Geographic Disparities in Diabetes Prevalence Across U.S. Counties: An Explainable Machine Learning Study

Cagatay Berke Erdas
Department of Computer Engineering, Baskent University, Ankara, Turkey

Abstract:

Background: Diabetes represents a substantial public health burden, with considerable geographic variation in prevalence. Identifying population-level factors associated with this variation may support data-driven public health surveillance and targeted prevention strategies. This study investigated county-level disparities in diagnosed diabetes prevalence across the United States and evaluated the ability of explainable machine learning methods to identify population-level characteristics associated with these disparities.

Methods: County-level data were obtained from the 2025 release of the Centers for Disease Control and Prevention (CDC) PLACES dataset. Diagnosed diabetes prevalence among adults was used as the outcome. Six population-level predictors were examined: obesity, no leisure-time physical activity, current cigarette smoking, short sleep duration, lack of health insurance among adults aged 18–64 years, and binge drinking. Descriptive statistics and Pearson and Spearman correlation analyses were performed. Nine regression algorithms—Linear Regression, Ridge, Lasso, Elastic Net, Random Forest, Extra Trees, XGBoost, LightGBM, and CatBoost—were evaluated using 10-fold cross-validation. Performance was assessed using R^2, mean absolute error (MAE), and root mean squared error (RMSE). State-grouped validation was additionally used to assess geographic generalizability. SHapley Additive exPlanations (SHAP) were used to interpret the best-performing model.

Results: A total of 2,956 U.S. counties were included. Mean diagnosed diabetes prevalence was 13.64±2.78%, ranging from 4.9% to 27.1%. No leisure-time physical activity showed the strongest positive correlation with diabetes prevalence (Pearson $r=0.853$; Spearman $\rho=0.843$), followed by current cigarette smoking ($r=0.714$) and obesity ($r=0.670$). Extra Trees achieved the highest performance in conventional 10-fold cross-validation ($R^2=0.880\pm0.013$, MAE=0.718±0.038 percentage points, RMSE=0.956±0.061 percentage points), followed by Random Forest ($R^2=0.876\pm0.012$) and LightGBM ($R^2=0.872\pm0.012$). Under state-grouped validation, Extra Trees retained substantial performance ($R^2=0.777\pm0.101$, MAE=0.950±0.230, RMSE=1.197±0.262). SHAP identified no leisure-time physical activity as the dominant predictor (mean $|SHAP|=1.124$), followed by binge drinking (0.756), current cigarette smoking (0.386), obesity (0.190), lack of health insurance (0.127), and short sleep duration (0.117).

Conclusion: Explainable machine learning demonstrated strong ability to characterize geographic variation in county-level diabetes prevalence. Physical inactivity emerged as the most prominent population-level correlate. The reduction in performance under state-grouped validation highlights the importance of accounting for geographic structure when evaluating public health prediction models. Ecological associations should not be interpreted as individual-level causal effects.

Keywords:

Diabetes, public health, epidemiology, machine learning, explainable artificial intelligence, SHAP, health disparities, CDC PLACES.