ADTA 5940 Section 003 - Analytics Capstone Experience (Fall 2025 1)

A Project Report on

Analyzing Diabetes Risk Patterns Across U.S. Census Tracts

A Comprehensive Study Using CDC PLACES Data

UNIVERSITY OF NORTH TEXAS

Under the Guidance of

Dr. Denise Philpot

Submitted by

Team N

Shabana Shaik (11766712)

Vishnu Vardhan Reddy Golamari (11682425)

TABLE OF CONTENTS

I. INTRODUCTION

Diabetes mellitus has remained one of the greatest challenges to public health in the USA since the condition currently affects approximately 37.3 million people in the country. That measures 11.3% of the whole population. Additionally, the economic costs of the disease have been estimated to be over $327 billion per year. Most of the targeted populations affected by the disease have type 2 diabetes, a form of the disease that comprises between 90-95% of people who have been diagnosed (Hu et al., 2025).

Substantial evidence also indicates that physical inactivity, smoking exposure, lack of quality sleep, and poor lifestyle habits remain major risk factors for the onset of diabetes via complications of inflammation, impaired glucose uptake, and impaired insulin function (Zhang et al., 2019; Qin et al., 2023; Lu et al., 2021). These habits also vary in different parts of the country. There is evidence presented by previous studies to show the effects of geographical variation in the neighborhood environment of regions to have a great influence on the onset of the disease (Nath & Odoi, 2024; Lord & Odoi, 2024).

The analysis relies on the PLACES dataset developed by the 2023 CDC. This dataset provides information at the census tract level for all 68,172 tracts in the country. Since the PLACES dataset relies more on the behavioral aspects of the communities instead of demographic or socioeconomic factors for analysis, there is a unique opportunity to analyze the influence of modifiable behaviors in the prevalence of Diabetes regardless of the demographic composition of the communities. Interestingly enough, the PLACES dataset lacks demographic factors.

In the context of the analysis of a number of behavior indicators simultaneously, the present research expands the boundaries of existing studies in the field in as much as a number of different components enter into the analysis. These include behavior analysis, spatial analysis, together with predictive modeling. Machine-learning algorithms together with cluster analysis also facilitate the examination of a number of significant indicators of the prevalence of diabetes in order to determine the feasibility of community formation based on their risk indicators.

II. LITERATURE REVIEW

Demographic and Social Determinants of Diabetes

Age represents one of the prominent unmodifiable risk factors for diabetes. Unlike habits that can be altered to some extent to promote positive health effects, age refers to a set of unalterable biological factors. The statistics clearly indicate the gradual rise of the prevalence of the disease proportionate to the advancement of ages. For example, only 4.2% of the members of the 18- to 44-year-old group have been diagnosed with the disease compared to 17.5% in the 45- to 64-year-old group. Interestingly enough, the corresponding figure for the 65+ group has reached 29.2% (Hu et al., 2025). All these factors point to the physiological transformations that occur in the body due to increasing ages. This includes the reduced activity of cells in the pancreas for the production of beta-cells. People also become less active physically. Their eating habits have already developed their own trajectory of raising the risk of developing the disease.

Diabetes prevalence also tends to vary significantly between racial/ethnic groups. This tends to be a function of systemic disparities in health conditions rather than a reflection of genetic variation. The prevalence for African Americans is 12.1%, significantly higher than the 6.9% of Non-Hispanic Whites. Hispanic/Latino prevalence rates at 11.7% fall below the highest-affected groups but remain significantly higher compared to the average. The largest affected groups in the country have a prevalence of 14.5%. Asian Americans also have a higher prevalence of 9.5%, which tends to be about 1.4 times the rates of White communities (Hu et al., 2025). These conditions can be chiefly linked to the social determinants of health. These may range from systemic discrimination that differentially affects the segregation of communities in terms of a healthy food environment in their vicinity.

There is also disparity between genders with respect to diabetes. Men tend to have a slightly higher incidence at 12.1% compared with women at 10.8%; however, women have predisposing factors such as gestational diabetes and Polycystic Ovary Syndrome (PCOS), which predispose them to Type 2 diabetes. It is worth noting that women with diabetes have specific cardiovascular morbidities compared with their male counterparts. Genetics make diabetes risk even more complex. If one parent has diabetes, the risk is increased to around 40%, and with two parents having diabetes, the risk is 70%. More than 400 gene variations leading to diabetes have been discovered; however, this accounts for no more than 10-20% of the overall risk. The dramatic rise in the incidence rate of diabetes cannot be merely due to genetics and thus the triggering factor for diabetes is the environment.

Access to health care is a significant predictor of diabetes outcomes. Some 8.7 million Americans do not know they have diabetes and comprise approximately 23% of all persons with diabetes. In this regard, the majority with undiagnosed diabetes lack health insurance/coverage. In addition to such troubles, rural areas encounter additional health care access constraints due to reduced local health care services from a lack of hospitals and health care professionals. Contrary to reduced diabetes prevalence within said areas, fewer screening tests for diabetes occur. However, Lord and Odoi (2024) proved health care access is an independent predictor for diabetes admissions when controlling for income and health insurance coverage.

A significant factor for diabetes risk is the food environment. The relative risk for diabetes is higher by 32% for food desert communities with limited availability and accessibility of healthy and fairly priced food compared to communities with convenient access to a supermarket. However, food swamps with higher ratios of fast-food establishments compared to grocery stores can be attributed as additional risk factors for diabetes risk. For instance, communities with limited income possess 2.4 more fast-food restaurants and 1.3 fewer supermarkets compared to affluent communities. Sharma (2023) supported that there was a substantial geographical variation for food insecurity and diabetes prevalence, thus validating the link between the local food environment and diabetes risk.

Behavioral Risk Factors

A sedentary lifestyle is one of the biggest and easiest risk factors to reverse for Type 2 diabetes. In 2024, Yang and colleagues found that active people with prediabetes lower their risk of developing diabetes by 44% compared to inactive people with prediabetes. Exercise benefits insulin sensitivity and glucose metabolism and decreases visceral fat - fat known to be closely associated with diabetic metabolism problems. In 2021, Wu and various other scientists found that regions with high obesity and physical inactivity tend to be clustered together and form so-called ‘diabetes hotspots.’ This is evidence that the geographical environment and facilities such as parks and sidewalks influence diabetes risk on a larger community basis. In 2023, Jayedi and various other scientists found that diabetes risk can be forecasted by your walking speed regardless of your exercise time and duration. This means that exercise intensity is as important as exercise duration.

Diabetes risk increases with U-shaped relationship length of sleep duration. There are two groups for this increase in diabetes risk. One group is people who do not get enough sleep (less than 6 hours), and the other group is people who sleep too much (more than 9 hours). According to Lu et al. (2021), both insufficient and excessive sleep have a higher risk for developing diabetes than the average amount of sleep of 7 to 8 hours Therefore, there may be a need for additional research to identify the pathways that these paths share in common and to understand how they differ from each other. Zhou and Tian (2024) found that short lengths of sleep predict future incidence of diabetes regardless of the activity level. So therefore, short lengths of sleep and activity levels are likely to represent separate pathways for diabetes. Silva et al. (2024) said that while there is a relationship between the length of sleep and negative consequences on one's overall health, it still is unknown whether an increased length of sleep actually causes the negative metabolic effects in people, or if the increased length of sleep reflects an underlying health issue.

Behavioral Risk Factors for Diabetes Among Smokers and Non-smokers. Another major behavioral risk factor for developing diabetes is smoking. According to a report published by Qin and colleagues in 2023, the risk of developing diabetes is increased by 22% for people who are exposed to secondhand smoke. In addition to increasing the risk of being diagnosed with diabetes, smoking causes chronic inflammation, damages beta cells that produce insulin, which are responsible for secreting and releasing insulin molecules into the bloodstream, and impairs the ability of blood vessels to work properly. Thus, the negative effects of smoking affect both the active smoker as well as those who are exposed to tobacco smoke in their environment. Therefore, it can be concluded that smoking is a significant risk factor for diabetes at both the individual level and the community level.

As discussed previously in this article, behavioral risk factors "tend to cluster" with one another, thus magnifying the effect of all behavioral risk factors on the likelihood of developing diabetes. In a study done by Zhang and colleagues in 2019, they demonstrated that people who participate in multiple healthy behaviors, such as exercise, eating a healthy diet, not smoking, and drinking moderate amounts of alcohol, had a much lower risk of diabetes than those who established only one healthy behavior. Zhu et al. (2023) reported that if a person with diabetes replaces one hour of sedentary time with light physical activity, the risk of death is greatly decreased. Additionally, Deng and colleagues (2022) conducted research using Mendelian randomization to obtain genetic evidence that supports the conclusion that there is a causal relation between sedentary behavior and diabetes risk. Thus, it supports the rationale for implementing behavioral interventions into the prevention of diabetes.

Machine Learning and Geographic Analysis

Recent developments in machine learning have allowed for improved prediction accuracy of Diabetes via methods utilized in the past, thus Relying on statistically based models (Traditional Methods) versus machine learning approach; In a study done by Chou et al (2023) it was found through Collaborative methods (Ensemble) like "Random Forest and Gradient Boosting" to identify the complexity/nonlinearities of relationships between predictors that would have been difficult or impossible to identify using Linear Model methods. Shin et al (2022) found an Improvement in Diabetes prediction accuracy of a Tree-based model (Decision Tree) compared to a Logistic Regression Model. In addition, Fu et al (2023) demonstrated using Machine Learning techniques to detect subtle but useful results through Clinical Data when using Classical Methods vs Machine Learning techniques. However, most of the studies that have examined Machine Learning in the field of Diabetes have been limited to Individual Clinical based approaches rather than using Population Geographic Analysis (Community). Therefore, it will be necessary to have improved Models within a Community to identify Neighborhood areas that may require additional Prevention Resources.

Geographic analysis identifies considerable spatial variability in diabetes occurrence beyond accounting for factors that affect individuals. Quiñones et al. (2021) found diabetes risk factors to be geographically dispersed, indicating that a single nationwide diabetes intervention may not reach all at-risk individuals across the country. Uddin et al. (2022) demonstrated that neighborhood characteristic variables (e.g., poverty rate, percent Black) predicted greater rates of diabetes occurrence beyond the individual behavior variables; therefore, one must consider the influence of the social structure on diabetes risk rather than the personal choices made by the individual. Wittman et al. (2024) indicated that tracking high concentration areas of diabetes may enable more effective distribution of resources related to diabetes prevention programs. Benavidez et al. (2024) reported extreme variability in chronic disease rates across different ZIP codes, suggesting that even in small geographic areas, there are significant differences.

According to Nath & Odoi (2024) geographic disparities in diabetes-related deaths continue to exist and have increased in amount over time as it appears, that statewide initiatives do not appear to be utilized equally by all communities. Thus, researchers recommend the use of spatial targeting for intervention development, focusing on local area conditions rather than simply providing general state or population-wide level approaches.

Research Gaps and Study Contributions

Although we have made great strides in figuring out the factors that lead to diabetes, we still know little about how individual/behavioral factors interact with geographic/predictive factors on a community level. Current machine learning research has examined individual clinical data rather than the population-level indicators of community/behavioral factors. The CDC PLACES dataset fills this knowledge gap by providing community/census tract level estimates of behaviors across 68,172 census tracts in the US. What PLACES is missing, however, is a comprehensive representation of the surrounding community demographics, such as age, race, income and education, as well as the ability to measure environmental factors, such as food availability and accessibility, as well as other access to health services.

