

Using Explainable Machine Learning to Examine Community, Food-Access, and Built-Environment Factors Associated with Census-Tract Diabetes Prevalence in Collin and Denton Counties, Texas

Eshaan Nidee¹

eshaan.nidee.772@k12.friscoisd.org

Liberty High School

Research Article

Keywords: diabetes, social determinants of health, census tract, machine learning, explainable artificial intelligence, spatial epidemiology, health equity

Posted Date: August 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10543051/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Diabetes affects tens of millions of Americans [1], and its prevalence varies markedly between neighboring communities; social determinants of health explain much of that variation independent of individual clinical risk [2,3]. Most predictive modeling in this space still relies on individual-level clinical data; community-level prediction using only public data remains comparatively rare, and few studies separate how much of a model's accuracy comes from actionable community characteristics versus from comorbidities that are themselves close correlates of diabetes.

Methods

We linked American Community Survey (ACS) socioeconomic indicators, USDA Food Access Research Atlas measures [10], OpenStreetMap built-environment densities, and CDC PLACES health estimates [14] across 412 census tracts in Collin and Denton Counties, Texas. Two feature sets were compared: a primary, actionable Model A (community, food-access, and built-environment variables only) and a Model B comparison that adds obesity, physical inactivity, and hypertension prevalence. Baseline, ordinary least squares (OLS), LASSO [17], and random forest [16] models were evaluated with an 80/20 split, 5-fold out-of-fold cross-validation, nested cross-validation with inner hyperparameter tuning [26], and geographic external validation (train Collin, test Denton). SHAP values [18,19] and OLS coefficients provided model interpretation; equity was assessed across income terciles; residual spatial structure was assessed with a Moran scatterplot [22].

Results

Mean tract-level diabetes prevalence was 9.4% (SD 1.9; range 3.6–17.4%). Model A explained a modest, consistent share of tract-to-tract variation: $R^2 = 0.519$ (95% bootstrap CI 0.402–0.624) by cross-validation, 0.523 by nested cross-validation, and 0.592 under geographic external validation. Adding comorbidities (Model B) raised out-of-fold R^2 to 0.935 (95% CI 0.908–0.953), but SHAP attributed most of that gain to hypertension ($r = 0.87$ with diabetes prevalence), so we treat Model B as a comorbidity ceiling rather than the study's central finding. Model A's error was roughly twice as large in low-income tracts (MAE = 1.29) as in high-income tracts (MAE = 0.63; gap 95% CI 0.44–0.89, permutation $p = 0.0001$).

Conclusions

Public community, food-access, and built-environment data predict roughly half of tract-level diabetes variation, comparable to several recently published community- and county-level models [4,5,6], with

reasonable stability out of county. Predictive accuracy is not equity-neutral, and that gap should temper any prioritization use of the resulting model.

Introduction

Diabetes mellitus affects hundreds of millions of people worldwide and remains a leading contributor to cardiovascular disease, kidney failure, blindness, amputation, and premature death [1]. In the United States, an estimated 40.1 million people were living with diabetes as of the most recent national surveillance data, with roughly one in four cases undiagnosed [1]. Diabetes prevalence also varies substantially by geography, even between adjoining neighborhoods within the same metropolitan area, in ways that individual clinical risk factors alone do not explain [2].

A large epidemiologic literature links diabetes risk to socioeconomic status, education, food access, and the built environment. Lower income, less education, and greater poverty are consistently associated with higher diabetes burden [2, 3], food-insecure and low-access neighborhoods have been linked to elevated diabetes and obesity risk in several analyses [10], and more walkable neighborhoods with better park and grocery access are associated with lower obesity and diabetes incidence [12, 13]. These associations are not without controversy: the food-desert literature in particular has been criticized for treating a policy-relevant hypothesis as more settled than the underlying evidence supports, with several UK analyses finding little difference in healthy-food price or availability between poor and affluent neighborhoods once directly measured [11], and prospective, well-controlled built-environment studies remain the exception rather than the rule [12, 13]. This tension motivates evaluating community and environmental predictors on their own, rather than assuming their importance from prior literature.

Machine learning offers tools for evaluating many interacting community characteristics without imposing a fixed functional form, and a growing number of studies have applied it to diabetes prediction. Individual-level models built on electronic health records or claims data often achieve strong discrimination [7, 8, 9], but require protected health information and offer limited guidance for community-level prevention. Community-level studies using public data are fewer, but several recent examples are directly comparable to the present analysis. Using a geographically weighted random forest on county-level CDC data, Quiñones and colleagues found six socioeconomic and lifestyle predictors explained meaningfully more variance locally than global OLS or random forest models, with substantial spatial heterogeneity in which predictor mattered most [4]. At the census-tract level in New York City, Tanner and colleagues used gradient-boosted models and SHAP to show that a 16-variable social-determinants feature set explained 67% of between-tract variance in diabetes prevalence, rising to 80% with 81 variables [5]. Allen applied explainable random forest methods to county-level obesity prevalence and reported 79% of variance explained, with physical inactivity, diabetes, and smoking prevalence as the dominant predictors. This again shows that very high accuracy in this literature is often driven by closely related comorbidities rather than upstream social determinants [6]. Egede and colleagues, using CDC PLACES and Opportunity Insights data across 11,457 tracts nationally, found

healthcare-access variables carried the strongest standardized associations with tract-level diabetes prevalence [3].

A recurring limitation across this literature is that studies combining socioeconomic and comorbidity predictors rarely separate how much apparent accuracy is attributable to each. A model that predicts diabetes largely from hypertension or obesity prevalence is not identifying new, modifiable community-level targets. It is exploiting the well-established clustering of cardiometabolic risk factors. Explainable methods such as SHapley Additive exPlanations (SHAP) let researchers check this distinction by quantifying each predictor's marginal contribution to individual predictions [18, 19], and Rudin has argued more broadly that for high-stakes decisions, models should be interpretable by design rather than explained after the fact [20]. Community-level public health applications sit squarely in this category: a prioritization tool built on an uninterpretable or comorbidity-driven model risks directing resources on the basis of a signal that is not actually actionable.

