Predicting County-Level Diabetes Prevalence in the United States Using Machine Learning and Geospatial Analysis

Joseph Grob

December 2024

Abstract

This study predicts diabetes prevalence in U.S. counties for 2019 using data from 2016-2018 and machine learning models. We compared XGBoost, Random Forest, Geographical Random Forest, and a baseline linear regression model. XGBoost was selected as the final model for its superior performance, achieving the highest R^2 on the test dataset ($R^2 = 0.4158$) and the lowest test error (RMSE test = 0.7643). SHAP values, derived from XGBoost, identified the most influential factors driving diabetes prevalence at the county level, including *physical inactivity*, *excessive drinking*, and *smoking*. This analysis provides insights into regional health disparities and key drivers of diabetes, aiding targeted public health strategies.

1 Introduction

Diabetes is a major U.S. public health challenge, affecting 37.3 million people (County Health Rankings, 2021). Its prevalence varies across regions due to demographic, geographic, and lifestyle factors, emphasizing the need for targeted interventions.

Research Question: How do demographic, geographic, environmental, and lifestyle factors, identified through interpretable machine learning models, contribute to explaining diabetes prevalence at the county level in the U.S., and how does their influence and variability differ across regions?

Objectives:

- To predict the percentage of people diagnosed with diabetes at the county level in the U.S. from 2016 to 2019 using machine learning models.
- To identify and quantify the key factors contributing to diabetes prevalence using interpretable methods.

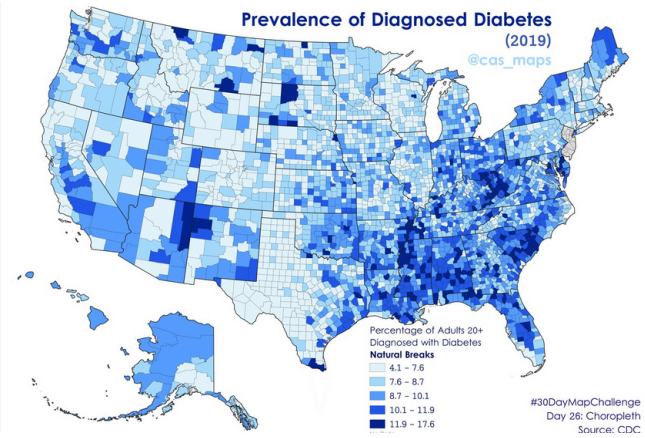

Figure 1: Prevalence of Diagnosed Diabetes in U.S. Counties (2019). Source: CDC.

2 Methodology

This study followed a systematic approach to analyze diabetes prevalence across U.S. counties using data from the US Census [3] and the 2024 County Health Rankings [2]. These datasets provide comprehensive county-level information on demographic, environmental, socioeconomic, and health-related variables. Data from 2016-2019 was merged, standardized, and explored using summary statistics, distribution plots, and correlation matrices. The data from 2016-2018 was split into 80% for training and 20% for validation, with 2019 reserved for testing to ensure generalization and capture temporal trends.

A baseline linear regression model was trained, with features selected using AIC and BIC criteria applied to an OLS regression. Predictive models, including Random Forest, Geographical Random Forest,

and XGBoost, were then trained on the selected features. SHAP values were used to interpret model predictions and identify key factors contributing to diabetes prevalence, emphasizing their geographical variability. K-means clustering grouped counties with similar diabetes-related characteristics, providing insights into regional disparities and interactions between factors.

