked HGB threshold 0.15356, the unweighted confusion matrix contained 262,095 true negatives, 95,057 false positives, 21,358 false negatives, and 56,149 true positives (Fig. 2). The analogous survey-weighted totals were 157.75, 56.46, 13.20, and 33.47 million population units.

Point estimates with 95% percentile confidence intervals from 200 stratified PSU-weight perturbation bootstraps. BA, balanced accuracy; PPV, positive predictive value; NPV, negative predictive value. Thresholds were locked from 2021 out-of-fold predictions.

Table 2: Survey-weighted performance on the untouched 2022 temporal test.

- Panel A. Discrimination and classification

Model	ROC-AUC	PR-AUC	BA	Sensitivity	Specificity	PPV	NPV	F1
HGB	0.801 (0.798–0.805)	0.504 (0.497–0.513)	0.727 (0.724–0.731)	0.717 (0.710–0.724)	0.736 (0.733–0.739)	0.372 (0.367–0.379)	0.923 (0.921–0.925)	0.490 (0.485–0.496)
Random forest	0.800 (0.796–0.803)	0.501 (0.494–0.510)	0.726 (0.723–0.730)	0.721 (0.715–0.727)	0.730 (0.727–0.734)	0.368 (0.364–0.374)	0.923 (0.921–0.925)	0.487 (0.483–0.493)
Extra-trees	0.798 (0.794–0.802)	0.500 (0.492–0.508)	0.723 (0.720–0.727)	0.707 (0.701–0.713)	0.740 (0.737–0.743)	0.372 (0.367–0.377)	0.920 (0.919–0.922)	0.487 (0.482–0.493)
Logistic regression	0.793 (0.789–0.797)	0.497 (0.489–0.506)	0.721 (0.717–0.724)	0.712 (0.706–0.719)	0.730 (0.726–0.733)	0.364 (0.360–0.371)	0.921 (0.919–0.923)	0.482 (0.477–0.489)
Decision tree	0.779 (0.775–0.783)	0.467 (0.459–0.475)	0.710 (0.706–0.713)	0.706 (0.698–0.712)	0.714 (0.711–0.717)	0.350 (0.345–0.355)	0.918 (0.915–0.919)	0.468 (0.462–0.473)

- Panel B. Overall probability accuracy and calibration

Model	Brier	Calibration intercept	Calibration slope	Threshold
HGB	0.1167 (0.1154–0.1180)	0.132 (0.102–0.165)	0.999 (0.984–1.017)	0.15356
Random forest	0.1172 (0.1160–0.1184)	0.209 (0.177–0.244)	1.075 (1.058–1.094)	0.16885
Extra-trees	0.1176 (0.1163–0.1188)	0.263 (0.231–0.299)	1.104 (1.087–1.122)	0.16657
Logistic regression	0.1179 (0.1166–0.1193)	0.078 (0.049–0.114)	0.964 (0.949–0.981)	0.15199
Decision tree	0.1213 (0.1199–0.1227)	-0.043 (-0.075–-0.007)	0.870 (0.856–0.888)	0.14054