In this study, we developed an interpretable machine learning framework to predict community-level diabetes prevalence across 412 census tracts in Collin and Denton Counties, Texas, deliberately separating a primary, actionable feature set (ACS socioeconomic and demographic variables, USDA food-access measures, and OpenStreetMap built-environment densities, called Model A) from a comorbidity-enhanced comparison that adds obesity, hypertension, and physical inactivity (Model B). Four algorithms (baseline, OLS, LASSO [17], and random forest [16]) were evaluated under repeated cross-validation, nested hyperparameter tuning [26], and independent geographic external validation. Our primary objective was to determine how well actionable community characteristics alone predict diabetes prevalence and whether that performance generalizes outside the county used for training. Secondary objectives were to quantify the incremental accuracy contributed by comorbidities, and to test whether prediction error is distributed evenly across income groups. This question was motivated by evidence that health-related prediction algorithms can quietly encode disparities even when race or income is never used as an explicit input [24, 25].

To our knowledge, this is among the first census-tract-level diabetes studies to explicitly separate actionable community predictors from downstream health comorbidities while combining geographic external validation, comorbidity-aware explainability, residual spatial analysis, and income-based fairness auditing within a single framework, rather than reporting any one of these checks in isolation.

Methods

Study Design and Population

This was a cross-sectional ecological machine learning analysis of all census tracts in Collin and Denton Counties, Texas, a rapidly growing, socioeconomically diverse metropolitan region. Census tracts were used as the unit of analysis because they approximate neighborhoods while maintaining comparable population sizes. Because this is an ecological design, associations describe population-level patterns

and cannot be interpreted as individual-level or causal relationships, a distinction with a long history in the social sciences, dating to Robinson's classic demonstration that correlations computed on aggregated data need not match, and can even reverse, the corresponding individual-level correlation [21]. After harmonization, the analytical dataset comprised 412 tracts, each with a CDC PLACES diabetes-prevalence estimate and a full set of community-level predictors.

Data Sources

Diabetes prevalence, along with obesity, hypertension, and physical-inactivity prevalence used in the Model B comparison, came from the CDC PLACES database, which produces model-based small-area estimates from the Behavioral Risk Factor Surveillance System, the decennial census, and ACS data using a consistent national methodology [14]. Socioeconomic and demographic predictors (median household income, poverty rate, bachelor's-degree attainment, unemployment rate, and median age) came from the American Community Survey five-year estimates. Food-access measures (a binary low-income/low-access flag and the low-access population share, both at the 1-mile threshold) came from the USDA Food Access Research Atlas [10]; the underlying methodology for defining low-access census tracts follows Rhone and colleagues' national analysis of food-store proximity, income, and vehicle access [10]. Built-environment density measures (grocery, clinic, pharmacy, and park density, plus distance to the nearest mapped park) were derived from an Overpass API snapshot of OpenStreetMap and should be read as an approximation of mapped amenities rather than a verified administrative inventory.

Feature Sets

Two feature sets were constructed to keep actionable and comorbidity-adjacent predictors clearly separated. Model A, the primary and actionable set, comprises the five ACS variables, two USDA food-access variables, and five OSM built-environment variables (12 predictors). Model B adds obesity, physical inactivity, and hypertension prevalence (15 predictors total) and is reported strictly as a comorbidity comparison; it was not used to generate tract-level priority scores.

Predictive Models

Four algorithms of increasing flexibility were compared for both feature sets: a mean-prediction baseline; OLS linear regression, chosen as the principal interpretable benchmark because its coefficients are directly interpretable; standardized LASSO regression [17], which shrinks less informative coefficients toward zero and was expected to help given correlated socioeconomic predictors; and random forest regression [16], evaluated to test whether nonlinear interactions improved on the linear methods. For Model A, random forest hyperparameters (tree count, depth, minimum leaf size, and features per split) were tuned inside a nested cross-validation procedure, following the guidance of Varma and Simon that hyperparameter selection must be confined to an inner loop to avoid optimistic bias in the reported error

[26]. Standard OLS diagnostics (variance inflation factors, residual normality, heteroscedasticity, and Cook's distance) for the Model A fit are reported in Supplementary Materials S1 rather than in the main text.

Model Training and Validation

Model A was evaluated under four complementary procedures rather than a single split: an 80/20 held-out split; 5-fold out-of-fold cross-validation, which generated the primary sensitivity comparison between Model A and Model B and the basis for bootstrap confidence intervals and the fairness analysis; nested 5-fold cross-validation for the tuned random forest [26]; and geographic external validation, training exclusively on Collin County's 219 tracts and testing on Denton County's 193 tracts, a substantially harder test than random splitting, since adjoining tracts within a county often share correlated characteristics. Uncertainty in the 5-fold out-of-fold estimates was further quantified with a paired bootstrap (2,000 resamples), yielding 95% confidence intervals for R^2 , RMSE, and MAE.

Model Evaluation and Explainability

Performance was summarized using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). For OLS, standardized coefficients quantified each predictor's estimated direction and relative magnitude. For the random forest, SHAP values were computed with a TreeExplainer [18, 19] and applied specifically to Model B, to characterize how much of that model's accuracy is carried by comorbidity variables rather than by the actionable community predictors; this framing follows Rudin's broader argument that models used for consequential decisions should be evaluated for what they actually rely on, not only for how accurately they predict [20].

Community Priority Score

A Community Priority Score (CPS) was constructed from Model A predictions only: each tract's predicted diabetes prevalence was converted to a 0–100 percentile rank, with no obesity, hypertension, or inactivity value added, so the score reflects actionable predictors and avoids double-counting known diabetes correlates. The CPS is an illustrative decision-support tool, not a diagnostic or causal instrument.

Spatial and Fairness Analyses

Residual spatial structure was assessed with a Moran scatterplot of standardized Model A residuals against their spatial lag, using Queen-contiguity weights among the 412 tracts [22]; Anselin's local-indicator framework motivates interpreting a positive slope as evidence that geographically neighboring tracts share similarly signed prediction error rather than being spatially independent [23]. Equity was assessed by splitting tracts into income terciles and comparing 5-fold out-of-fold MAE across groups,

with a permutation test (10,000 resamples, two-sided) and paired bootstrap (2,000 resamples) for the low-minus-high MAE gap. This analysis was motivated by evidence, most prominently Obermeyer and colleagues' demonstration that a widely used commercial health-risk algorithm systematically underestimated illness burden in Black patients because it optimized for a cost proxy rather than need [24], and by Rajkomar and colleagues' broader framework for auditing distributive justice in clinical machine learning [25], that predictive accuracy in health-related models cannot be assumed to be uniform across subgroups and must be checked explicitly.

Software

Analyses used Python with scikit-learn for baseline, OLS, LASSO, and random forest models [27]; the SHAP library for explainability [18]; and GeoPandas/PySAL for spatial analysis. Random seeds were fixed (random_state = 42) throughout, and identical preprocessing was applied across algorithms and feature sets.