Methodology used in this study

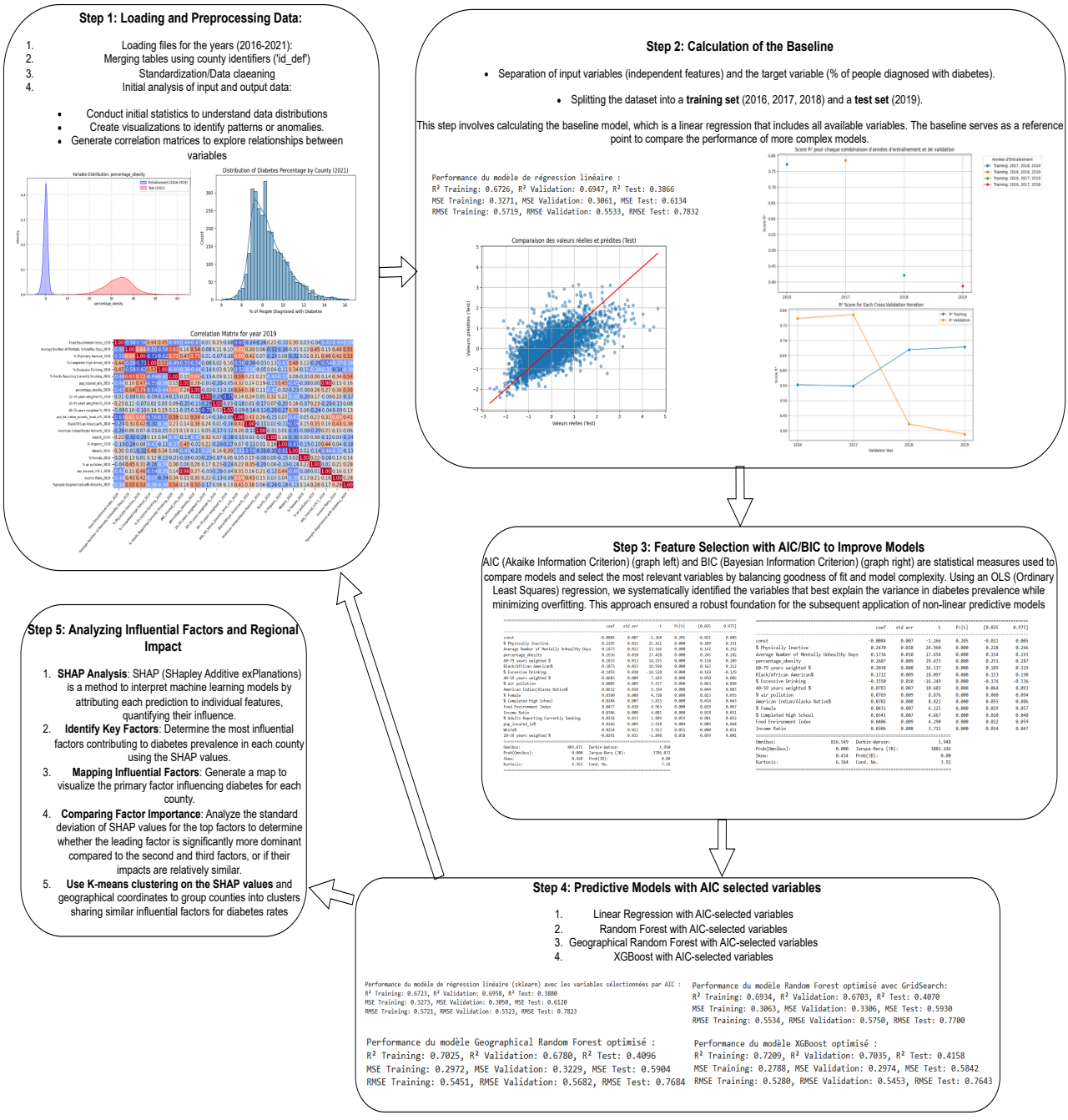

Figure 2: Overview of the methodology used in the study.

3 Results

Linear Regression Performance

A Linear Regression model using AIC-selected variables demonstrated a slight improvement over the baseline but remained limited in performance (R^2 test = 0.3880). Overfitting persisted, as indicated by the gap between training and test performance (R^2 train = 0.6723, R^2 test = 0.3880). Additionally, the training, validation, and test errors remained relatively high (MSE train = 0.3273, MSE val = 0.3050, MSE test = 0.6120; RMSE train = 0.5721, RMSE val = 0.5523, RMSE test = 0.7823), suggesting limited generalizability.