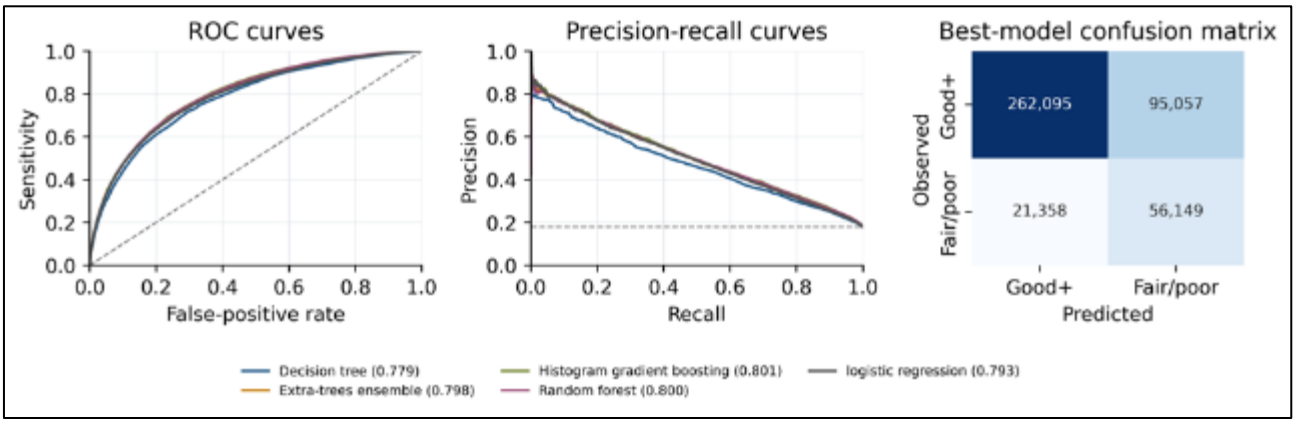

Figure 2: Survey-weighted ROC and precision-recall curves for the 2022 temporal test, plus the unweighted confusion matrix for the best model at the locked threshold. Legend values are weighted ROC-AUCs; the dashed PR baseline is weighted prevalence.

4.3 Domain contribution, parsimony, and sensitivity

The socioeconomic/demographic/access-only HGB achieved ROC-AUC 0.773, PR-AUC 0.474, and Brier 0.1208; the behavioral-only model achieved 0.707, 0.343, and 0.1345. Combining all 17 predictors improved ROC-AUC by 0.028 and PR-AUC by 0.030 over the structural/access domain alone, and by 0.094 and 0.161 over behavior alone. The eight-predictor model—employment, physical activity, cost barrier, education, BMI, age, income, and personal-doctor status—retained ROC-AUC 0.792, PR-AUC 0.494, balanced accuracy 0.720, and Brier 0.1180, a 0.009 ROC-AUC loss from the full model.

Complete-case data included 290,526 development participants (67.49%) and 286,048 test participants (65.81%). The complete-case HGB achieved ROC-AUC 0.807, PR-AUC 0.507, balanced accuracy 0.728, Brier 0.1121, calibration intercept 0.126, and slope 0.997. Slightly better performance does not establish superior missing-data handling because the complete-case target population differed.

4.4 Interpretability and temporal stability

Permutation importance ranked employment first (weighted ROC-AUC decrease 0.0442 in 2022), followed by physical activity (0.0277), cost barrier (0.0168), BMI category (0.0162), income (0.0134), age group (0.0132), education (0.0109), and personal-doctor status (0.0097) (Fig. 3). Rankings were highly stable between the 2021 state holdout and 2022 (Spearman ρ=0.978, $p<0.001$). Binge drinking and health coverage had near-zero conditional importance given the other variables; this does not imply that they are unimportant to health.

The ridge-logistic model provided concordant directional signals within its penalized full-dummy parameterization. Large positive coefficients included unable to work (β=1.066), cost barrier present (β=0.427), income <$15,000

(β=0.434), less than high-school education (β=0.395), inactivity (β=0.321), obesity (β=0.265), and current daily smoking (β=0.212). Employed status (β=−0.534), active status (β=−0.442), college graduation (β=−0.376), and income ≥$200,000 (β=−0.671) were negative. Because all category indicators were retained under ridge regularization, these coefficients must not be read as traditional reference-group odds ratios or intervention effects.

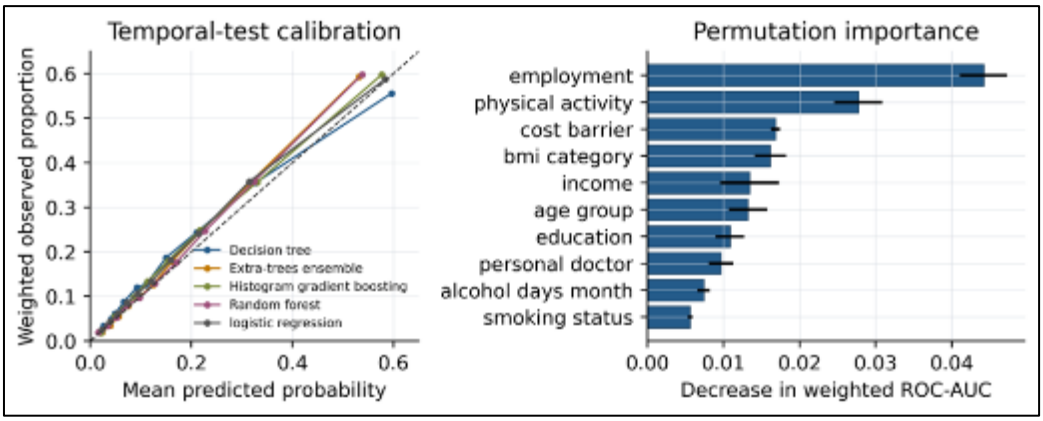

Figure 3: Weighted decile calibration and 2022 permutation importance. Importance is the mean decrease in weighted ROC-AUC; error bars show the standard deviation across five permutations in a fixed 30,000-person sample.