Results

Dataset Characteristics

The 412 tracts had a mean diabetes prevalence of 9.4% (SD 1.9), ranging from 3.6% to 17.4% (Table 1). This is comparable in spread to the 11.5% mean (SD 3.7%; range 1.9–42.8%) reported across 2,096 New York City tracts by Tanner and colleagues [5], though this Texas sample's narrower range likely reflects the more limited socioeconomic and racial heterogeneity of a two-county suburban region relative to a dense, demographically varied city.

Table 1. Dataset characteristics across 412 census tracts in Collin and Denton Counties, Texas.

Variable	Mean	SD	Min	Max
Diabetes prevalence (%)	9.43	1.88	3.6	17.4
Median household income (\$)	121,080	46,615	8,325	250,001
Poverty rate (%)	7.25	7.92	0.0	80.52
Bachelor's degree or higher (%)	51.67	16.44	6.95	82.73
Unemployment rate (%)	4.02	2.69	0.0	18.18
Median age (years)	38.4	6.46	20.2	74.0
USDA low-income/low-access flag (1-mi)	0.03	0.17	0	1
USDA low-access population share (% , 1-mi)	29.21	31.38	0.0	100.0
Grocery density (per km ²)	0.49	0.75	0.0	5.75
Clinic density (per km ²)	0.32	0.80	0.0	5.87
Pharmacy density (per km ²)	0.16	0.30	0.0	2.03
Park density (per km ²)	0.70	0.93	0.0	11.46
Distance to nearest park (km)	0.90	1.02	0.03	8.06
Obesity prevalence (%)	33.22	3.90	22.3	46.3
Physical inactivity prevalence (%)	20.66	4.56	13.5	39.7
Hypertension prevalence (%)	29.55	4.20	16.2	53.7

Correlation Structure

Diabetes prevalence correlated most strongly with hypertension ($r = 0.87$) and physical inactivity ($r = 0.66$), moderately with obesity ($r = 0.47$), and more weakly with the Model A socioeconomic predictors: bachelor's-degree attainment ($r = -0.49$), median age ($r = 0.36$), and median household income ($r = -0.20$). Poverty rate ($r = 0.11$) and, notably, the unemployment rate ($r = 0.01$) showed little univariate association, a reminder that not every socioeconomic indicator flagged in the broader social-determinants literature [2] carries equal weight in every setting, and consistent with Egede and colleagues' finding that healthcare-access measures, not employment status, carried the strongest standardized association with tract-level diabetes prevalence in their national sample [3]. Substantial multicollinearity among income, education, and poverty motivated including LASSO alongside OLS.

Model Comparison

Table 2 reports performance from the initial 80/20 split across all four algorithms and both feature sets.

Table 2. Model performance, 80/20 held-out split.

Model	Feature set	RMSE	MAE	R ²
Baseline	Model A	2.100	1.453	−0.064
LASSO (standardized)	Model A	1.523	1.040	0.441
Linear regression (OLS)	Model A	1.523	1.040	0.441
Random forest	Model A	1.509	0.996	0.451
Random forest	Model B	0.676	0.474	0.890
Linear regression (OLS)	Model B	0.483	0.376	0.944

All three algorithms clearly outperformed the baseline for Model A, but explained less than half of tract-level variance ($R^2 = 0.44\text{--}0.45$) on this split. Random forest performed marginally better than the two linear approaches ($R^2 = 0.451$ vs. 0.441); LASSO and OLS were essentially identical, indicating shrinkage did not meaningfully change accuracy for a feature set already limited to 12 variables. Table 3 reports four validation procedures to check whether this single-split result was representative.

Table 3. Model A validation across four procedures.

Validation procedure	R ²	RMSE	MAE
80/20 held-out split (linear)	0.441	1.523	1.040
5-fold out-of-fold CV (linear)	0.519	1.301	0.919
Nested 5-fold CV, tuned random forest	0.523	1.268	0.929
Geographic external validation (train Collin, test Denton; linear)	0.592	1.237	0.912

Performance was stable to slightly higher under repeated and out-of-county validation: 5-fold cross-validation ($R^2 = 0.519$), nested tuning of the random forest ($R^2 = 0.523$) [26], and geographic external validation ($R^2 = 0.592$) all exceeded the single split (Fig. 2). Training exclusively on Collin County and testing on the entirely unseen Denton County tracts still explained roughly 59% of variance (RMSE = 1.24, MAE = 0.91 points), a materially harder benchmark than the county-level cross-validation reported by Quiñones and colleagues, whose geographically weighted models were fit and evaluated within the same national county sample rather than transferred to a held-out region [4]. This transfer is a meaningful test of generalizability precisely because the two counties are demographically distinct rather than interchangeable suburbs of the same metropolitan area: tracts in Collin County have significantly higher educational attainment than those in Denton County (55.4% vs. 47.5% with a bachelor's degree or higher, t -test $p < 0.001$), while Denton County tracts have significantly greater USDA low-access population shares (35.9% vs. 23.3%, $p < 0.001$), consistent with Denton's larger rural and exurban fringe outside the Dallas–Fort Worth core. Median household income, poverty rate,

unemployment, and median age did not differ significantly between the two counties (all $p > 0.18$), so the successful transfer specifically demonstrates that Model A generalizes across a meaningful shift in educational composition and food-retail geography rather than merely being re-tested on a similar population under a different label.

Uncertainty

Table 4 gives paired bootstrap 95% confidence intervals for the 5-fold out-of-fold estimates. Model A's R^2 interval (0.402–0.624) is fairly wide, a reminder that with 412 tracts the point estimate of roughly 0.52 carries real sampling uncertainty.

Table 4. Bootstrap 95% confidence intervals, 5-fold out-of-fold predictions, linear regression (2,000 resamples).

Model	Metric	Estimate	95% CI
Model A (community + food/built environment)	R^2	0.519	0.402 – 0.624
Model A (community + food/built environment)	RMSE	1.301	1.135 – 1.468
Model A (community + food/built environment)	MAE	0.919	0.827 – 1.007
Model B (+ health comorbidities)	R^2	0.935	0.908 – 0.953
Model B (+ health comorbidities)	RMSE	0.479	0.406 – 0.563
Model B (+ health comorbidities)	MAE	0.348	0.318 – 0.383

Incremental Value of Health Comorbidities

Adding obesity, physical inactivity, and hypertension (Model B) raised out-of-fold R^2 from 0.519 to 0.935. This pattern mirrors other community-level diabetes and obesity models: Tanner and colleagues' NYC model reached 0.80 only once a much larger, less strictly actionable 81-variable set was used [5], and Allen's county-obesity model reached 0.79 while explicitly flagging that physical inactivity and diabetes prevalence, not socioeconomic variables, drove most of the signal [6]. Because hypertension alone correlates with diabetes prevalence at $r = 0.87$ here, and SHAP confirmed it dominates Model B's predictions (see below), we read Model B as a ceiling on predictability given comorbidity data, not as evidence that community and environmental factors explain 93.5% of diabetes prevalence.