Optimized Model Performance

Table 1 lists the hyperparameters used for Random Forest, Geographical Random Forest, and XGBoost, trained with AIC-selected variables. Table 2 compares their performances. XGBoost achieved the highest R^2 test score (0.4158) and the lowest test error (RMSE test = 0.7643), outperforming Geographical Random Forest (RMSE test = 0.7684) and Random Forest (RMSE test = 0.7700).

Random Forest with GridSearch performed reasonably well but exhibited larger training-test discrepancies (RMSE train = 0.5534, RMSE test = 0.7700). Geographical Random Forest showed good performance (R^2 test = 0.4096) but was slightly outperformed by XGBoost in both accuracy and generalization.

While XGBoost achieved the best results among the models tested, its R^2 score of 0.4158 is relatively low. This indicates that a substantial portion of the variance in diabetes prevalence remains unexplained, reflecting the complexity of the data and the need for further model improvements or additional features.

Model Selection

XGBoost was selected as the best model for its superior performance and generalization. It achieved the highest R^2 on the test dataset (R^2 = 0.4158) and the lowest test error (RMSE test = 0.7643). Its validation performance (R^2 val = 0.7035, RMSE val = 0.5453) further supports its robustness.

Random Forest and Geographical Random Forest also showed good predictive capabilities but were slightly less accurate than XGBoost. For instance, Geographical Random Forest achieved R^2 test = 0.4096 with RMSE test = 0.7684, while Random Forest achieved R^2 test = 0.4070 with RMSE test = 0.7700.

XGBoost emerged as the most robust model, capturing complex patterns in diabetes prevalence while maintaining better performance across training, validation, and test datasets.

SHAP Analysis and Factor Interpretation

Dominant Factors: The most influential factors identified by SHAP (Figure 3) include: % Physically Inactive, % Excessive Drinking, % percentage obesity, Average Number of Mentally Unhealthy Days, 60–79 years weighted %, Adults Reporting Currently Smoking, and 20–39 years weighted %. These factors are strongly linked to diabetes prevalence through physical, behavioral, and demographic pathways. Among these, % Physically Inactive is the most significant predictor, followed by % Excessive Drinking and % percentage obesity.

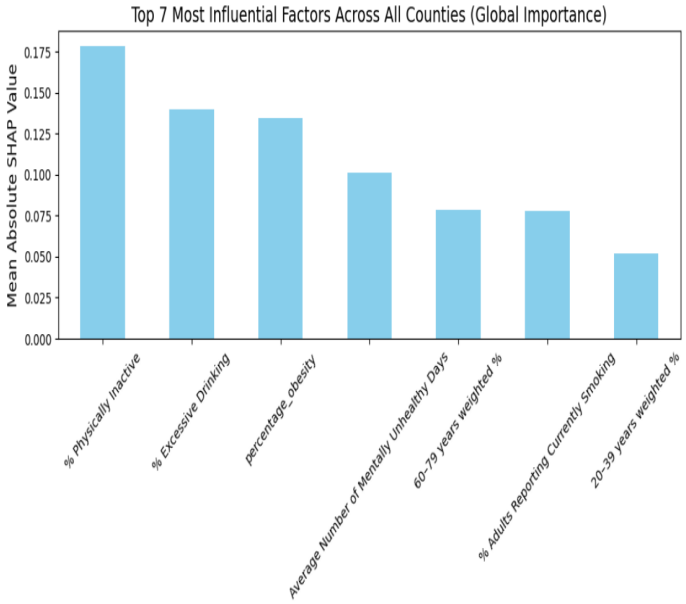

Figure 3: Top 7 Most Influential Factors Across All Counties (Global Importance) based on SHAP values.