4.5 Subgroup and geographic robustness

Table 3 shows heterogeneous operating characteristics at the single fixed threshold. Age-group ROC-AUC ranged from 0.737 (18–34 years; 95% CI 0.727–0.747) to 0.839 (50–64 years; 0.834–0.843). Sensitivity was 0.447 among adults 18–34 (false-negative rate 0.553) versus 0.832 among adults ≥65, whereas specificity changed in the opposite direction (0.843 versus 0.554). Female and male ROC-AUCs were 0.806 and 0.795. Across race/ethnicity, ROC-AUC ranged from 0.740 among non-Hispanic Asian adults (0.716–0.765) to 0.810 among non-Hispanic White adults (0.807–0.813); the non-Hispanic American Indian/Alaska Native interval was wider (0.732–0.789), reflecting smaller effective sample size.

Income-group ROC-AUCs were similar (0.746–0.762), yet threshold behavior differed because outcome prevalence and score distributions differed: sensitivity was 0.924 below $25,000 and 0.293 at ≥$100,000, with specificity 0.330 and 0.948, respectively. Education showed the same prevalence-linked tradeoff. Regional ROC-AUC ranged from 0.789 in the West to 0.809 in the South. Across 51 state/DC units, median ROC-AUC was 0.798 (IQR 0.790–0.813), ranging from 0.731 in Hawaii to 0.833 in North Carolina. These state endpoints lack separate multiplicity-adjusted inference and should motivate local validation, not ranking of jurisdictions.

ROC-AUC and PR-AUC show 95% design-aware bootstrap intervals (100 replicates). Other entries are survey-weighted point estimates at the single 2021-derived threshold. Prev, fair/poor-SRH prevalence; Sens, sensitivity; Spec, specificity; Cal I/S, calibration intercept/slope; AIAN, American Indian/Alaska Native.

Table 3: Subgroup performance of the locked HGB model on the 2022 temporal test.

Dimension	Group	n	Prev	ROC-AUC (95% CI)	PR-AUC (95% CI)	Sens	Spec	Brier	Cal I/S
Age	18-34	72,843	0.121	0.737 (0.727–0.747)	0.291 (0.276–0.306)	0.447	0.843	0.097	0.11/0.95
	35-49	84,853	0.154	0.805 (0.797–0.812)	0.463 (0.445–0.480)	0.681	0.773	0.106	0.10/0.99
	50-64	112,066	0.213	0.839 (0.834–0.843)	0.605 (0.594–0.615)	0.807	0.718	0.119	0.16/1.04
	65+	156,051	0.246	0.777 (0.772–0.783)	0.538 (0.525–0.553)	0.832	0.554	0.151	0.15/1.02