Model Interpretability

OLS Coefficients (Model A)

Socioeconomic disadvantage was generally associated with higher predicted diabetes prevalence, and greater educational attainment and income with lower predicted prevalence, consistent with the broader social-determinants literature [2,3]. Because several predictors are correlated, individual coefficients should be read as each variable's estimated contribution net of the others in this specific model rather than as isolated causal effects [21].

SHAP Analysis (Model B)

SHAP values (Fig. 3) were dominated by hypertension, with obesity and physical inactivity contributing smaller effects and the ACS, USDA, and OSM variables comparatively modest marginal effects once comorbidities were present. This is itself a cautionary result: it shows Model B's high accuracy is carried by hypertension rather than by actionable community variables, reinforcing why it was not used for tract prioritization and echoing Rudin's broader concern that a model's apparent performance can obscure what it is actually using to generate that performance [20].

Community Priority Score

Using Model A predictions only, Table 5 lists the ten highest-priority tracts by percentile-rank Community Priority Score.

Table 5. Ten highest-priority tracts by Community Priority Score (Model A only).

Tract FIPS	County	Model A predicted diabetes (%)	Priority score
48121021535	Denton	16.4	100.0
48085031420	Collin	14.5	99.8
48121020310	Denton	14.4	99.5
48121021414	Denton	13.8	99.3
48085031325	Collin	13.3	99.0
48121021743	Denton	12.9	98.8
48121021744	Denton	12.8	98.5
48085031008	Collin	12.8	98.3
48121021516	Denton	12.7	98.1
48085030903	Collin	12.7	97.8

High-priority tracts by this score appeared in both counties rather than clustering in one (Fig. 4); the ranking is intended as a starting point for local investigation, not a finished resource-allocation tool.

Fairness Across Income Groups

Table 6. Model A out-of-fold prediction error by income tercile.

Income tercile	N tracts	Mean absolute error (points)
Low income	138	1.29
Middle income	136	0.84
High income	138	0.63

Model A's error was roughly twice as large in low-income tracts (MAE 1.29) as in high-income tracts (MAE 0.63) (Table 6); the gap was 0.66 points (95% CI 0.44–0.89; permutation $p = 0.0001$), unlikely to be due to chance. This is a finding about differential model error, not a causal claim about income or healthcare quality, but it echoes a broader pattern documented in health-related algorithms: Obermeyer and colleagues showed a widely deployed cost-prediction algorithm systematically underestimated illness burden among Black patients because it optimized a proxy correlated with, but distinct from, actual need [24], and Rajkomar and colleagues argue that distributive-justice checks of this kind should be a standard, not optional, part of clinical and public health model development [25]. Any use of the Community Priority Score should carry the same caveat: it is least reliable in exactly the lower-income tracts most likely to be flagged by it.

Spatial Structure of Model Residuals

The scatterplot (Fig. 5) showed a positive relationship between each tract's residual and its geographic neighbors' average residual, indicating clustering of under- and over-prediction rather than spatial randomness [22,23]. This suggests some regional structure, plausibly shared healthcare-system boundaries, transportation networks, or development history, is not fully captured by Model A, and argues for regionally coordinated rather than single-tract intervention strategies.

Discussion

A twelve-variable actionable model built entirely from public ACS, USDA, and OpenStreetMap data explained roughly half of tract-to-tract variation in diabetes prevalence ($R^2 \approx 0.52$ by cross-validation, 0.59 under geographic external validation), with reasonably stable performance across four validation schemes. This sits in the same range as other recent community- and county-level efforts: Quiñones and colleagues' geographically weighted random forest of six county-level predictors [4], and below the higher figures reported once larger comorbidity-adjacent feature sets are used, as in Tanner and colleagues' 81-variable NYC model ($R^2 = 0.80$) [5] or Allen's county obesity model ($R^2 = 0.79$), both of which, like ours, found that health-behavior variables rather than socioeconomic ones carried most of

the predictive weight once included [6]. Our results extend this pattern by explicitly separating the two contributions rather than reporting a single blended figure.

Adding obesity, hypertension, and physical inactivity raised out-of-fold R^2 to 0.935, but we deliberately do not treat this as the headline result. Hypertension alone correlates with diabetes prevalence at $r = 0.87$ here, and SHAP confirmed it dominates Model B's predictions, consistent with Allen's finding that diabetes and physical inactivity, not income or education, drove county-level obesity predictions [6], and with the well-established clustering of cardiometabolic risk documented across the social-determinants literature [2]. A model that predicts diabetes largely from a closely correlated comorbidity is not identifying new, actionable targets; it is closer to noticing that two related conditions tend to occur together than to finding a new, upstream cause. We report Model A and Model B separately specifically to avoid the common practice, visible across parts of this literature, of quoting a single high R^2 without clarifying how much of it is attributable to comorbidity variables that are not independently modifiable. That said, Model B's predictors are not immutable either: the Diabetes Prevention Program trial showed that structured lifestyle intervention targeting weight and physical activity reduced incident type 2 diabetes by 58% relative to placebo [15], so these comorbidities remain clinically actionable through individual-level intervention even though they fall outside the community-level policy levers this study was designed to isolate.

Educational attainment showed the strongest univariate association with diabetes prevalence among the Model A predictors, ahead of age and income; poverty and, especially, unemployment carried little independent signal. This partially diverges from Egede and colleagues' larger, national census-tract analysis, where healthcare-access variables (insurance coverage, recent checkups) carried the strongest standardized associations [3]. This difference is plausibly attributable to this study's built-environment and food-access variables substituting for some of that access signal, or simply to genuine regional variation, which the national scientific review of social determinants and diabetes explicitly anticipates given how differently individual determinants are weighted across settings and populations [2]. Our built-environment findings are broadly consistent with, if more modest in magnitude than, the walkability and park-access associations summarized by Howell and Booth [13] and by den Braver and colleagues' meta-analysis of built-environment characteristics and diabetes risk [12], both of which caution that most underlying evidence remains cross-sectional and cannot establish causal direction, a limitation den Braver and colleagues make explicit and one that applies equally here.