The application of Random Forest regression and K-means clustering to census tract-level behavioral data adds to the existing literature on the assessment of community risk profiles and the measurement of the strength of nonlinear correlations between behavioral characteristics and health-related characteristics through a predictive modeling approach (i.e., Random Forest). The absence of any demographic or socioeconomic data in the predictive modeling, however, means that efforts to use this modeling for resource allocation will be limited due to the inability to differentiate between the variance in our study and that of the unobserved structural determinants associated with the community. Future studies should consider combining behavioral characteristics with demographic characteristics and/or access to environmental and/or health care resources in order to provide more precise attribution of community risk factors to either modifiable individual behavior or structural inequity.

III. RESEARCH QUESTIONS

RQ1: What are the relationships between the lifestyle determinants (such as smoking, physical inactivity, poor sleep, and binge drinking) and diabetes prevalence rates within census tracts?

This exploratory question explores the correlation between various key risk factors and diabetes prevalence among 68,172 census tracts. Based on correlation analysis and data explorations, various behaviors that show strong correlation to diabetes prevalence are identified. The results are then further verified to check if there is uniform correlation with various geographical areas. Based on exploratory analysis, key risk factors that are most influential in diabetes prevalence are identified.

RQ2: Can machine learning models accurately predict diabetes prevalence with accuracy using lifestyle factors such as physical inactivity, smoking, lack of sleep, and binge drinking?

Starting with the results derived from the relations identified, this question was designed to assess the predictive accuracy of four machine learning models (Linear Regression Model, Decision Tree Model, Random Forest Model, and Gradient Boosting Model) to predict diabetes prevalence rates in the census tracts. The models are tested using R-Squared values to determine how well the models fit the data, Adjusted R-Squared to see how well the models fit and take data complexity into account, RMSE to check prediction accuracy, and MAE to check average errors. The analysis provides the best algorithm to use, determines the importance ranking of various features, and examines possible thresholds that indicate significant risk factors to diabetes.

RQ3: Which states and counties experience highest diabetes prevalence, and what are the geographic distributions by region?

This question highlights regional variations in diabetes prevalence by systematically ranking the prevalence rates among the states and counties. The findings illustrate regional clustering mainly among Southern states and Appalachia, with higher diabetes prevalence rates that are more than double the national average. The regional variations are mainly due to underlying structural challenges such as poverty rates, poor healthcare accessibility, and cultural factors rather than individual factors informing targeted prevention program implementation and healthcare resource allocation.

RQ4: Can areas with similar health patterns be grouped together and what are the factors that distinguish these groups?

The application of k-means clustering analysis on these four variables (physically inactivity, smoking, lack of sleep, binge drinking) and diabetes prevalence divides the 68,172 census tracts into various groups based on risk factors. These clusters are comprised of groups with similar health factors but driven by different factors. The reason different groups are addressed differently is that those with higher rates of inactivity need to be treated differently from those with higher smoking rates or sleep<7 hours. Creating these clusters allows health departments to develop community-specific, data-based public health interventions, as opposed to a "one-size-fits-all" approach.

IV. METHODS

Data Overview and Preparation

The CDC PLACES dataset is a collection of chronic disease risk factor data for small areas using multilevel regression and poststratification modeling methods on data from the Behavioral Risk Factor Surveillance System (BRFSS) and U.S. Census population estimates. At the census tract level, it includes data for 68,172 tracts across all 50 states and Washington D.C., with an average population of 3,965 people per tract.

The PLACES dataset is limited to the study of behavioral health measures only and does not include the following demographic information at the census tract level: age, race/ethnicity, income, education. There are no estimates of healthcare access or environmental variables, such as food environment quality. While this limited scope enables researchers to look at potential modifiable behavioral risk factors, it also means researchers cannot account for demographic or structural confounders. The extent to which this limitation affects analysis is discussed in the discussion section.

Variable Selection

Our selection of variables was systematic and data-driven: we listed all the available Health Risk Behaviors in the PLACES dataset; further, we conducted a correlation analysis between all the risk factors and the prevalence of diabetes. Variables were selected based on the strength of correlation, biological plausibility, and completeness of data.

Dependent Variable: Diabetes Prevalence

Model-based estimates of diagnosed diabetes prevalence among adults aged 18 years and over. The variable was chosen because diabetes prevalence has been shown to average 10.9% among census tracts, ranging from 0.7% to 46.1%, making it one of the most significant challenges within the public health sector.

Independent Variables:

In this process, correlation analysis through the use of the EDA heatmap identified that there are four Behavioral Risk Factors that have the strongest correlation with diabetes:

1. Physical inactivity (r = 0.86) is the number of people aged 18 and older who say they did not do any fun physical activity in the last month. This measure gets picked because it connects to diabetes. Inactivity can affect how the body uses glucose. It can also affect how insulin works fast. Regular activity can lower the risk by 30 to 50 percent (Yang et al., 2024).

2. Current smoking (r = 0.73) is the number of adults aged 18 and older who smoke cigarettes now. It is the second strongest link that we find. Smoking can harm beta-cells in the pancreas. This causes a 30 to 40 percent higher risk of diabetes because of inflammation and oxidative stress (Lu et al., 2021).

3. Sleep less than 7 hours (r = 0.70) is the number of adults aged 18 and older who say they sleep less than seven hours a day. This measure has a strong link in the study. Sleep problems can affect important hormones. These hormones are cortisol, leptin, and ghrelin (Qin et al., 2023).

4. Binge drinking (r = -0.70) is the number of adults aged 18 and older who said they drank five or more drinks for males or four or more drinks for females in one occasion in the last 30 days. A strong opposite link was found that needs more exploration. This surprising finding shows the "sick quitter" effect where people with diabetes reduce how much alcohol they drink (Zhang et al., 2019).

Other variables were excluded because they did not represent modifiable Health Risk Behaviors, demographic variables cannot be modified through intervention, and these four captured 81.5% of variance in our predictive model. The selection process thus focused attention on factors that are actionable, evidence-based risk factors that have strong statistical associations with diabetes prevalence.

Exploratory Data Analysis

Data Cleaning Process

In the initial dataset, there were 2,555,113 rows encompassing various measures of health by census tract. The dataset was cleaned as follows:

- The dataset was filtered to only include diabetes and the four predictors,
- One row was removed due to missing diabetes prevalence information.
- We eliminated 10 columns of redundant data from the analysis, including the Year, DataSource and footnotes.
- The LocationID for the census tract is standardized to ensure that each census tract has an unambiguous geographic identifier.
- The final analytical dataset contains 68,172 census tracts that have complete data available to them.

Descriptive Statistics

Variable	**Mean (%)**	**Std Dev**	**Min**	**Max**	**Correlation with Diabetes**
Diabetes	10.9	3.7	0.7	46.1	1.00
Physical Inactivity	23.1	7.2	2.8	68.9	0.86
Current Smoking	16.8	5.9	1.1	57.4	0.73
Sleep <7 Hours	33.4	5.4	14.1	61.2	0.7
Binge Drinking	15.9	4.3	2.3	41.8	-0.70

Table*: Summary Statistics for Key Variables*

Distribution Analysis

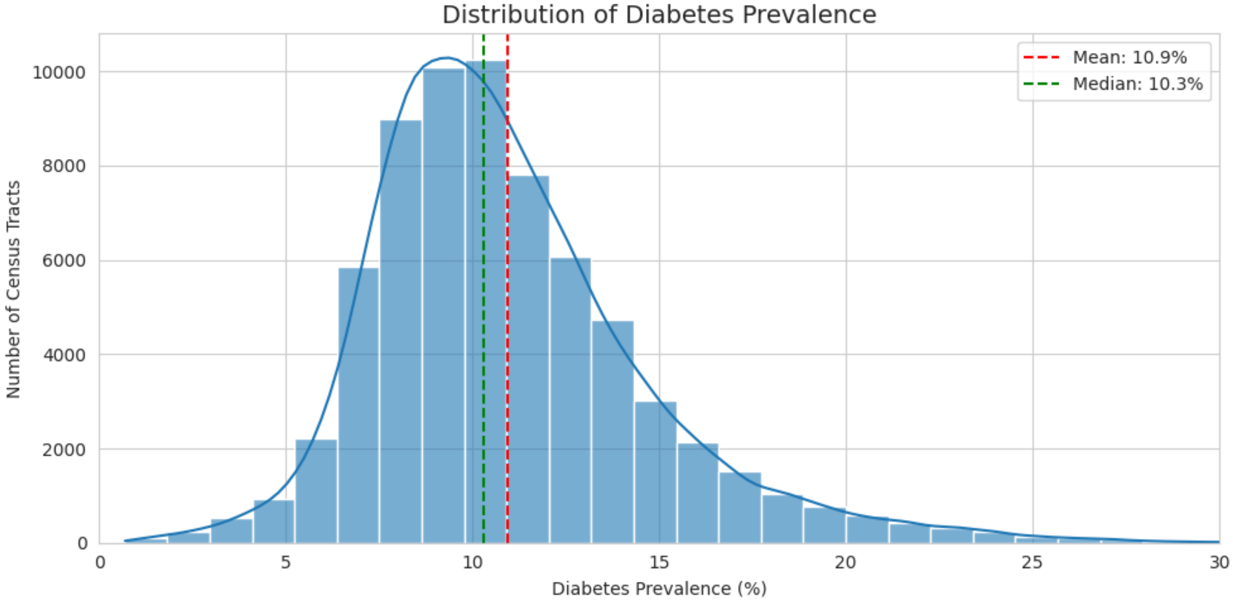
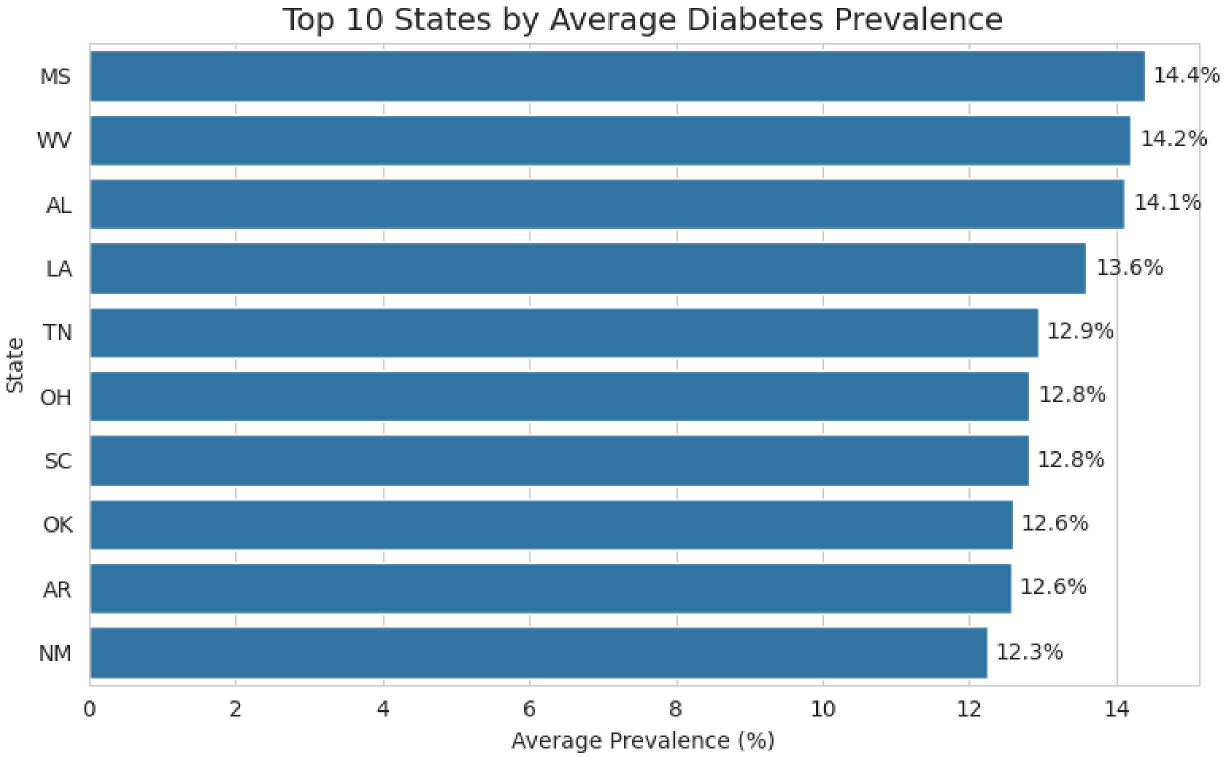