Sex	Female	229,974	0.186	0.806 (0.802–0.811)	0.521 (0.511–0.530)	0.746	0.720	0.119	0.14/1.01
	Male	204,685	0.171	0.795 (0.791–0.800)	0.484 (0.473–0.494)	0.684	0.753	0.114	0.13/0.99
Race/ethnicity	AIAN, non-Hisp.	7,081	0.291	0.764 (0.732–0.789)	0.603 (0.568–0.636)	0.766	0.607	0.170	0.31/0.90
	Asian, non-Hisp.	12,750	0.100	0.740 (0.716–0.765)	0.273 (0.227–0.339)	0.438	0.863	0.082	-0.19/0.89
	Black, non-Hisp.	34,660	0.210	0.777 (0.765–0.785)	0.496 (0.477–0.516)	0.754	0.655	0.137	0.04/0.94
	Hispanic	37,070	0.233	0.775 (0.764–0.786)	0.533 (0.511–0.554)	0.779	0.610	0.146	0.13/0.97
	Other/multiracial	10,908	0.217	0.808 (0.793–0.822)	0.566 (0.538–0.596)	0.748	0.730	0.133	0.40/1.05
	White, non-Hisp.	332,190	0.161	0.810 (0.807–0.813)	0.498 (0.490–0.507)	0.695	0.773	0.106	0.14/1.02
Income	$25k–$49,999	86,820	0.218	0.748 (0.740–0.756)	0.476 (0.464–0.494)	0.758	0.589	0.145	0.04/0.94
	$50k–$99,999	106,165	0.126	0.746 (0.735–0.755)	0.343 (0.326–0.361)	0.519	0.820	0.098	0.08/0.98
	<$25,000	52,408	0.384	0.755 (0.745–0.761)	0.649 (0.635–0.661)	0.924	0.330	0.194	0.16/0.93
	≥$100,000	95,524	0.069	0.762 (0.746–0.772)	0.255 (0.234–0.279)	0.293	0.948	0.058	0.34/1.09
Education	College grad.	183,864	0.091	0.785 (0.778–0.791)	0.338 (0.327–0.350)	0.418	0.918	0.071	0.25/1.05
	High school	106,057	0.200	0.767 (0.760–0.772)	0.484 (0.473–0.494)	0.737	0.659	0.133	0.04/0.95
	<High school	24,789	0.378	0.748 (0.737–0.758)	0.639 (0.623–0.656)	0.931	0.285	0.194	0.18/1.01
	Some college	117,683	0.172	0.777 (0.771–0.783)	0.467 (0.456–0.480)	0.677	0.733	0.118	0.13/1.00
Region	Midwest	115,575	0.171	0.802 (0.796–0.809)	0.495 (0.483–0.509)	0.704	0.750	0.113	0.15/1.01

	Northeast	83,278	0.163	0.797 (0.790–0.805)	0.471 (0.455–0.492)	0.686	0.766	0.110	0.09/0.98
	South	124,829	0.192	0.809 (0.804–0.814)	0.537 (0.527–0.548)	0.755	0.709	0.121	0.14/1.02
	West	110,977	0.175	0.789 (0.781–0.799)	0.474 (0.459–0.491)	0.683	0.745	0.118	0.14/0.97

5 DISCUSSION

5.1 Principal findings and comparison with literature

Recent behavioral, socioeconomic, demographic, and access factors classified contemporaneous fair/poor SRH with useful but incomplete discrimination. HGB was the best overall model, but its 0.009 ROC-AUC and 0.0012 Brier improvements over weighted logistic regression were modest. This agrees with systematic evidence that flexible algorithms often do not decisively outperform well-specified regression under rigorous validation [[29]]. It also closely matches the 0.80–0.81 ROC-AUC range reported by Clark et al. [[19]], despite our exclusion of chronic diagnoses, disability, depression, and health-day variables that are more proximal to SRH. The present contribution is therefore not a claim that machine learning newly predicts SRH; it is temporal, survey-weighted, calibration-aware validation with domain ablation, parsimonious modeling, and broad subgroup/geographic auditing.

Employment, physical activity, healthcare affordability, BMI, income, education, and age dominated predictions. This pattern accords with research showing that behavior contributes to SRH [[14]], that social conditions shape health opportunities [[16]], and that the meaning of SRH reflects integrated information [[9]]. Employment was most influential partly because categories such as “unable to work” may encode functional limitation; consequently, importance is predictive, not an estimate of the health benefit of employment. Physical activity and smoking are potentially modifiable, whereas income, education, employment constraints, and cost barriers are structural or policy-sensitive. The combined model’s improvement over either domain indicates complementarity, while the strong eight-predictor model offers a lower-burden surveillance option.

5.2 National public-health and health-equity relevance

BRFSS is used across U.S. jurisdictions, making a harmonized open workflow scalable beyond a single health system. At aggregate level, calibrated risk distributions could supplement prevalence surveillance, guide where richer community assessment is needed, and support planning for prevention, benefits navigation, physical-activity access, or affordability outreach. This aligns with Healthy People 2030’s emphasis on upstream determinants, equitable well-being, and accessible data for targeted action [[18]]. The model should not be used to deny services, set premiums, penalize individuals, or substitute for direct measurement. Population algorithms can reproduce harmful proxy relationships even without explicit intent [[32]]; agencies should document purpose, examine subgroup error and calibration, involve affected communities, and retain human governance [[31]].