The successful geographic external validation, training only on Collin County and testing on entirely unseen Denton County tracts, is among the more informative results here because random splits within a metropolitan county tend to overstate transferability, given how strongly adjoining tracts share demographic and socioeconomic characteristics. That Model A performed as well or slightly better out of county ($R^2 = 0.592$) than under within-sample cross-validation ($R^2 = 0.519$) suggests the captured relationships are not artifacts of a single county's demographic mix, though this remains a two-county comparison within one metropolitan region.

The income-related error gap identified here (permutation $p = 0.0001$) fits a broader pattern documented elsewhere, most prominently Obermeyer and colleagues' racial-bias finding in a widely deployed health-risk algorithm [24], showing that predictive accuracy in health-related models is not automatically distributed evenly across groups even when group membership is never used as an input. Rajkomar and colleagues' framework for auditing distributive justice in clinical machine learning [25] is written primarily for individual-level clinical prediction, but the underlying principle, that accuracy differentials must be checked and disclosed rather than assumed, applies just as directly to a community-prioritization tool like the CPS presented here. Any downstream use of this or a similar framework should disclose that its predictions are systematically less reliable in the lower-income communities most likely to be flagged by it.

Positive spatial autocorrelation in Model A's residuals indicates some regional structure in diabetes prevalence that tract-level predictors alone do not capture, consistent with a spatial-epidemiology literature dating to Moran's original formulation of spatial autocorrelation statistics [22] and Anselin's subsequent local-indicator framework for characterizing where such clustering occurs [23]. This argues for regionally coordinated prevention strategies rather than interventions aimed at isolated high-priority tracts.

Finally, this study's cross-sectional, ecological design constrains every claim made above to the population level. Robinson's classic demonstration that aggregate correlations need not reproduce, and can even reverse, individual-level correlations [21] is a caution worth restating plainly: nothing here should be read as evidence about any individual resident's risk, and the modifiable-versus-fixed framing of Model A's predictors describes community-level policy targets, not individual behavior change.

Limitations

Cross-sectional, ecological design. Associations describe population-level, not individual-level or causal, relationships [21].

Modeled outcome and comorbidity data. CDC PLACES estimates are themselves small-area model outputs rather than direct measurements [14], and Model B's comorbidity variables share that same modeling lineage with the diabetes outcome, which likely inflates their apparent independence as predictors.

Built-environment data currency. OpenStreetMap densities are a single-date crowd-sourced snapshot, not a verified administrative facility inventory, with coverage that can vary by tract.

Geographic scope. Findings are limited to two counties in one North Texas metropolitan region; external validation across rural areas, other metros, and other states is needed before generalizing further.

No causal or quasi-experimental component. This analysis is purely predictive and correlational and makes no natural-experiment or before/after claim about any specific local intervention.

Differential accuracy by income. As detailed above, Model A predicts less accurately in lower-income tracts, a constraint that should accompany any downstream use of the Community Priority Score [24, 25].

OLS assumption violations. The full-sample OLS fit underlying the Community Priority Score and spatial residual analysis shows statistically significant heteroscedasticity and mild residual non-normality, driven partly by a small number of high-leverage tracts (Supplementary Materials S1). Variance inflation factors were low, so multicollinearity is not a concern, but the heteroscedasticity is consistent with, and likely partly explained by, the income-related error gap documented in the fairness analysis.

Conclusion

Public community, food-access, and built-environment data predict a moderate, not near-total, share of tract-to-tract variation in diabetes prevalence in this two-county Texas sample ($R^2 \approx 0.52$ by cross-validation, 0.59 out of county), broadly in line with comparable community- and county-level models published elsewhere [3, 4, 5, 6]. Adding established comorbidities substantially raises apparent accuracy, but that increase is carried by hypertension and should be read as a ceiling for comparison, not as evidence that socioeconomic and environmental factors alone explain the bulk of diabetes burden. Prediction error is not distributed evenly across income groups, a disparity that should inform, and temper, any prioritization use of this or similar frameworks [24, 25]. Future work should extend external validation to additional regions, investigate the income-related error gap directly, and add explicitly spatial modeling terms to address the residual geographic clustering identified here [22, 23].

Declarations

Ethics approval and consent to participate

Not applicable. This study used only publicly available, de-identified, aggregate census-tract-level data from CDC PLACES, the U.S. Census Bureau American Community Survey, the USDA Food Access Research Atlas, and OpenStreetMap. No individual human subjects data were collected, accessed, or analyzed.

Consent for publication

Not applicable.

Availability of data and materials

All data used in this study are publicly available. CDC PLACES data are available at <https://www.cdc.gov/places>. American Community Survey data are available from the U.S. Census

Bureau at <https://www.census.gov/programs-surveys/acs>. USDA Food Access Research Atlas / SNAP-authorized Retailer Access Map data are available at <https://www.ers.usda.gov>. OpenStreetMap data were queried via the Overpass API (<https://overpass-api.de>). The merged tract-level dataset and analysis code generated for this study are available from the corresponding author upon reasonable request.

Competing interests

The author declares no competing interests.

Funding

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

Authors' contributions

E.N. is the sole author of this work and was responsible for all aspects of the study, including conceptualization, data acquisition, methodology, formal analysis, and writing of the manuscript.

Acknowledgements

Not applicable.

References

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report. Atlanta, GA: US Department of Health and Human Services; report updated January 2026. Available at: <https://www.cdc.gov/diabetes/php/data-research/index.html>
2. Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, Thornton PL, Haire-Joshu D. Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*. 2021;44(1):258–79.
3. Egede LE, Walker RJ, Linde S. Association between population level social risk factors and prevalence of diabetes. *BMC Health Serv Res*. 2025;25:1252.
4. Quiñones S, Goyal A, Ahmed ZU. Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. *Sci Rep*. 2021;11:6955.
5. Tanner D, Zhang Y, Chang JE, Speyer P, Adamson E, Aerts A, Lavista Ferres JM, Weeks WB. Machine learning to evaluate the relationship between social determinants and diabetes prevalence in New