Figure*: Distribution of Diabetes Prevalence*

The histogram shows the frequency distribution of diabetes cases by census tract so we can evaluate whether our data are normally distributed thus suitable for regression analysis. The frequency distribution shows a positive skew, with a mean of 10.9% and a median of 10.3%. Most census tracts lie in the range of 8% - 13% of the total number of census tracts that were analyzed. However, there is a substantial long positive tail of the frequency distribution showing numbers as high as 46.1%. The interquartile range (IQR) method identifies 96.5% of the census tracts in the range of 1.8% - 19.4% to be normative or "one of the usual" values, demonstrating that 3.46% of the census tracts analyzed are statistical outliers. The right skew of the frequency distribution demonstrates that most communities throughout America have moderate levels of diabetes prevalence, while some census tract regions are experiencing disproportionately high levels of diabetes prevalence and thus need focused interventions.

Geographic Patterns



[Figure: Top States by Diabetes Prevalence]

Most states have similar diabetes outcomes within their borders, due to geographical similarities. Of all the states, Mississippi was the highest state for people who had been diagnosed with diabetes at 14.4% followed closely by West Virginia at 14.2%, Alabama at 14.1%, Louisiana at 13.6%, Tennessee at 12.9%, while

Colorado, Utah, Vermont, Montana, and Alaska had the lowest rates of diabetes at 7.1%, 7.8%, 8.0%, 8.2%, and 8.3%. The state with the highest prevalence of diabetes is approximately double the rate of the state with the lowest, demonstrating a strong correlation between the "Diabetes Belt" and poverty, health care infrastructure, and built environments.

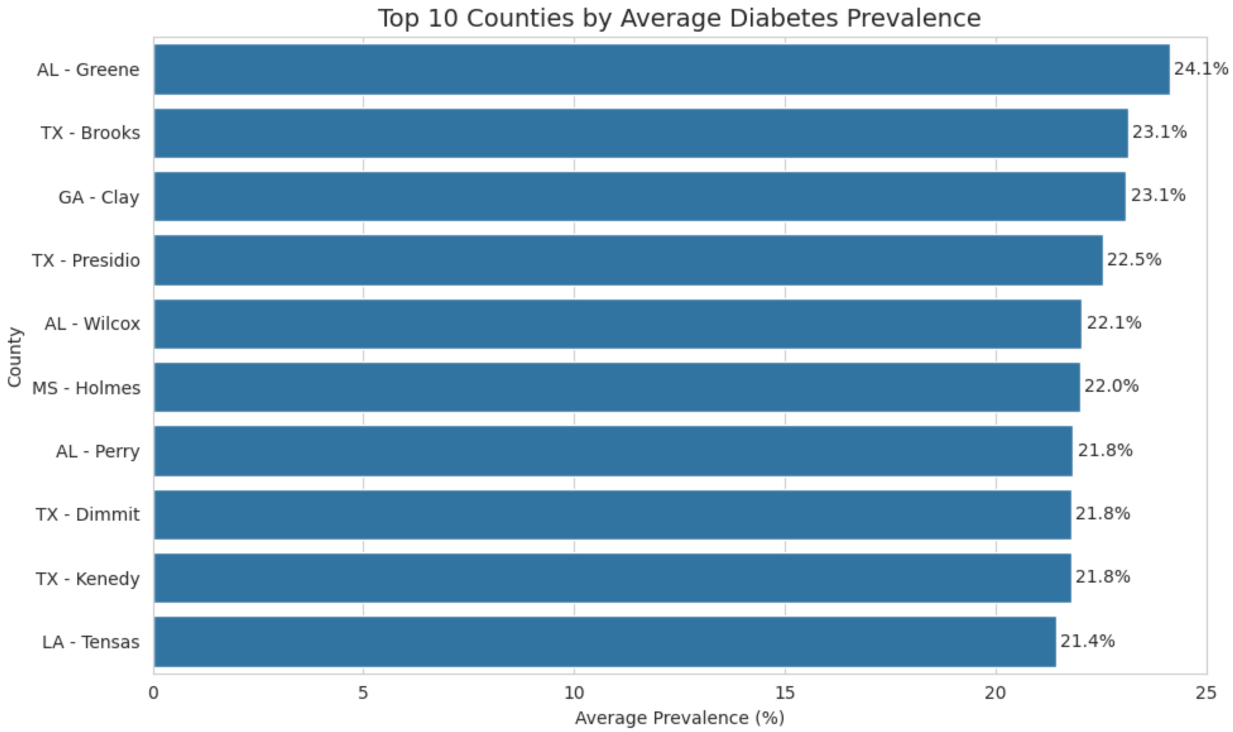


Figure*: Top Counties by Diabetes Prevalence*

The county-level analysis shows part of the higher level of risk in these counties may be hidden in the state average. The counties with the highest indicated rates (greater than twice the national average) are Greene County, Alabama, (24.1%); Brooks County, Texas (23.2%); and Clay County, Georgia (23.1%).

The counties included have small populations, which makes it easier to get a high percentage diagnosed with diabetes in those areas compared to large-population counties where there are relatively few people being diagnosed with diabetes. This means that while the percentage of diagnosed with diabetes is high for the counties included, we remain concerned; therefore, caution should be taken when comparing large-population counties and small-population counties due to differences in population size. The high number of people diagnosed with diabetes living in rural counties where they may have little or no access to healthcare and limited physical activity is likely influenced by factors beyond the individual level.

Risk Factor Correlation Analysis

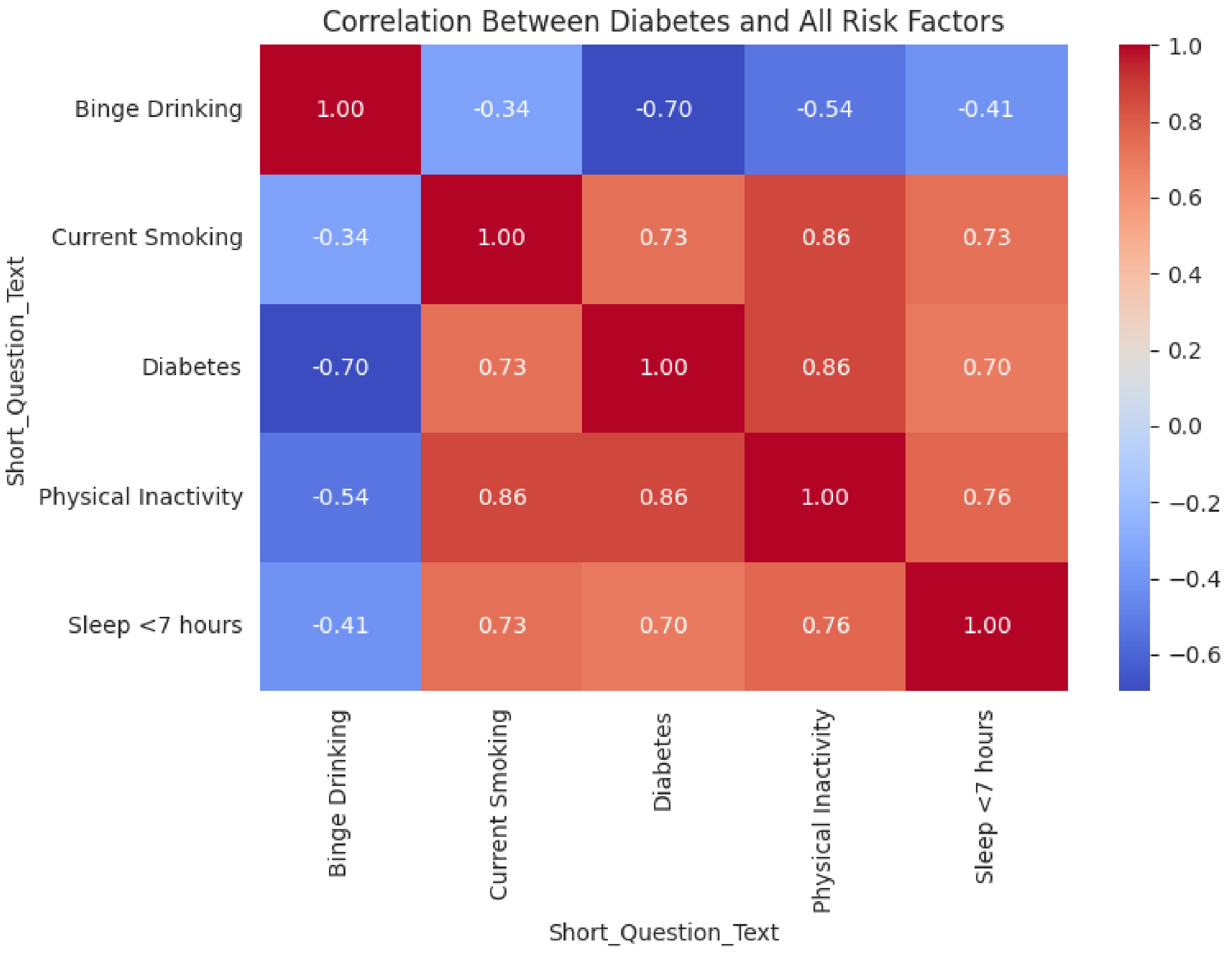


***Figure**: Correlation Heatmap*

The correlation heatmap offers an extensive overview of all correlation values in one glance, making it easy to identify correlation values among variables and check for multicollinearity among predictors to be used in regression analysis. The correlation heatmap shows that physical inactivity has the strongest correlation with diabetes with a correlation coefficient of 0.86, followed by current smoking with 0.73 and then lack of sleep with 0.70. The most conspicuous area on the correlation heatmap is the negative correlation with binge drinking with values -0.70. Notably, the correlation heatmap shows that there are no severe multicollinearity among the predictors since the greatest correlation among them is 0.76 between physical inactivity and current smoking.