Subgroup results show why a single national threshold is not neutral in operation. The lower sensitivity among younger and higher-income adults does not, by itself, establish bias: fair/poor prevalence, score distributions, reporting behavior, and predictor-outcome relationships differ. Likewise, the lower ROC-AUC among Asian adults has uncertainty and may reflect heterogeneous cultural interpretation of SRH, sample composition, or omitted factors. Still, these gaps are decision-relevant. Any implementation should specify the relative harms of false negatives and false positives, validate locally, consider calibrated group-agnostic or context-specific operating rules only with governance review, and monitor intersectional groups. Geographic variation is consistent with prior evidence that state context matters [[15]] and supports state-level recalibration before operational use.

5.3 Pathway for responsible continuation

Next steps are prospective or repeated temporal validation across multiple years; testing in independent national surveys and administrative/community data; assessment of measurement invariance; finer uncertainty for state and intersectional groups; and comparison of missing-data strategies. Agencies should predefine use cases, decision consequences, refresh intervals, performance floors, and rollback criteria. Public code, versioned data provenance, calibration monitoring, and model cards can make reuse auditable. Impact evaluation would then ask whether aggregate targeting improves reach or resource alignment without worsening disparities—an intervention question this classification study cannot answer.

LIMITATIONS

First, BRFSS measures outcome and predictors contemporaneously. “Prediction” denotes statistical classification; the study does not forecast a future health event, establish temporality, or estimate causal effects. Second, SRH, behaviors, BMI inputs, and socioeconomic variables are self-reported and subject to recall, social-desirability, cultural-interpretation, and common-method error [[12]], [[13]]. Third, landline/cell-phone coverage, nonresponse, state sampling, raking, and national aggregation assumptions remain despite weighting [[3]]–[[6]]. Fourth, one-year temporal validation shares the BRFSS infrastructure and is stronger than a random respondent split but is not independent external validation.

Fifth, harmonization prioritized core availability and omitted sleep, diet, environment, mental health, diagnoses, and disability. This reduced leakage and geographic missingness but limits discrimination and may leave structurally important context unmeasured. Sixth, explicit missing-category and median imputation may be imperfect when missingness is informative; the complete-case analysis selected a different, more complete population. Seventh, 200 overall and 100 subgroup bootstrap replicates provide usable intervals but less tail precision than larger replication; subgroup intervals did not cover every threshold metric, and small or intersectional groups require more data. Eighth, state estimates were descriptive without state-specific confidence intervals or multiple-comparison adjustment. Finally, thresholds optimized balanced accuracy in 2021, not costs, capacity, fairness constraints, or outcomes of an actual program. Prospective validation and governance are required before clinical, eligibility, or individual-level use.

6 CONCLUSION

A survey-weighted HGB model using 17 accessible factors temporally classified fair/poor SRH with ROC-AUC 0.801 and near-ideal calibration slope, but only modestly outperformed weighted logistic regression and retained average underprediction. Structural/access and behavioral information were complementary; eight predictors preserved most performance. Subgroup and state variation cautions against uniform operational deployment. The reproducible framework is most defensible for aggregate public-health surveillance, hypothesis generation, and resource-planning research, followed by local validation, calibration monitoring, community input, and prospective impact evaluation.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The author acknowledges the U.S. Centers for Disease Control and Prevention and the participating state and territorial health departments for collecting and making the Behavioral Risk Factor Surveillance System data publicly available. The author also acknowledges the developers and maintainers of the open-source Python and machine-learning libraries used in this study.

Funding Source

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosure of conflict of interest

The author declares no conflict of interest.

Institutional Review Board Statement

This study was conducted in accordance with the Declaration of Helsinki. Institutional review board review and approval were not required because the study involved a secondary analysis of publicly available, de-identified Behavioral Risk Factor Surveillance System data and did not involve direct interaction with participants or access to identifiable private information.