Regional Patterns: The map (Figure 4) highlights patterns in the dominant factors influencing diabetes prevalence across U.S. counties:

- **Physical inactivity** (orange) is the most widespread factor, particularly dominant in the western U.S.

- **Excessive drinking** (purple) is significant in the central and eastern U.S., overlapping with economically disadvantaged areas.

- **Smoking** (red) is concentrated in states like Indiana and Ohio.

- **Obesity** (blue) is most visible in urban centers such as New York, Los Angeles, and Chicago.

Urban areas show a varied distribution of dominant factors, reflecting the complexity of health determinants.

The Southeast, known as the "Diabetes Belt," faces high diabetes prevalence due to a combination of obesity (blue), smoking (red), and excessive drinking (purple) (Figure 1).

These observations emphasize the geographic variability of dominant factors and the need for regionally tailored public health strategies.

Dominant factors in counties, USA

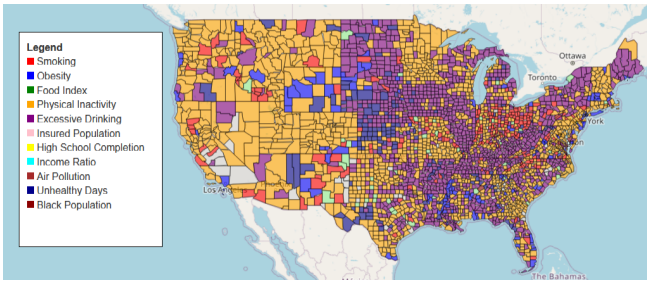

Figure 4: Map showing dominant factors in urban and rural counties.

Consistency of Factors: SHAP scatter plots (Figure 5) reveal that most counties have low standard deviations between dominant and secondary factors, showing co-occurrence of multiple factors influencing diabetes prevalence.

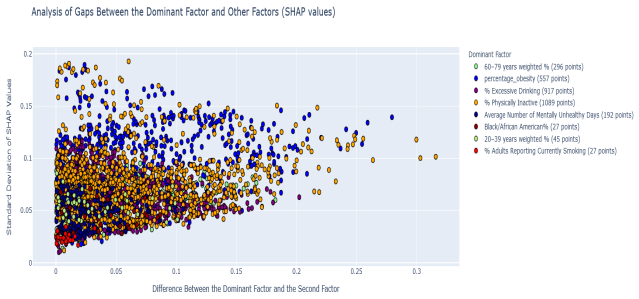

Figure 5: SHAP scatter plot showing consistency between dominant and secondary factors.

Cluster Analysis

The clusters (Figure 6) are based on the 7 most influential factors for each county, ranked by importance, providing a grouping of counties with similar health profiles.

- **Cluster 0 (Blue):** Predominantly in the Midwest and Southeast, these counties are dominated by physical inactivity and smoking, with moderate contributions from mental health issues and obesity. Interventions promoting active lifestyles and reducing smoking are needed.

- **Cluster 1 (Red):** Found mainly in the Southeast, these counties are strongly influenced by excessive drinking, along with physical inactivity and smoking. Public health programs should focus on reducing alcohol consumption and improving lifestyles.

- **Cluster 2 (Pink):** Located across the western and northern U.S., these counties show minimal or negative influences for most factors, indicating lower health risks. These areas can serve as benchmarks for strategies in higher-risk regions.

- **Cluster 3 (Cyan):** Concentrated in the Southeast and Midwest, these counties face high obesity and physical inactivity, with moderate contributions from mental health issues and smoking. Comprehensive interventions are needed to address combined lifestyle and socioeconomic challenges.

Clusters 0, 1, and 3, concentrated in the Southeast, highlight dominant factors such as alcohol consumption, smoking, and physical inactivity. These align with regions where diabetes prevalence is highest, emphasizing significant health disparities.