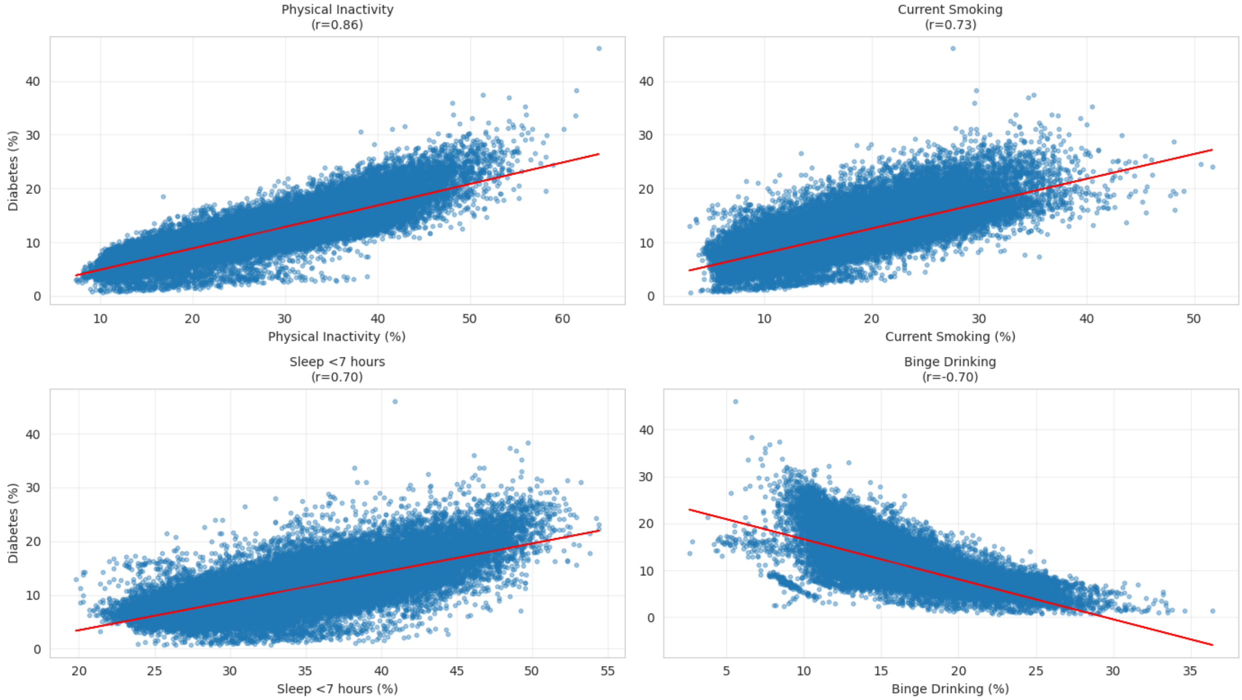


***Figure**: Scatter Plot Matrix (Diabetes vs Risk Factors)*

The scatter plot matrix with regression lines was created to validate the assumptions to be satisfied before implementing regression analysis to identify possible outliers and non-linear relationships. Inactivity shows the strongest linear trend with the lowest scatter about the regression line, again confirming it was the strongest predictor. Smoking now shows a linear trend that is strongly positively correlated however there is increased scatter about the higher values which could infer heteroscedasticity. Sleep deprivation shows a moderately strong linear trend with evidence of possible clustering. The scatter plot for binge drinking clearly shows the negative slope, again emphasizing the opposite trend to the correlation analysis. All three show that there are no considerable non-linear trends that would infer transformation is required to allow for linear modeling approaches.

Predictive Modeling: Methodology

Model Selection

A total of four supervised machine learning models were used for predicting the prevalence of diabetes:

1. Linear Regression: Used as a baseline model assuming linear relationships between predictors and outcome.
2. Decision Tree (max_depth=10): A non-parametric algorithm that can identify nonlinear relationships and interactions between features.
3. Random Forest (n_estimators=100, max_depth=15): It is a type of ensemble method wherein multiple decision trees are created and their forecasts are then averaged out. It prevents overfitting.
4. Gradient Boosting (n_estimators=100, max_depth=5): This is a sequential form of Boosting that focuses on correcting mistakes made by previous models to reach higher precision on well-structured datasets.

Model Configuration

- Train/test split: 80% training (54,537 tracts), 20% testing (13,635 tracts)
- Random state: 42 (ensuring reproducibility)
- Evaluation metrics: R^2, Adjusted R^2, Root Mean Square Error (RMSE), Mean Absolute Error (MAE)
- Cross-validation is not used when the sample is large and therefore gives a good representation of the variability that exists within the population.

Model Evaluation

To determine its applicability, the predicted prevalence rate for diabetes was then categorized on the basis of four risk categories as follows:

- Low Risk: Below Q1 (8.4%)
- Moderate Risk: Q1 to Q2 (8.4% – 10.3%)
- High Risk: Q2 to Q3 (10.3% – 12.8%)
- Very High Risk: Above Q3 (12.8%)

On classification evaluation, Confusion Matrices, Accuracy, Precision, Recall, and F1-Score were employed with emphasis on balanced classification performances across levels of risk.

Cluster analysis: Methodology

K-means clustering was utilized to identify categories of census tracts with similar profiles for behavioral health.

Optimal Cluster Selection

Elbow method was employed to find the optimal number for the number of clusters at which the rate of reduction of within-cluster variability (inertia) became smaller once more clusters were added. The values for inertia were: 147,187 for k=2, 113,122 for k=3, 93,132 for k=4 and then finally reduced to 80,694 when k=5. A relative reduction from k-3 to k-4 is much sharper (by 23%) than from k-4 to k-5 (by 13%), so the elbow point is at k-4. This is consistent with typical risk categorizations for public health (Low, Moderate, High and Very High).

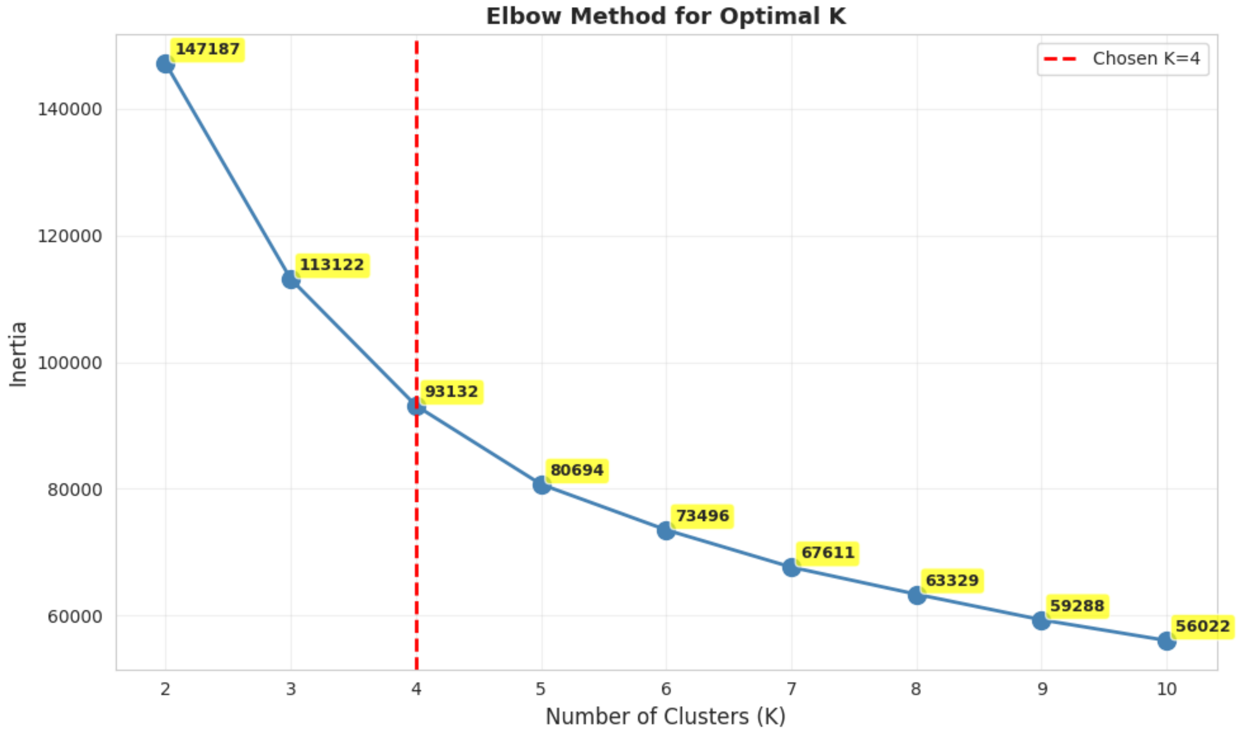


***Figure**: Elbow Plot for K-means Clustering*

Implementation

1. Feature Standardization: For feature standardization, the StandardScaler from scikit-learn is employed so that all features can be scaled equally.
2. Optimal K Selection: We tested k from 2 to 10 for optimal k-value selection using the elbow method.
3. Final Model: Elbow analysis showed that the number of clusters is K = 4.
4. Cluster Labeling: We assign a label to the clusters based on the ordered mean diabetes prevalence: Low, Moderate, High, and Very High Risk.

Clustering Parameters

- Algorithm: K-means

- n_clusters: 4
- n_init: 10 (number of initializations to ensure stable results)
- random_state: 42 (reproducibility)

V. RESULTS

<u>Predictive Modeling: Results</u>

Model Performance Comparison

The performances of four machine learning algorithms were assessed for predicting diabetes prevalence at the census-tract level. On this task, the Random Forest model was found to be the best-performing one. It produced R-squared of approximately 0.8728. This means that approximately 87.3% variation was explained by the model.

Model	**R²**	**Adjusted R²**	**RMSE**	**MAE**
Random Forest	0.8728	0.8728	1.326	0.962
Gradient Boosting	0.8688	0.8688	1.347	0.992
Decision Tree	0.8528	0.8528	1.426	1.027
Linear Regression	0.8314	0.8313	1.527	1.123

Table*: Predictive Model Performance Metrics*

The model with the lowest RMSE was the Random Forest model with a value of 1.326 percentage points and MAE of 0.962 percentage points, which means that on average the deviation between the model's forecast and diabetes prevalence is less than one percentage point. This is expected given the complex correlations between the risk factors and diabetes prevalence that cannot be predicted by linear models. Both ensemble methods (Random Forest and Gradient Boosting) performed better than other models.

The similarity between R² and Adjusted R² for all models is a function of the sample size (68,172 census tracts). Looking at large datasets, the second part of the formula for the Adj (R-squared) becomes irrelevant for additional predictors. This explains why R² and Adjusted R² equalize for large datasets. It is expected behavior for a sample size this large.