- York City. *BMJ Public Health*. 2024;2(2):e001394.
6. Allen B. An interpretable machine learning model of cross-sectional U.S. county-level obesity prevalence using explainable artificial intelligence. *PLoS ONE*. 2023;18(10):e0292341.
 7. Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int J Environ Res Public Health*. 2021;18(6):3317.
 8. Afsaneh E, Sharifdini A, Ghazzaghi H, Ghobadi MZ. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetol Metab Syndr*. 2022;14:196.
 9. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput Struct Biotechnol J*. 2017;15:104–16.
 10. Rhone A, Ver Ploeg M, Williams R, Breneman V. Understanding Low-Income and Low-Access Census Tracts Across the Nation: Subnational and Subpopulation Estimates of Access to Healthy Food. *Economic Information Bulletin 209*. Washington, DC: USDA Economic Research Service; 2019.
 11. Cummins S, Macintyre S. Food deserts—evidence and assumption in health policy making. *BMJ*. 2002;325(7361):436–8.
 12. den Braver NR, Lakerveld J, Rutters F, Schoonmade LJ, Brug J, Beulens JWJ. Built environmental characteristics and diabetes: a systematic review and meta-analysis. *BMC Med*. 2018;16:12.
 13. Howell NA, Booth GL. The Weight of Place: Built Environment Correlates of Obesity and Diabetes. *Endocr Rev*. 2022;43(6):966–83.
 14. Greenlund KJ, Lu H, Wang Y, Matthews KA, LeClerc JM, Lee B, et al. PLACES: Local Data for Better Health. *Prev Chronic Dis*. 2022;19:210459.
 15. Knowler WC, Barrett-Connor E, Fowler SE, Diabetes Prevention Program Research Group, et al. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. *N Engl J Med*. 2002;346(6):393–403.
 16. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32.
 17. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Ser B*. 1996;58(1):267–88.
 18. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. 2017:4765–74.
 19. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
 20. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206–15.
 21. Robinson WS. Ecological Correlations and the Behavior of Individuals. *Am Sociol Rev*. 1950;15(3):351–7.
 22. Moran PAP. Notes on Continuous Stochastic Phenomena. *Biometrika*. 1950;37(1–2):17–23.

23. Anselin L. Local Indicators of Spatial Association—LISA. *Geogr Anal.* 1995;27(2):93–115.
24. 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):447–53.
25. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med.* 2018;169(12):866–72.
26. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 2006;7:91.
27. 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 E. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825–30.

Figures

Correlation Among Diabetes Prevalence, Community Predictors, and Health Comorbidities (n = 412 Census Tracts)

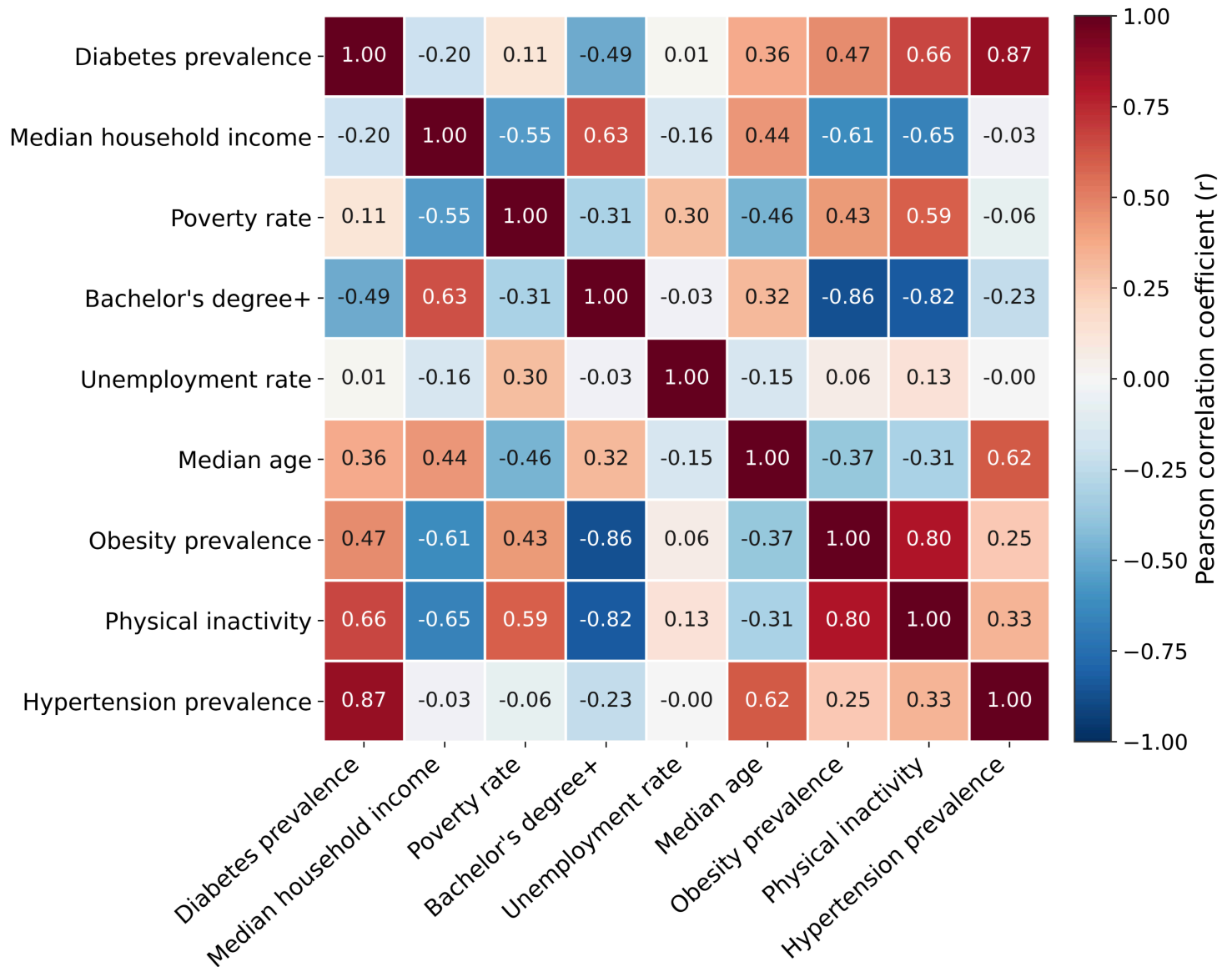


Figure 1

Pearson correlation matrix among diabetes prevalence, community predictors, and health comorbidities.