Clusters identified in US counties, based on the 7 most influential factors

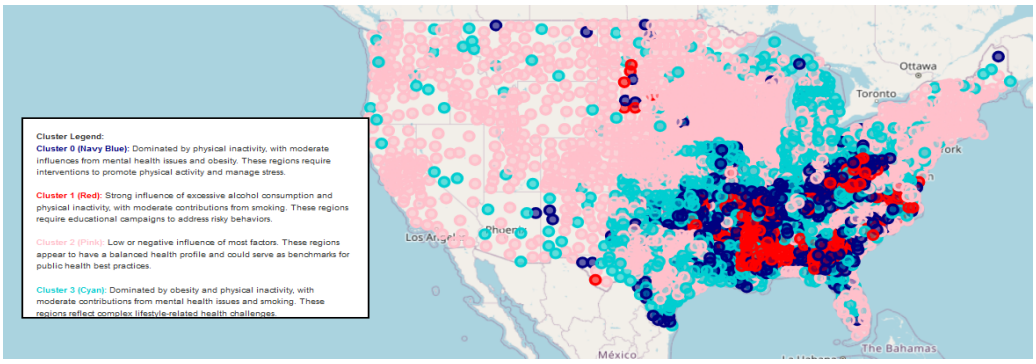

Figure 6: Clustering analysis highlighting regional distributions of dominant health factors.

Hyperparameters search for XGBoost, Random Forest and Geographical Random Forest

Model	Hyperparameter	Optimal Value	Range of Hyperparameters	Justification
XGBoost	Learning Rate	0.05	[0.01, 0.05]	A lower learning rate ensures gradual optimization, reducing the risk of overshooting minima.
	Max Depth	4	[1, 4]	A smaller depth captures simple patterns and prevents overfitting.
	Number of Estimators	75	[50, 75]	A smaller number of trees stabilizes predictions and avoids over-complexity.
	Subsample	0.7	[0.5, 0.6, 0.7]	Randomly selecting subsets of data prevents overfitting.
	Colsample by Tree	0.6	[0.5, 0.6, 0.7]	Sampling a subset of features ensures diversity and reduces overfitting.
	Gamma	0	[0, 1, 5]	Penalizes splits with low gain, encouraging meaningful splits.
	Regularization (L1, α)	0.1	[0, 0.1, 0.5, 1]	Moderate L1 regularization adds sparsity, reducing noise.
	Regularization (L2, λ)	10	[5, 10, 20]	Stronger L2 regularization penalizes large weights, stabilizing the model.
Random Forest	Number of Estimators	75	[50, 75]	A smaller number of trees reduces computational cost without sacrificing accuracy.
	Max Depth	6	[2, 6]	Limiting depth ensures simpler, generalizable models.
	Min Samples Split	10	[10, 15]	A higher minimum split size reduces overfitting on small partitions.
	Min Samples Leaf	5	[5, 10]	Ensures larger leaves, avoiding noise in predictions.
	Max Features	sqrt	[sqrt, log2]	Limits the number of features per tree to ensure diversity.
	Bootstrap	True	[True]	Bootstrapping ensures diversity in sampling, reducing overfitting.
Geographical Random Forest	Number of Estimators	75	[50, 75]	A smaller number of trees enhances stability without overfitting.
	Max Depth	6	[2, 6]	Captures regional trends without overfitting on smaller data.
	Min Samples Split	10	[5, 10]	Allows flexibility in capturing granular splits while avoiding noise.
	Min Samples Leaf	4	[2, 4]	Ensures larger leaf sizes for better generalization.
	Features Used (lat, lng)	Latitude, Longitude	[Latitude, Longitude]	Includes spatial coordinates to capture geographical relationships in predictions.
	Bootstrap	True	[True]	Ensures spatial diversity in sampling, capturing more general patterns.

Table 1: Hyperparameter search, ranges, and optimal values for XGBoost, Random Forest, and Geographical Random Forest models, with justifications for selected values.