Classification Performance

To estimate the ability of the approach to facilitate resource allocation, we categorized the forecasts for ongoing prevalences within four risk levels with respect to the quartiles from the training data (Q1 = 8.4%, Q2 = 10.3%, Q3 = 12.8%). The results of the Random Forest classifier were as follows:

- Overall accuracy: 71.8%
- Weighted precision: 72.3%
- Weighted recall: 71.8%
- F1-score: 72.0%

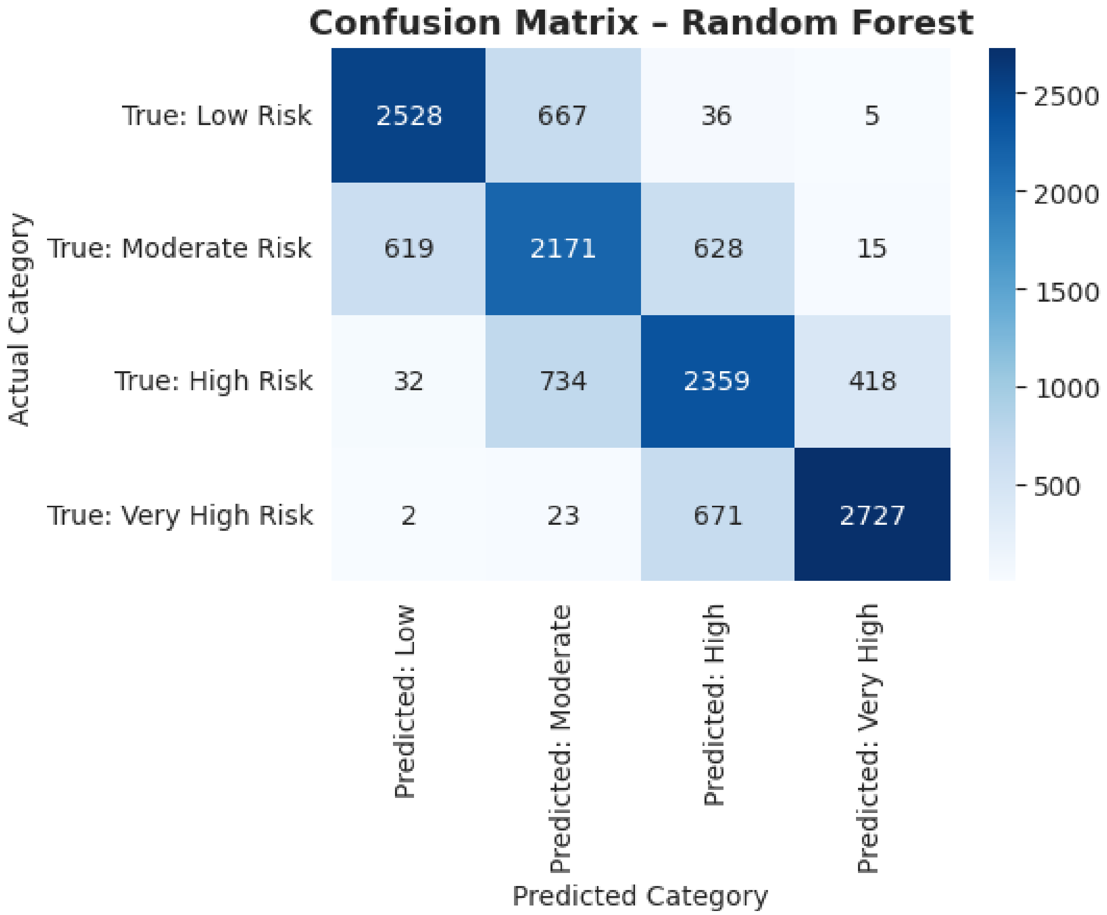


***Figure**: Confusion Matrix for Random Forest Risk Classification*

Classification performance was strongest at the extremes of the risk distribution:

Risk Category	Precision	Recall	Support (n)
Very High Risk	86.2%	79.7%	3,423
Low Risk	79.5%	78.1%	3,236
High Risk	63.9%	66.6%	3,543
Moderate Risk	60.4%	63.2%	3,433

It successfully predicted and clustered 2,727 out of 3,423 tracts belonging to Very High Risk and 2,528 out of 3,236 Low Risk tracts. The model made mistakes mostly within consecutive categories and not between the highest and lowest categories. This is satisfactory for real-world uses because consecutive risk tracts can be handled under equal treatment methods. The model made more mistakes for Moderate and High-Risk tracts as compared to other tracts due to overlapping behavioral characteristics found between tracts close to boundaries.

Regression Performance:	**Value**
R²	0.8728
RMSE	1.326
MAE	0.962
Classification Performance:	**Value**
Overall Accuracy	71.80%
Weighted Precision	72.30%
Weighted Recall	71.80%
F1-Score	72.00%

***Table**: Model Performance Summary*

Feature Importance Analysis

The feature importances from the Random Forest model demonstrate a strong ranking among the four behavior predictors.

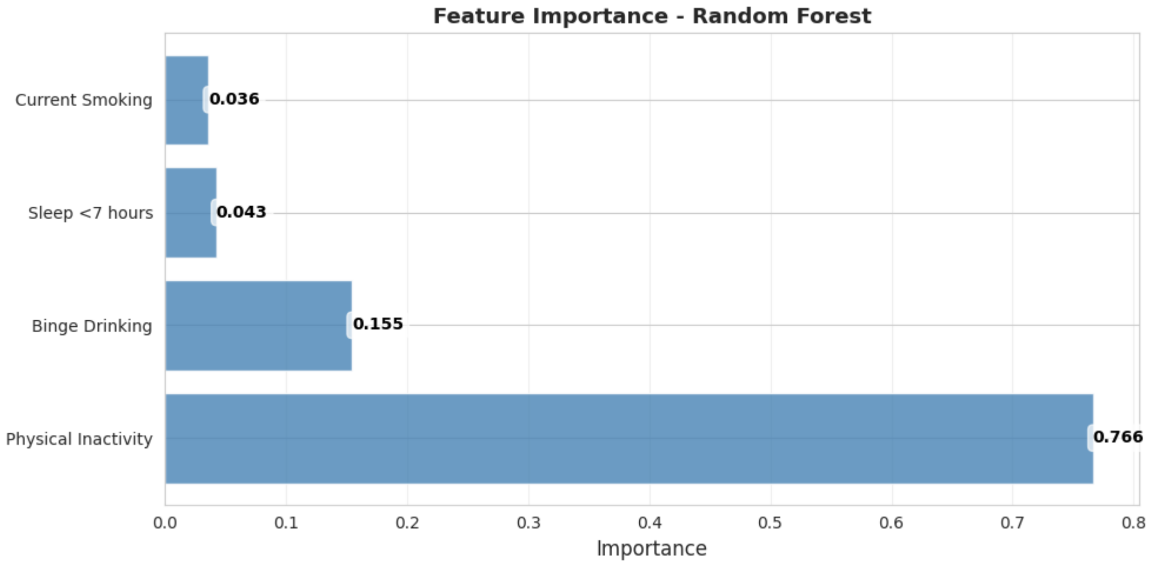


***Figure**: Feature Importance from the Random Forest Model*

Feature	**Importance**
Physical Inactivity	76.6%
Binge Drinking	15.5%
Sleep <7 Hours	4.3%
Current Smoking	3.6%

Physical Inactivity is the leading predictor with 76.6% total feature importance much higher than would be expected given its correlation coefficient (r = 0.86). This result shows that inactivity is a major contributing cause for diabetes risk within the community and may be having an influence on other behavior-related factors.

Binge drinking was the second (15.5%) factor even though it is negatively correlated with diabetes, indicating that either the model is detecting non-linear relationships well or binge drinking is acting as a proxy for unknown socioeconomic factors. Inadequate sleep (4.3%) and smoking (3.6%) were contributing less than would be expected given their correlations with diabetes (r = 0.70 and r = 0.73, respectively), which may be measuring similar variance as physical inactivity or may lack discrimination on the population level.

Risk Thresholds and Decision Points

Cluster and model analysis identify dominant predictive features and critical thresholds above which diabetes incidence rises steeply. Decision thresholds represent the point at which communities fall within a particular risk category:

Physical Inactivity Level	**Diabetes Prevalence**	**Risk Category**	**Interpretation**
<18%	~7.3%	Low Risk	Baseline community health
18–20%	~9.2%	Moderate Risk	Transition zone
~20%	**~10%**	**Critical Threshold**	**Tipping point**
20–28%	~11.5%	High Risk	Elevated risk
>28%	>12%	Very High Risk	Intervention priority
>38%	>16%	Extreme Risk	Urgent intervention needed

The point of critical inflection is at around 20% physical inactivity. Below this point, communities tend to keep diabetes prevalence below 10%, and above this point, communities tend to see the rate of diabetes rise steeply. The gradient is steepest between the Moderate Risk (18.8% inactivity and 9.2% diabetes) and High Risk (27.7% inactivity and 11.5% diabetes) groups.

For other behavioral factors, the following thresholds emerged:

Risk Factor	**Low Risk Threshold**	**High Risk Threshold**
Physical Inactivity	<18%	>27%

Current Smoking	<13%	>18%
Sleep <7 Hours	<31%	>35%
Binge Drinking	>20% (inverse)	<14% (inverse)

Before these thresholds are established, there should be a clear methodology to determine areas of community need or intended goals for preventative strategies.

Model Diagnostics

Systematic errors do not exist in the Random Forest method as proven through a detailed residual analysis.

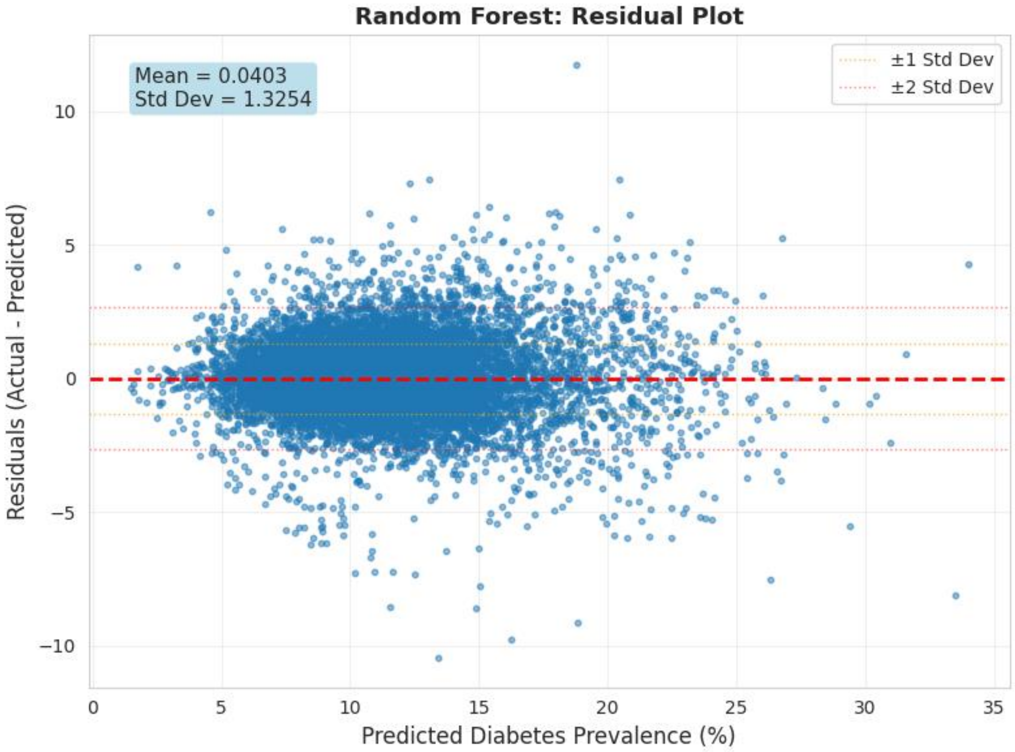


***Figure**: Residual Plot for Random Forest Model*

Residual Statistic	**Value**
Mean	0.0403

Standard Deviation	1.3254

The randomness of the residuals around zero suggests that systematic prediction error does not exist with this model. All but a few of the residuals compared to the predicted data exist within ±2 standard deviations from the mean predicted value, although there is a small amount of heteroskedasticity present in the model at higher predicted values. The lack of identifying patterns indicates that this model will be able to generalize effectively across all census tracts (unit of geography) characteristics.

<u>Clustering Analysis Results</u>

Cluster Profiles

K-means Clustering (k = 4) generated four unique classifications of risk for census tracts based primarily on their respective levels of physical inactivity and incidence of Diabetes.