Model A Validation Across Four Procedures (Community, Food-Access, and Built-Environment Predictors; n = 412 Tracts)

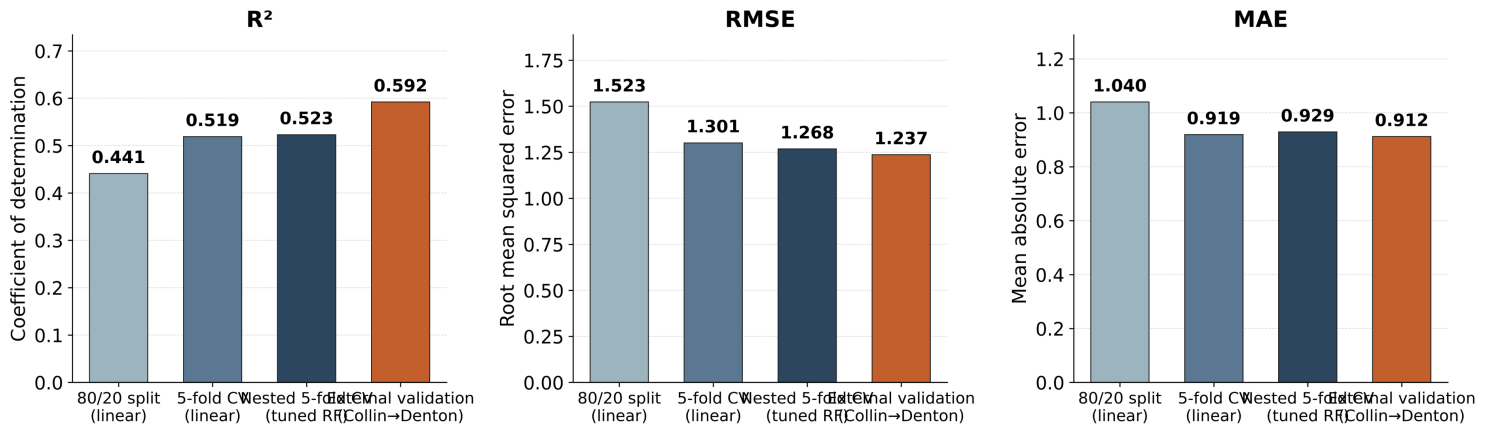
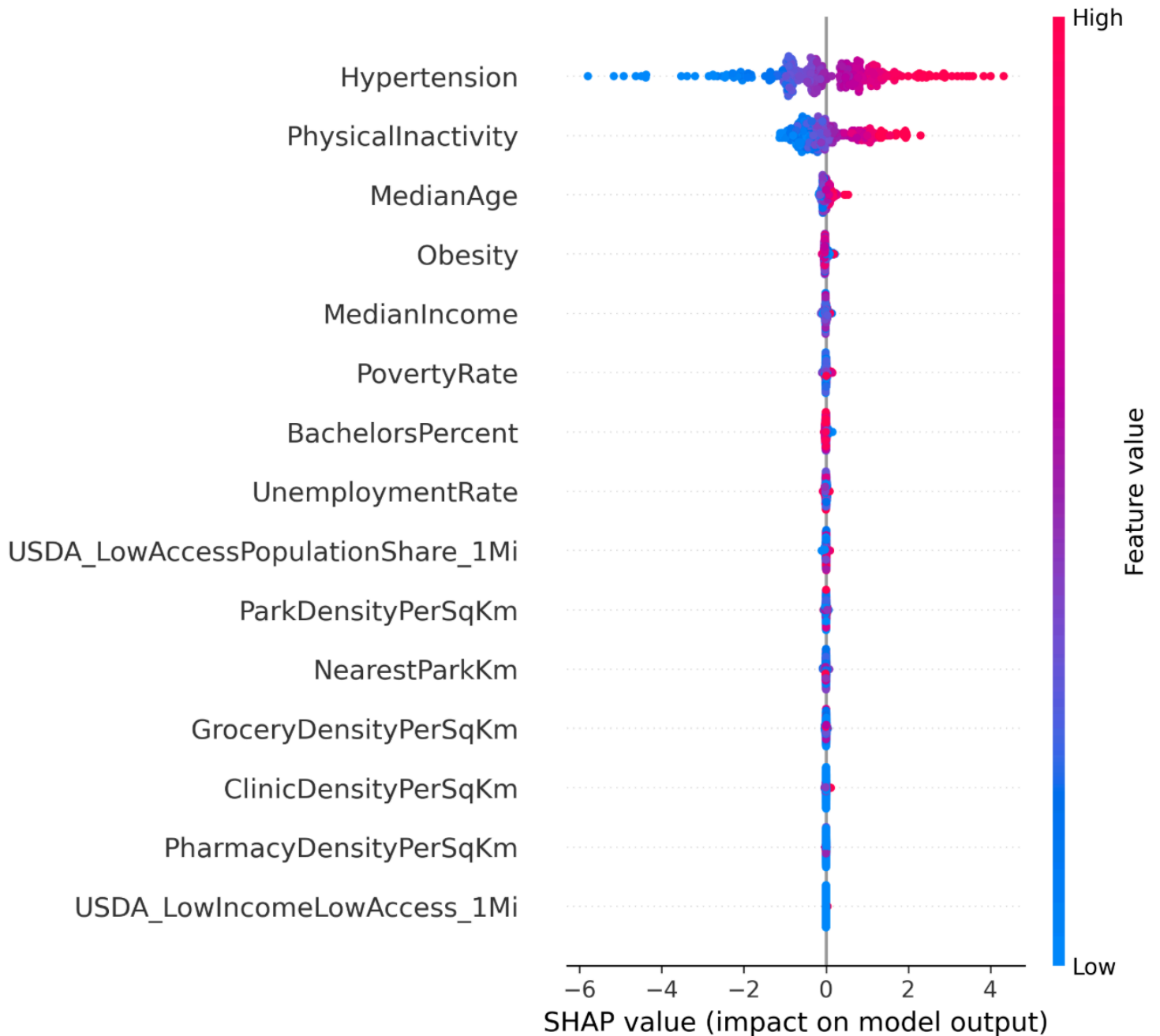


Figure 2

Model performance comparison across algorithms and validation schemes.

SHAP: Model B (community + health comparison)



Interpret as a comorbidity comparison; Model B is not used for tract prioritization.

Figure 3

SHAP summary plot, random forest fit to Model B (community + food/built environment + comorbidities).