Data Availability Statement

The 2021 and 2022 Behavioral Risk Factor Surveillance System datasets and accompanying documentation analyzed in this study are publicly available from the U.S. Centers for Disease Control and Prevention at (Link 1) , (Link 2). The analysis code, preprocessing scripts, model configurations, and supplementary outputs are available from the corresponding author upon reasonable request.

AI Use Statement

During the preparation of this manuscript, the author used OpenAI's ChatGPT and Grammarly for language refinement, manuscript organization, and editorial assistance. The author critically reviewed, verified, and revised all AI-assisted content and accepts full responsibility for the accuracy, integrity, analysis, interpretation, and conclusions presented in the manuscript.

REFERENCES

[1] Bombak, A.E. Self-rated health and public health: A critical perspective. *Frontiers in Public Health* 2013, 1, 15. https://doi.org/10.3389/fpubh.2013.00015

[2] Braveman, P.; Gottlieb, L. The social determinants of health: It's time to consider the causes of the causes. *Public Health Reports* 2014, 129(Suppl. 2), 19–31. https://doi.org/10.1177/00333549141291S206

[3] Breiman, L. Random forests. *Machine Learning* 2001, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[4] Centers for Disease Control and Prevention. 2021 BRFSS survey data and documentation. Available online: https://www.cdc.gov/brfss/annual_data/annual_2021.html (accessed on 24 August 2026).

[5] Centers for Disease Control and Prevention. 2022 BRFSS survey data and documentation. Available online: https://www.cdc.gov/brfss/annual_data/annual_2022.html (accessed on 24 August 2026).

[6] Chen, Y.; Zhang, X.; Grekousis, G.; Huang, Y.; Hua, F.; Pan, Z.; Liu, Y. Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (XGBoost) approach. *Journal of Cleaner Production* 2023, 413, 137432. https://doi.org/10.1016/j.jclepro.2023.137432

[7] Chetty, R.; Stepner, M.; Abraham, S.; Lin, S.; Scuderi, B.; Turner, N.; Bergeron, A.; Cutler, D. The association between income and life expectancy in the United States, 2001–2014. *JAMA* 2016, 315(16), 1750–1766. https://doi.org/10.1001/jama.2016.4226

[8] Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 2019, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[9] Clark, C.R.; Ommerborn, M.J.; Moran, K.; Brooks, K.; Haas, J.; Bates, D.W.; Wright, A. Predicting self-rated health across the life course: Health equity insights from machine learning models. *Journal of General Internal Medicine* 2021, 36(5), 1181–1188. https://doi.org/10.1007/s11606-020-06438-1

[10] Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[11] DeSalvo, K.B.; Bloser, N.; Reynolds, K.; He, J.; Muntner, P. Mortality prediction with a single general self-rated health question: A meta-analysis. *Journal of General Internal Medicine* 2006, 21(3), 267–275. https://doi.org/10.1111/j.1525-1497.2005.00291.x

[12] DeSalvo, K.B.; Jones, T.M.; Peabody, J.; McDonald, J.; Fihn, S.; Fan, V.; He, J.; Muntner, P. Health care expenditure prediction with a single item, self-rated health measure. *Medical Care* 2009, 47(4), 440–447. https://doi.org/10.1097/MLR.0b013e318190b716

[13] Franks, P.; Gold, M.R.; Fiscella, K. Sociodemographics, self-rated health, and mortality in the US. *Social Science & Medicine* 2003, 56(12), 2505–2514. https://doi.org/10.1016/S0277-9536(02)00281-2

[14] Friedman, J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 2001, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[15] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Machine Learning* 2006, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[16] Gumà-Lao, J.; Arpino, B. A machine learning approach to determine the influence of specific health conditions on self-rated health across education groups. *BMC Public Health* 2023, 23, 131. https://doi.org/10.1186/s12889-023-15053-8

[17] Hoekstra, J.; Lenssen, E.S.; Wong, A.; Loef, B.; Herber, G.-C.M.; Boshuizen, H.C.; Strak, M.; Verschuren, W.M.M.; Janssen, N.A.H. Predicting self-perceived general health status using machine learning: An external exposome study. *BMC Public Health* 2023, 23, 1027. https://doi.org/10.1186/s12889-023-15962-8

[18] Iachan, R.; Pierannunzi, C.; Healey, K.; Greenlund, K.J.; Town, M. National weighting of data from the Behavioral Risk Factor Surveillance System (BRFSS). *BMC Medical Research Methodology* 2016, 16, 155. https://doi.org/10.1186/s12874-016-0255-7