Cluster	**Tracts (n)**	**% of Total**	**Diabetes (%)**	**Physical Inactivity (%)**	**Smoking (%)**	**Sleep <7h (%)**	**Binge Drinking (%)**
Very High Risk	10,717	15.7%	16.8	38.2	25.7	41.2	13.6
High Risk	26,607	39.0%	11.5	27.7	18.2	35.2	16.3
Moderate Risk	17,863	26.2%	9.2	18.8	11.1	30.2	15.8
Low Risk	12,985	19.0%	7.3	18.1	12.7	30.6	21.0

***Table**: Health Risk Profiles by Cluster*

The largest classification representing 39.0% (26,607) of all census tract is identified as High Risk, suggesting that there is a wide scope of high-risk people for Diabetes in America and that most communities contain many people at moderate to high risk for developing Diabetes. The Very High-Risk census tracts, comprising 15.7% of the data, had rates of all four behavioral metrics nearly double than the female population rates measured except for binge drinking. Of the Very High-Risk communities, 38.2% were identified as Physically Inactive and 41.2% reported getting insufficient amounts of Sleep.

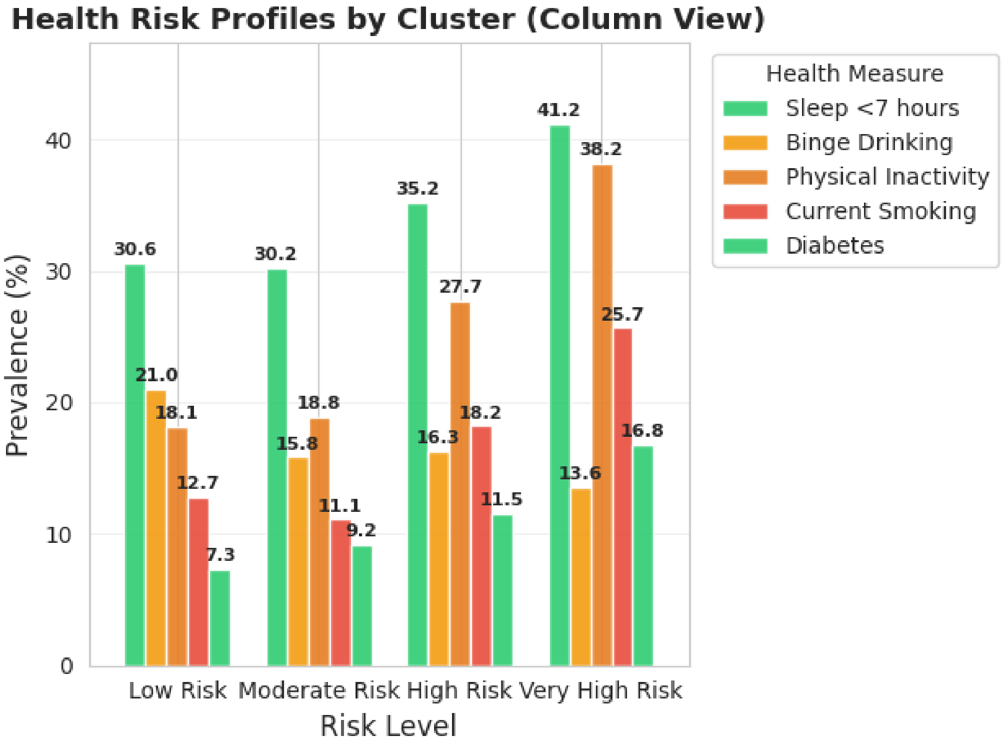


***Figure**: Health Risk Profiles by Cluster*

The results from cluster visualization include good overall gradients across all risk categories. The variable for physical inactivity has the steepest gradient, as its values increase from an 18.1% (low risk) prevalence to a 38.2% (very high risk) prevalence. This result supports the conclusion that physical inactivity is the most important variable in the predictive/modeling process. In addition, the cluster representing Low-Risk status has the highest prevalence of binge drinking (21.0%) but the lowest prevalence of diabetes (7.3%). This observation supports the conclusion that the negative association between binge drinking and diabetes is a result of unmeasured confounding variables rather than a protective effect.

Geographic Distribution of Clusters

Very High-Risk clusters concentrate heavily in Southern states:

State	**% of Tracts in Very High Risk**
Mississippi	42%
Alabama	38%
Louisiana	35%

By contrast, Colorado, Vermont, and Utah have less than five percent of their tracts classified as Very High Risk for diabetes. This clustering of the states geographically demonstrates that there are state-level patterns of prevalence and supports that there are multiple regional structural factors such as poverty, health care infrastructure, and built environment that have an impact on diabetes risk at the regional level in addition to individual behaviors.

VI. DISCUSSION

This examination of 68,172 US census tracts has shown the most dominant area of community-level diabetes prevalence is physical inactivity, comprising 76.6% of the total feature importance in the Random Forest model (R^2=0.8728). This is significantly greater than the contributions of correlation analyses alone and confirms that physical inactivity is a fundamental gateway factor in population-level diabetes risk. As such, it is reasonable to consider that promoting physical activity will yield greater results for diabetes prevention than dispersing resources amongst many different behavioral preventative measures.

Risk Distribution Across Communities:

Based on the clustering analysis, four different community risk profiles were identified, each with a distinct gradient in the Prevalence of Diabetes. These are Low Risk (7.3%), Moderate Risk (9.2%), High Risk (11.5%%), and Very High Risk (16.8%). A very important finding is that 39% (26,607) of all census tracts are in the High-Risk category, which means that there are a lot more than just extreme cases of diabetes risk, but also a large population of diabetes risk across the majority of American communities. The fact that so many areas are classified as High-Risk demonstrates the need for interventions focused on populations rather than individuals who fall into the most extreme categories.

Critical Threshold at 20% Physical Inactivity:

Based on an analysis of several clusters, including all the communities included in the analysis of cluster profiles, there is evidence of a tipping point for physical inactivity around the point at which a population is physically inactive more than 20% of the time. The clusters with a proportion of physically inactive individuals less than 20% are found to have less than 10% diabetes prevalence; however, if greater than 20%, the rate of diabetes rises sharply. There is a very steep gradient observed between Moderate and High-Risk communities (18.8% → 27.7% inactivity/9.2% → 11.5% diabetes) on the diabetes prevalence curve, indicating that a 47% increase in the proportion of physically inactive individuals among Moderate Risk communities equates with approximately a 25% increase in diabetes prevalence when outlooked on this curve. This information provides an idea of a threshold from which to set intervention types, and those communities that reduce community-level rates of physical inactivity to less than 20% may experience greater changes in diabetes prevention than would otherwise.

Geographic Disparities and Structural Factors:

Even when controlling behavioral factors, geographic clustering of diabetes risk remains high. Mississippi has the highest state-wide prevalence (14.4%) whereas Colorado has the lowest (7.1%), which means that there is more than a two-fold difference. At the county level, there are significant concentrations of diabetes including Greene County, AL (24.1%), Brooks County, TX (23.2%), and Clay County, GA (23.1%). All three exceed twice the national average. However, it should be noted that while these counties have a high percentage prevalence of diabetes, they also have relatively small populations, so a higher percentage may simply reflect the statistical effect of having a small denominator that helps to skew the data rather than an increased overall disease burden. Nevertheless, these communities should be noted for their high concentration of need.

Although the ongoing geographic clustering of Very High-Risk tracts is evident in Mississippi (42%), Alabama (38%), and Louisiana (35%), it also indicates that structural influences on poverty are a major contributor to the variance in risk between these regions. Further strengthening this assertion is the fact that the Behavioral factors alone are able to capture only 87.3% of the total risk variance while the remaining 12.7% is most likely attributable to factors not measured in this study, such as poverty, healthcare infrastructure, food environment quality, and built environment characteristics, all of which were not included in the PLACES database.

The Binge Drinking Paradox:

One problem arises from the negative correlation between binge drinking and diabetes (r = - 0.70). Clusters with low risk for developing diabetes showed the highest prevalence of binge drinking (21.0%) while having the lowest prevalence of diabetes (7.3%). This suggests that consuming alcohol does not seem to provide protection from developing diabetes, as many researchers have documented. However, this finding points to the possibility that either (1) the sick quitter effect may occur when people diagnosed with diabetes reduce their alcohol intake on medical advice, (2) the relationship is confounded by unmeasured socioeconomic status, where higher-income areas tend to have higher rates of alcohol consumption along with lower rates of diabetes because of other protective factors or (3) the relationship exists as a result of cultural variations based upon population characteristics that cannot be explained by behavioral variables alone. Therefore, this finding demonstrates that negative correlations should not be interpreted in a simplistic manner and emphasize the need for including both demographic and socioeconomic characteristics in future studies that use this data.

Implications for Public Health Interventions:

The merging of findings from predictive modeling and clustering offers a clear path forward for future diabetes prevention networks. Three major implications can be drawn from this work:

The primary target for improving interventions around diabetes prevention should be physical activity. A full 7.6% of prediction importance is derived from physical activity and has the greatest slope difference across all 4 behaviors; therefore, any program dedicated to preventing diabetes within the community has the potential for maximum return through increased movement/exercise participation, not that other risk factors are unimportant, but rather that they should be a second priority when funds are limited.

Secondly, the accessibility of community health interventions must be extended to include all high-risk communities, rather than just those considered the most extreme. For community health interventions that target high risk communities, the greatest population level impact will occur from the 26,607 high risk census tracts (39% of all community census tracts). Implementation of moderately intensive interventions designed to decrease the levels of physical inactivity from 27.7% down to below 20%, within these communities, will result in shifting entire populations into a classification of lower demographic risk. By concentrating on only the census tracts falling within the designation of very high risk (15.7% of all tracts); we will lose sight of the vast majority of communities that experience elevated and addressable levels of risk.

Thirdly, the alignment between cluster profile type and tiered intervention strategies would provide a possible structure for the allocation of resources.

Each cluster exhibits a distinct constellation of risk factors warranting tailored approaches:

Cluster	**Priority Intervention Focus**
Very High Risk	Intensive, multi-factor interventions addressing physical inactivity (38.2%), smoking (25.7%), and sleep (41.2%); structural changes to built environment

High Risk	Physical activity promotion as primary focus; reduce inactivity from 27.7% to <20%
Moderate Risk	Maintenance and prevention; reinforce existing healthy behaviors
Low Risk	Surveillance and best-practice documentation; investigate protective factors

Model Utility for Screening and Resource Allocation:

Having a strong model performance (R^2=0.8728), these four modifiable behavioral factors can accurately predict the prevalence of diabetes amongst Communities and help in targeting resources and screening for diabetes without having to collect an extensive set of demographic data. By using the model from this study, public health agencies can locate high-priority communities using the PLACES dataset. They could then assess the local conditions to ensure they are appropriate prior to implementing any interventions. However, as discussed in the Limitations section, the absence of information regarding already existing programs and facilities in the identified communities means that public health agencies must validate their sustainability and appropriateness in the local context before implementing recommendations.

Comparison with Prior Literature:

Our research expands and corroborates previous studies. The association between inactivity and risk of developing diabetes is similar to that of Yang et al. (2024) who found that there was a 44% reduction in risk of developing diabetes among prediabetic persons who are physically active. Evidence in the Southern United States supports previous research by Wittman et al. (2024) and Benavidez et al. (2024) showing that there are continuing regional disparities in the outcomes from diabetes. We add to these studies by providing a quantification of decision thresholds for these factors and demonstrating that on average, the majority of the variance in community-level risk could be accounted for by behaviors indicating a critical input for intervention design.