Community Priority Score (Model A: food/built-environment features)

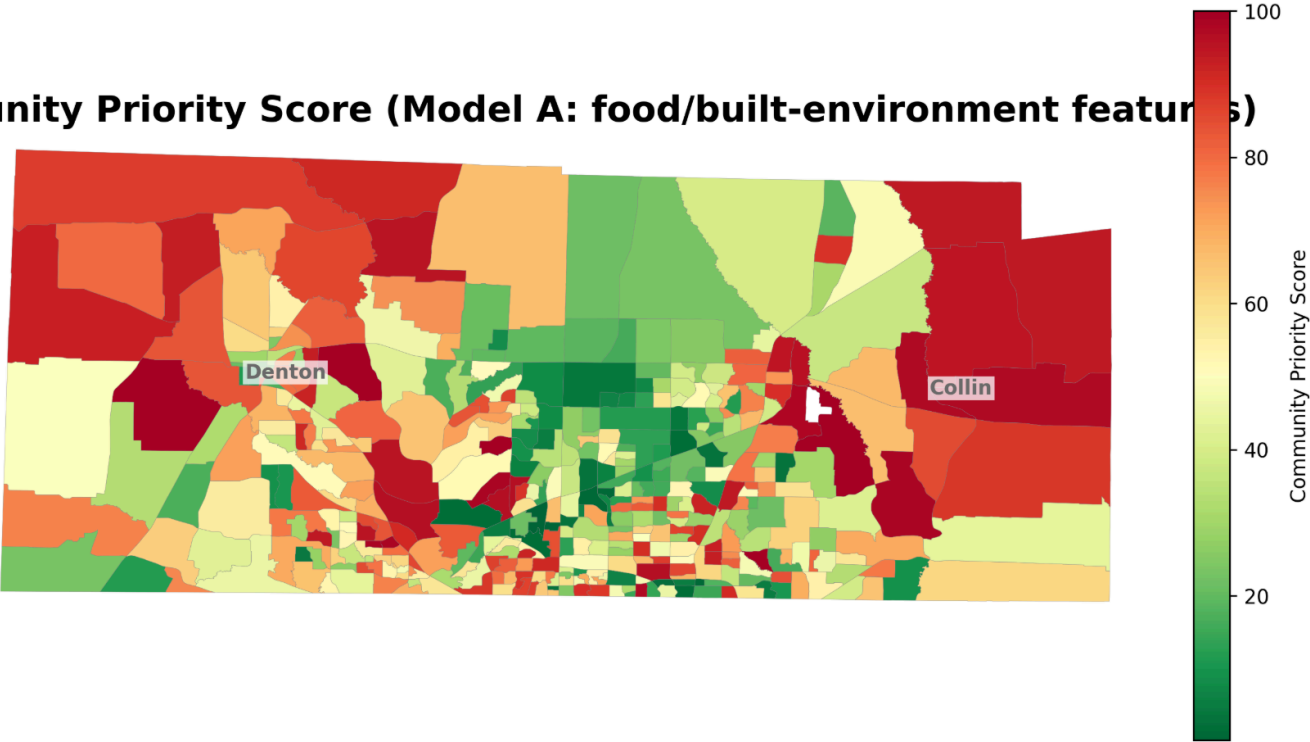


Figure 4

Community Priority Score by census tract, Collin and Denton Counties.

Moran Scatterplot

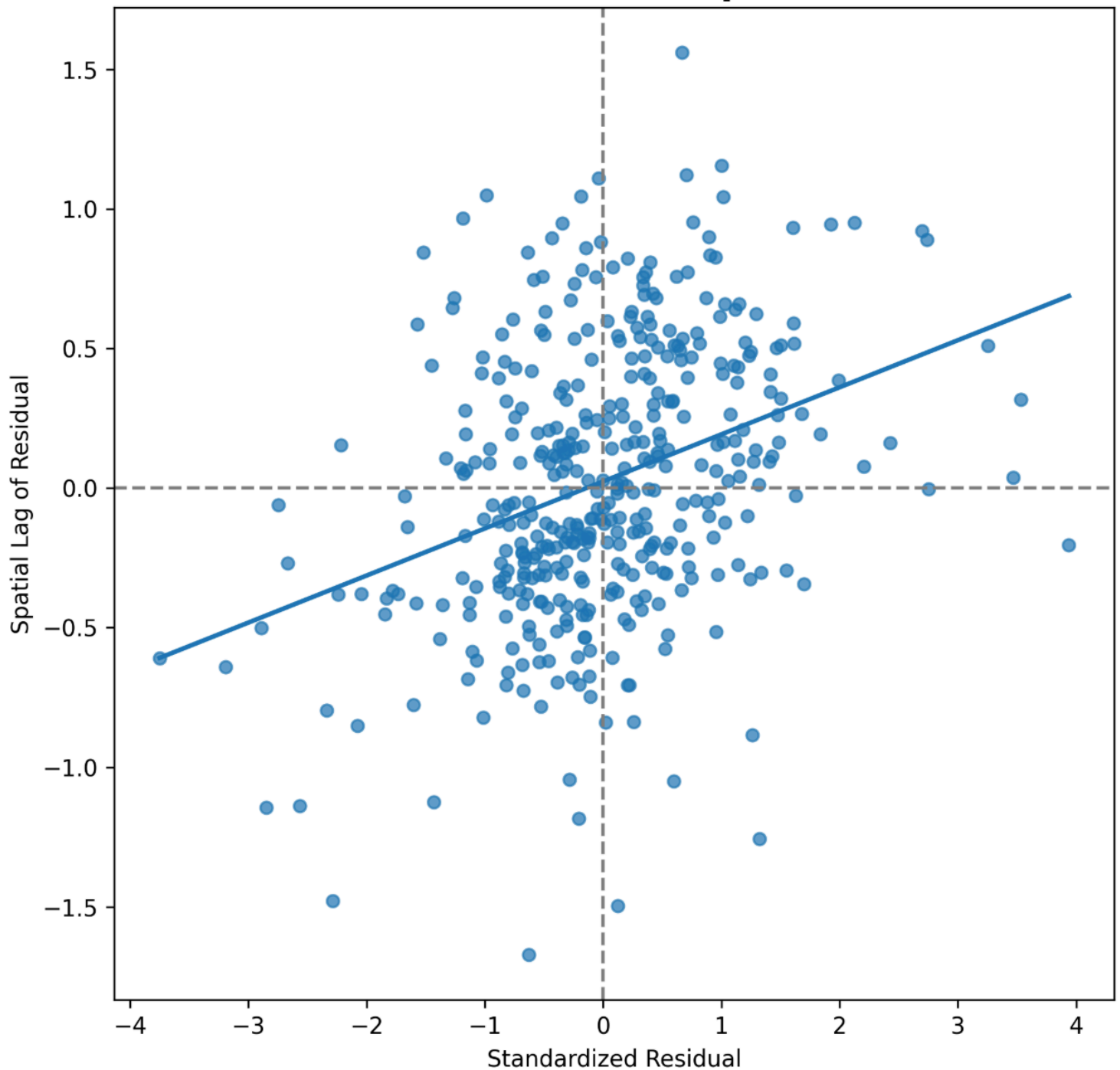


Figure 5

Moran scatterplot of standardized Model A residuals against their spatial lag.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.docx](#)