[19] Idler, E.L.; Benyamini, Y. Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior* 1997, 38(1), 21–37. https://doi.org/10.2307/2955359

[20] Jylhä, M. What is self-rated health and why does it predict mortality? Towards a unified conceptual model. *Social Science & Medicine* 2009, 69(3), 307–316. https://doi.org/10.1016/j.socscimed.2009.05.013

[21] Manderbacka, K.; Lundberg, O.; Martikainen, P. Do risk factors and health behaviours contribute to self-ratings of health? *Social Science & Medicine* 1999, 48(12), 1713–1720. https://doi.org/10.1016/S0277-9536(99)00068-4

[22] National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health, Population Health Surveillance Branch. *Behavioral Risk Factor Surveillance System: 2022 Summary Data Quality Report*; Centers for Disease Control and Prevention: Atlanta, GA, USA, 2023. Available online: https://stacks.cdc.gov/view/cdc/141898 (accessed on 24 August 2026).

[23] Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

[24] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830. Available online: https://www.jmlr.org/papers/v12/pedregosa11a.html (accessed on 24 August 2026).

[25] Pierannunzi, C.; Town, M.; Garvin, W.; Shaw, F.E.; Balluz, L. Methodologic changes in the Behavioral Risk Factor Surveillance System in 2011 and potential effects on prevalence estimates. *Morbidity and Mortality Weekly Report* 2012, 61(22), 410–413. Available online: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6122a3.htm (accessed on 24 August 2026).

[26] Pierannunzi, C.; Hu, S.S.; Balluz, L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology* 2013, 13, 49. https://doi.org/10.1186/1471-2288-13-49

[27] Pronk, N.P.; Kleinman, D.V.; Richmond, T.S. Healthy People 2030: Moving toward equitable health and well-being in the United States. *eClinicalMedicine* 2021, 33, 100777. https://doi.org/10.1016/j.eclinm.2021.100777

[28] Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine* 2018, 169(12), 866–872. https://doi.org/10.7326/M18-1990

[29] Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* 2015, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[30] Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010, 21(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

[31] Subramanian, S.V.; Kawachi, I.; Kennedy, B.P. Does the state you live in make a difference? Multilevel analysis of self-rated health in the US. *Social Science & Medicine* 2001, 53(1), 9–19. https://doi.org/10.1016/S0277-9536(00)00309-9

[32] Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W.; Topic Group "Evaluating Diagnostic Tests and Prediction Models" of the STRATOS Initiative. Calibration: The Achilles heel of predictive analytics. *BMC Medicine* 2019, 17, 230. https://doi.org/10.1186/s12916-019-1466-7

[33] Wadekar, A.S.; Reiter, J.P. Evaluating binary outcome classifiers estimated from survey data. *Epidemiology* 2024, 35(6), 805–812. https://doi.org/10.1097/EDE.0000000000001776

[34] Wiemken, T.L.; Kelley, R.R. Machine learning in epidemiology and health outcomes research. *Annual Review of Public Health* 2020, 41, 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437

[35] Al Kium, A.; Sarker, S.; Shikha, S.A.; Kamal, M.A.T.; Jabed, M.I.K.; Munifa, N.K.; Mitu, M.A.; Pritom, M.H.; Sharmin, F.; John, D.B. Health equity and digital disparities in cancer screening and cardiovascular care across socioeconomic and ethnic groups: A systematic review. *Vascular and Endovascular Review* 2023, 6(2), 35–44. Available online: https://verjournal.com/index.php/ver/article/view/1604 (accessed on 24 August 2026).

[36] Wolff, R.F.; Moons, K.G.M.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S.; PROBAST Group. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine* 2019, 170(1), 51–58. https://doi.org/10.7326/M18-1376

[37] Zajacova, A.; Dowd, J.B. Reliability of self-rated health in US adults. *American Journal of Epidemiology* 2011, 174(8), 977–983. https://doi.org/10.1093/aje/kwr204

[38] Zajacova, A., & Dowd, J. B. (2011). Reliability of self-rated health in US adults. American Journal of Epidemiology, 174(8), 977–983. https://doi.org/10.1093/aje/kwr204