The unexpected finding with binge drinking is consistent with concerns voiced by Zhang et al. (2019) regarding the complex interactions between alcohol use and diabetes and suggests that caution should be exercised when using ecological relationships to inform behaviors of individuals.

VII. RECOMMENDATIONS

We have developed an tiered intervention framework that is consistent with community risk classifications using predictive modeling and cluster analysis. However, it is important to remember that the PLACES data does not have any information about current available resources for physical activity or diabetes prevention programs, as well as health care facilities located within the community. Thus, before implementing these recommendations, it is important to compare the suggested interventions with what currently exists within the communities and to evaluate the gaps between them by conducting local community assessments.

Priority 1: High Risk Clusters (39.0% of tracts; 26,607 communities)

High Risk communities are considered to be the target for intervention because of a combination of high-risk levels, population coverage numbers, and ease of potential improvement. They are very close to 20%, physical inactivity levels beyond which serious physical problems might occur.

Current Status:

- Physically inactive: 27.7%

- Diabetes prevalence: 11.5%

- Target: Less than 20% inactive

Our calculation shows that decreasing physical inactivity from 27.7% to the level of Moderate Risk of 18.8% may lower diabetes prevalence from 11.5% to 9.2%, yielding a relative decrease of 20%. This achieves the maximum ROI (return of investment) among all risk factors.

Recommended Interventions (subject to local assessment of existing resources):

- Community-based physical activity programs designed to address local needs and preferences
- Investments in infrastructure upgrades to promote walking and cycling
- Corporate wellness programs with physical activity elements
- Increased access to recreational infrastructure and parks
- Community health worker programs for promoting behavioral changes

Priority 2: Very High-Risk Clusters (15.7% of tracts; 10,717 communities)

A Very High-Risk community refers to areas with extremely high levels of behavioral risk factors. Here, behavioral changes would be insufficient in order to attain a reasonable level of reductions.

Current Status:

- Physical inactivity: 38.2%
- Diabetes prevalence: 16.8%
- Gap to threshold: 18.2 percentage points (versus 7.7 points for High Risk)

Given the scale of transformation that needs to be achieved, such communities necessitate a wide-ranging intervention to:

- Increased access to healthcare: via mobile healthcare clinics and telemedicine programs.
- Environmental design for physical activity (sidewalks, parks, lighting, and security upgrades)

- Economic development programs for overcoming poverty-related factors affecting healthy behavior
- Multi-sector collaborations such as healthcare, housing, transportation, and employment organizations

Geographic targeting is indicated for federal funding: Mississippi (42 percent of Census tracts in Very High Risk), Alabama (38 percent), and Louisiana (35 percent) would be priorities for a collaborative effort.

Priority 3: Moderate Risk Clusters (26.2% of tracts; 17,863 communities)

Moderate Risk communities are below the critical level of 20% physical inactivity (18.8%) and need to maintain and prevent any advancement to a higher risk level.

Recommended approaches:

- Maintain existing physical activity programs and facilities
- Secondary risk factors: address sleep hygiene improvement (30.2% with fewer than 7 hours sleep) and smoking quit assistance (11.1% prevalence)
- Monitor community health indicators for potential warning signs of risk progression
- Launch low-intensity population-wide messages for health promotion

Priority 4: Low Risk Clusters (19.0% of tracts; 12,985 communities)

Low Risk communities (18.1% inactivity, 7.3% diabetes) are models of best practices and should prioritize surveillance and documentation of protective factors.

Recommended approaches:

- Document characteristics of successful communities for replication in other communities with higher risk factors.
- Explore binge drinking paradox (21.0% prevalence with low diabetes prevalence) to address possible confounding variables.
- Maintain current programming while ensuring resource availability
- Exchange best practices via public health networks and learning collaborative meetings

Implementation Priorities

1. Improve physical activity levels: Concentrate efforts on 26,607 High-Risk census tracts where a modest decrease in physical inactivity (from 27.7% to less than 20%) may provide maximum benefits with moderate levels of intervention intensity
2. Federal intervention: Target those counties with a diabetes prevalence ratio above 20% (Greene County, AL: 24.1%; Brooks County, TX: 23.2%; Clay County, GA: 23.1%) for aggressive intervention and investment
3. Regional Coordination: Implement regional programs for the Diabetes Belt where the 10 most prevalent diabetes states cluster.
4. Local validation: Assessments of existing conditions in the community prior to introducing any intervention with a focus on existing assets and specific needs

VIII. LIMITATIONS

There are a number of limitations in this study which must be taken into consideration when drawing inferences and applying recommendations.

Data and Variable Constraints

The CDC PLACES data does not contain demographic variables such as age composition, ethnicity/race, income, education level, and employment. All of these variables are clearly identified risk factors for diabetes based on our literature review and would likely explain a large portion of the 12.7% unexplained variance in our model. Such factors could confound any observed relationships between behavioral variables and diabetes prevalence because demographic variables would be unobserved.

Also, data is missing for:

- Access to healthcare (provider supply, coverage, screening rates)
- Environmental factors (food environment quality, built environment factors)
- Existing community programs and facilities.

This final limitation has particular significance for our recommendations section: we are unable to determine any existing physical activity-related resources, diabetes prevention programs, and healthcare systems for vulnerable communities identified as high-risk. All recommendations made would need validation before being put into action.

The behavioral variables are based on survey data (BRFSS), which tends to underestimate true prevalence because of social desirability bias. Subjects may underestimate disfavored behaviors such as smoking and physical inactivity. They may also overestimate more desirable behaviors. This could affect estimates of prevalence as well as intervariable association.

Methodological Constraints

The cross-sectional design does not allow itself to causal relationships. Though a lack of physical activity strongly correlates with diabetes prevalence (r = 0.86), we cannot determine whether physical inactivity leads to diabetes prevalence in a community. A community with a high level of diabetes prevalence might have low levels of physical activity because of its disease prevalence.

Ecological fallacy tends to be a major flaw in any form of tract level analysis. The results that are obtained for a group of individuals may not always apply to those individuals solely. This particular study aims to address factors based at population level. A community with 27.7% physical inactivity does contain active and inactive individuals. Individual behavior may lead to diabetes differently when checked for a population.

Model Specificity

The 76.6% feature importance for physical inactivity might be considered a biased estimate because of omitted variable bias. Physical inactivity could be proxying variables such as poverty levels, age composition, and access to healthcare because physical inactivity was not adjusted for demographic variables. The strong negative association between binge drinking and diabetes (r = -0.70) does indicate that some socioeconomic variables are unmeasured.

A potential problem with a four-cluster solution based on the optimal solution determined using the elbow method (k = 4) may be that important differences may be masked. Neighborhoods that belong to a particular risk category may vary appreciably regarding their needs for intervention based on their profiles.

Geographic Considerations

The counties with a high prevalence of diabetes were identified (Greene County, AL; Brooks County, TX; Clay County, GA), and their population sizes are relatively small. A possible reason for a high percentage prevalence in those counties could perhaps be a small denominator effect because a small number of cases might lead to a high prevalence percentage. Even then, such counties might be considered for public health focus but with a focus on their populations.

The results of this study apply only to the United States environment with its distinct healthcare system, physical environment trends, and culture. The 20% level of physical inactivity and associated clusters determined for this study may not be universally used for other countries until validated. There might be variations even in America regarding healthcare systems, culture, and economic conditions for a particular region.

IX. FUTURE RESEARCH DIRECTIONS

Immediate Priorities

1. Track communities longitudinally to determine whether a community crosses the 20% physical inactivity threshold before a numerically measurable increase in diabetes prevalence.
2. Connect data from behavioral data source PLACES with variables for demographic factors (age, race, income, education level) available via American Community Survey data and variables for healthcare access available via Area Health Resources Files data to distinguish between behavioral and socioeconomic influences affecting diabetes rates.
3. Assess communities where changes in the built environment (new parks, sidewalks, recreation facilities, complete street projects) were made to measure changes in levels of physical activity and diabetes prevalence for 5 to 10 years.
4. Construct detailed data sets regarding current physical activity facilities, diabetes prevention programs, and healthcare resources within census tracts to facilitate analysis for intervention.

Validation Studies

1. Conduct community-randomized trials to test whether using a risk model based on a four-cluster risk classification leads to more effective intervention targeting than other methods of classification.
2. Test whether a target of decreasing physical inactivity levels from 27.7% to below 20% in High-Risk communities leads to a predicted 20% reduction in diabetes prevalence and determine when results can be observed.

3. Examine using individual data with full socioeconomic adjustment to put forward a theory for why a contradictory negative association between binge drinking and diabetes emerges in community data separating out 'sick quitter' effects from socioeconomic factors and possible observation errors.
4. Investigate whether thresholds of inactivity of 20% and importance levels regarding features differ among communities with variations in demographic characteristics (age composition, racial and ethnic composition, income levels).

X. CONCLUSION

This data analysis of 68,172 U.S. census tracts illustrates that physical inactivity is a prominent modifiable predictor for diabetes community prevalence because physical inactivity explains 76.6% of feature importance to a Random Forest model with 87.3% explained variance (R^2=0.8728). This model directly informs diabetes prevention resource investment because more may be achieved in population-level prevention with a focus on physical activity than with a multipronged approach towards diverse behavior modification.

The finding of four community risk profiles indicates that communities with a high risk for diabetes exist beyond those communities ranked as outliers. As 39 percent of census tracts are ranked as High-Risk communities (socioeconomic factors: 11.5 percent diabetes prevalence; behavioral factors: 27.7 percent physical inactivity), diabetes prevention proves to be a challenge for a major plurality of communities in America. The High-Risk community group with 26,607 census tracts represents the largest opportunity for mass prevention. As this group lies marginally above the critical level for 20 percent physical inactivity levels, a modest improvement toward a lower risk for diabetes can be achieved. By contrast, Very High-Risk communities with 15.7 percent census tracts necessitate intense environmental modification.

One of the most important findings of this study is the determination of about 20% physical inactivity as a tipping point. Communities below this tipping point keep diabetes levels below 10%, but beyond this level, communities witness a sharp rise from 9.2% for Moderate Risk to 11.5% for High Risk and then to 16.8% for Very High-Risk communities.

Regional inequities are still apparent, with a twofold risk ratio separating those with the highest state prevalence (14.4%, Mississippi) and those with the lowest (7.1%, Colorado). Hotspots in counties such as Greene in Alabama (24.1%) are more than double the national average risk but may be influenced by low population numbers in those counties. The presence of strong determinants/SBR beyond behavioral risk factors indicates a need for focus beyond such determinants. The high number of Very High-Risk tracts in Alabama (38%), MS (42%), and LA (35%) may necessitate a state and nationally collaborative effort to target those living in the Diabetes Belt region.

The strong model performance proves that behavioral variables alone are capable of optimizing screening and resource allocation for communities based purely on available PLACES data. But several critical caveats exist. The fact that no demographic variables (age, race, income, and education level) are included means that relationships can be susceptible to bias from unobserved variables, potentially driving the 76.6 percent importance assigned to physical inactivity somewhat above where it might be. A possible presence of existing community resources necessitates that any recommended interventions be checked for appropriateness before being acted upon.

As a contrast to a single nationwide intervention effort, this analysis finds strength in precision public health strategies that target risk profiles. Whereas Low Risk communities may be adequately managed with maintenance and surveillance alone, communities with a Moderate Risk profile would be better off maintaining existing gains while managing secondary risk factors. As for High-Risk communities, behavioral strategies focusing on physical activity would be most effective. Finally, communities identified as Very High Risk would follow a comprehensive approach stressing changes in healthcare access, economic opportunities, and environmental factors alongside behavioral targets.