Performance comparison of models

Model	R^2 Training	R^2 Validation	R^2 Test	MSE Training	MSE Validation	MSE Test	RMSE Training	RMSE Validation	RMSE Test
Baseline (Linear Regression without AIC/BIC)	0.6726	0.6947	0.3866	0.3271	0.3061	0.6134	0.5719	0.5533	0.7832
Linear Regression with AIC/BIC	0.6723	0.6958	0.3880	0.3273	0.3050	0.6120	0.5721	0.5523	0.7823
Random Forest with GridSearch	0.6934	0.6703	0.4070	0.3063	0.3306	0.5930	0.5534	0.5750	0.7700
Geographical Random Forest with GridSearch	0.7025	0.6780	0.4096	0.2972	0.3229	0.5904	0.5451	0.5682	0.7684
Optimized XGBoost with RandomizedSearch	0.7209	0.7035	0.4158	0.2788	0.2974	0.5842	0.5280	0.5453	0.7643

Table 2: Performance comparison of models with R^2, MSE, and RMSE across training, validation, and test sets.

Conclusion

This study identified key factors driving diabetes prevalence in U.S. counties, with physical inactivity, excessive drinking, and obesity as the most significant predictors. Regional clustering highlighted compounded risks in the Southeast and Midwest, emphasizing the need for targeted interventions. While XGBoost performed best ($R^2 = 0.4158, \text{RMSE test} = 0.7643$), its modest R^2 score indicates room for improvement. Future work should integrate additional data and advanced techniques to enhance predictive accuracy and inform public health strategies. Expanding these methods to other chronic diseases could provide broader insights into regional health disparities.

Code Availability

The code used for this project, including the full pipeline and analysis scripts, is publicly available on GitHub (take the last code on the commit history). The code is contained in a Jupyter Notebook format (`.ipynb`), which must be downloaded to be visualized and executed. You can access it at the following link:

GitHub Repository: MLEES Final Code Project.

Dataset Availability

The dataset used in this study, including all four Excel files and CSV "latitude/longitude" file, is publicly available on GitHub. You can access it at the following link:

GitHub Repository: Dataset for Final Project.

Bibliography

References

[1] Centers for Disease Control and Prevention (CDC). (n.d.). *National Center for Health Statistics: Health, United States Resources*. Division of Analysis and Epidemiology. Retrieved from `https://www.cdc.gov/nchs/hus/resources.htm`.

[2] County Health Rankings & Roadmaps. (n.d.). *Health Data*. County Health Rankings & Roadmaps. Retrieved November 5, 2024, from `https://www.countyhealthrankings.org/health-data`.

[3] United States Census Bureau. (n.d.). *Census.gov*. United States Census Bureau. Retrieved November 5, 2024, from `https://www.census.gov/`.

[4] Santos, E. C., Ramírez-Reyes, C., Müller, C. S., & Torres, B. (2019). A Geographically Weighted Random Forest Approach for Evaluating Forest Change Drivers in the Northern Ecuadorian Amazon. *PLoS ONE, 14*(12), e0226224. Retrieved from `https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0226224`.

[5] Sun, Y., Hu, Y., & Wang, J. (2024). PyGRF: An Improved Python Geographical Random Forest Model and Case Studies in Public Health and Natural Disasters. *Transactions in GIS*. Retrieved from `https://www.acsu.buffalo.edu/~yhu42/papers/2024_TGIS_PyGRF.pdf`.

[6] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*. CRC Press.

[7] UT Southwestern Medical Center. (n.d.). *UT Southwestern Catalog*. UT Southwestern Medical Center. Retrieved November 5, 2024, from `https://www.utsouthwestern.edu/education/utsw-catalog/`.

[8] Wang, Q., Hu, Y., & Li, X. (2022). GWRBoost: A Geographically Weighted Gradient Boosting Method for Explainable Quantification of Spatially-Varying Relationships. *arXiv preprint*. Retrieved from `https://arxiv.org/abs/2212.05814`.