XI. REFERENCES

Benavidez, G. A., Zahnd, W. E., Hung, P., & Eberth, J. M. (2024). Chronic disease prevalence in the US: Sociodemographic and geographic variations by ZIP code tabulation area. *Preventing Chronic Disease, 21*, E14. https://doi.org/10.5888/pcd21.230267

Chen, Y., Jin, X., Chen, G., Wang, R., & Tian, H. (2024). Dose-response relationship between physical activity and the morbidity and mortality of cardiovascular disease among individuals with diabetes: Meta-analysis of prospective cohort studies. *JMIR Public Health and Surveillance, 10*, e54318. https://doi.org/10.2196/54318

Chou, C., Hsu, D., & Chou, C. (2023). Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine, 13*(3), 406. https://doi.org/10.3390/jpm13030406

Deng, M., Cui, H., Lan, Y., Nie, J., Liang, Y., & Chai, C. (2022). Physical activity, sedentary behavior, and the risk of type 2 diabetes: A two-sample Mendelian randomization analysis in the European population. *Frontiers in Endocrinology, 13*, 964132. https://doi.org/10.3389/fendo.2022.964132

Fu, X., Wang, Y., Cates, R. S., Li, N., Liu, J., Ke, D., Liu, J., Liu, H., & Yan, S. (2023). Implementation of five machine learning methods to predict the 52-week blood glucose level in patients with type 2 diabetes. *Frontiers in Endocrinology, 13*, 1061507. https://doi.org/10.3389/fendo.2022.1061507

Hu, M., Le, M. H., Yeo, Y. H., Wijarnpreecha, K., Likhitsup, A., Kim, D., & Chen, V. L. (2025). Diabetes prevalence and management patterns in US adults, 2001–2023. *Acta Diabetologica*. https://doi.org/10.1007/s00592-025-02572-6

Jayedi, A., Zargar, M., Emadi, A., & Aune, D. (2023). Walking speed and the risk of type 2 diabetes: A systematic review and meta-analysis. *British Journal of Sports Medicine, 58*(6), 334–342. https://doi.org/10.1136/bjsports-2023-107336

Lord, J., & Odoi, A. (2024). Determinants of disparities of diabetes-related hospitalization rates in Florida: A retrospective ecological study using a multiscale geographically weighted regression approach. *International Journal of Health Geographics, 23*(1), 1. https://doi.org/10.1186/s12942-023-00360-5

Lu, H., Yang, Q., Tian, F., Lyu, Y., He, H., Xin, X., & Zheng, X. (2021). A meta-analysis of a cohort study on the association between sleep duration and type 2 diabetes mellitus. *Journal of Diabetes Research, 2021*, 1–15. https://doi.org/10.1155/2021/8861038

Nath, N. D., & Odoi, A. (2024). Geographic disparities and temporal changes of diabetes-related mortality risks in Florida: A retrospective study. *PeerJ, 12*, e17408. https://doi.org/10.7717/peerj.17408

Qin, G., Chen, L., Zheng, J., Wu, X., Li, Y., Yang, K., Liu, T., Fang, Z., & Zhang, Q. (2023). Effect of passive smoking exposure on risk of type 2 diabetes: A systematic review and meta-analysis of prospective cohort studies. *Frontiers in Endocrinology, 14*, 1195354. https://doi.org/10.3389/fendo.2023.1195354

Quiñones, S., Goyal, A., & Ahmed, Z. U. (2021). Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. *Scientific Reports, 11*(1), 6955. https://doi.org/10.1038/s41598-021-85381-5

Sharma, A. (2023). Exploratory spatial analysis of food insecurity and diabetes: An application of multiscale geographically weighted regression. *Annals of GIS, 29*(4), 485–498. https://doi.org/10.1080/19475683.2023.2208199

Shin, J., Kim, J., Lee, C., Yoon, J. Y., Kim, S., Song, S., & Kim, H. (2022). Development of various diabetes prediction models using machine learning techniques. *Diabetes & Metabolism Journal, 46*(4), 650–657. https://doi.org/10.4093/dmj.2021.0115

Silva, N., Choi, H., & Goldman, J. (2024). The impact of sleep in diabetes mellitus. *ADCES in Practice, 12*(5), 8–11. https://doi.org/10.1177/2633559x241263069

Uddin, J., Malla, G., Long, D. L., Zhu, S., Black, N., Cherrington, A., Dutton, G. R., Safford, M. M., Cummings, D. M., Judd, S. E., Levitan, E. B., & Carson, A. P. (2022). The association between neighborhood social and economic environment and prevalent diabetes in urban and rural communities: The Reasons for Geographic and Racial Differences in Stroke (REGARDS) study. *SSM - Population Health, 17*, 101050. https://doi.org/10.1016/j.ssmph.2022.101050

Van Dyke, M. E., Chen, T. J., Nakayama, J. Y., Moore, L. V., & Whitfield, G. P. (2023). Changes in physical inactivity among US adults overall and by sociodemographic characteristics, Behavioral Risk Factor Surveillance System, 2020 versus 2018. *Preventing Chronic Disease, 20*, E65. https://doi.org/10.5888/pcd20.230012

Wittman, J. T., Alexander, D. S., Bing, M., Montierth, R., Xie, H., Benoit, S. R., & Bullard, K. M. (2024). Identifying priority geographic locations for Diabetes Self-Management Education and Support Services in the Appalachian region. *Preventing Chronic Disease, 21*, E27. https://doi.org/10.5888/pcd21.230297

Wu, J., Wang, Y., Xiao, X., Shang, X., He, M., & Zhang, L. (2021). Spatial analysis of incidence of diagnosed type 2 diabetes mellitus and its association with obesity and physical inactivity. *Frontiers in Endocrinology, 12*, 755575. https://doi.org/10.3389/fendo.2021.755575

Yang, W., Wu, Y., Chen, Y., Chen, S., Gao, X., Wu, S., & Sun, L. (2024). Different levels of physical activity and risk of developing type 2 diabetes among adults with prediabetes: A population-based cohort study. *Nutrition Journal, 23*(1), 107. https://doi.org/10.1186/s12937-024-01013-4

Zhang, Y., Pan, X., Chen, J., Xia, L., Cao, A., Zhang, Y., Wang, J., Li, H., Yang, K., Guo, K., He, M., & Pan, A. (2019). Combined lifestyle factors and risk of incident type 2 diabetes and prognosis among individuals with type 2 diabetes: A systematic review and meta-analysis of prospective cohort studies. *Diabetologia, 63*(1), 21–33. https://doi.org/10.1007/s00125-019-04985-9

Zhou, Z., & Tian, X. (2024). Prevalence and association of sleep duration and different volumes of physical activity with type 2 diabetes: The first evidence from CHARLS. *BMC Public Health, 24*(1), 3331. https://doi.org/10.1186/s12889-024-20743-y

Zhu, P., Lao, G., Li, H., Tan, R., Gu, J., & Ran, J. (2023). Replacing of sedentary behavior with physical activity and the risk of mortality in people with prediabetes and diabetes: A prospective cohort study. *International Journal of Behavioral Nutrition and Physical Activity, 20*(1), 81. https://doi.org/10.1186/s12966-023-01488-0

XII. APPENDICES

APPENDIX A: TABLES

Table A1: Summary Statistics for Key Variables

Variable	Mean (%)	Std Dev	Min	Max	Correlation with Diabetes
Diabetes	10.9	3.7	0.7	46.1	1.00
Physical Inactivity	23.1	7.2	2.8	68.9	0.86
Current Smoking	16.8	5.9	1.1	57.4	0.73
Sleep <7 Hours	33.4	5.4	14.1	61.2	0.7
Binge Drinking	15.9	4.3	2.3	41.8	-0.70

N = 68,172 census tracts across 51 states/territories and 1,839 counties

Table A2: Top 10 States by Average Diabetes Prevalence

Rank	State	Mean Prevalence (%)	Std Dev	Census Tracts (n)
1	Mississippi	14.4	4.5	658

2	West Virginia	14.2	2.8	484
3	Alabama	14.1	4.8	1,175
4	Louisiana	13.6	4.7	1,124
5	Tennessee	12.9	4.0	1,480
6	Ohio	12.8	4.3	2,940
7	South Carolina	12.8	4.1	1,089
8	Oklahoma	12.6	3.5	1,045
9	Arkansas	12.6	3.5	684
10	New Mexico	12.3	3.3	498

Table A3: Top 10 Counties by Average Diabetes Prevalence

Rank	**County**	**State**	**Mean Prevalence (%)**	**Census Tracts (n)**	**Total Population (approx.)**
1	Greene	AL	24.1	3	8,000
2	Brooks	TX	23.2	2	7,000
3	Clay	GA	23.1	1	3,000
4	Presidio	TX	22.6	2	6,500
5	Wilcox	AL	22.1	2	10,500
6	Holmes	MS	22.0	4	17,000
7	Perry	AL	21.8	3	9,000
8	Dimmit	TX	21.8	2	10,000

9	Kenedy	TX	21.8	1	400
10	Tensas	LA	21.4	2	4,500

The population figures are estimated. The low population in the respective counties might exaggerate the rate of prevalence because a low denominator will be involved in calculating the rates, for instance, a population of a few hundred patients with diabetes might present a high rate of the disease.

Table A4: Predictive Model Performance Comparison

Model	**R²**	**Adjusted R²**	**RMSE**	**MAE**
Random Forest	0.8728	0.8728	1.326	0.962
Gradient Boosting	0.8688	0.8688	1.347	0.992
Decision Tree	0.8528	0.8528	1.426	1.027
Linear Regression	0.8314	0.8313	1.527	1.123

Train/test split = 80/20 (54,537 training, 13,635 test). Random state = 42 for reproducibility.

Table A5: Random Forest Classification Performance by Risk Category

Risk Category	**Precision**	**Recall**	**F1 Score**	**Support (n)**
Very High Risk	0.862	0.797	0.828	3,423
High Risk	0.639	0.666	0.652	3,543
Moderate Risk	0.604	0.632	0.618	3,433
Low Risk	0.795	0.781	0.788	3,236
Overall	**0.723**	**0.718**	**0.720**	**13,635**

Risk categories defined by training data quartiles (Q1=8.4%, Q2=10.3%, Q3=12.8%)

Table A6: Random Forest Feature Importance

Feature	Importance	Rank
Physical Inactivity	76.6%	1
Binge Drinking	15.5%	2
Sleep <7 Hours	4.3%	3
Current Smoking	3.6%	4

Table A7: Health Risk Profiles by Cluster

Cluster	Tracts (n)	% of Total	Diabetes (%)	Physical Inactivity (%)	Smoking (%)	Sleep <7h (%)	Binge Drinking (%)
Very High Risk	10,717	15.7%	16.8	38.2	25.7	41.2	13.6
High Risk	26,607	39.0%	11.5	27.7	18.2	35.2	16.3
Moderate Risk	17,863	26.2%	9.2	18.8	11.1	30.2	15.8
Low Risk	12,985	19.0%	7.3	18.1	12.7	30.6	21.0

Table A8: Risk Thresholds and Decision Points

Physical Inactivity Level	Diabetes Prevalence	Risk Category	Interpretation
<18%	~7.3%	Low Risk	Baseline community health
18–20%	~9.2%	Moderate Risk	Transition zone

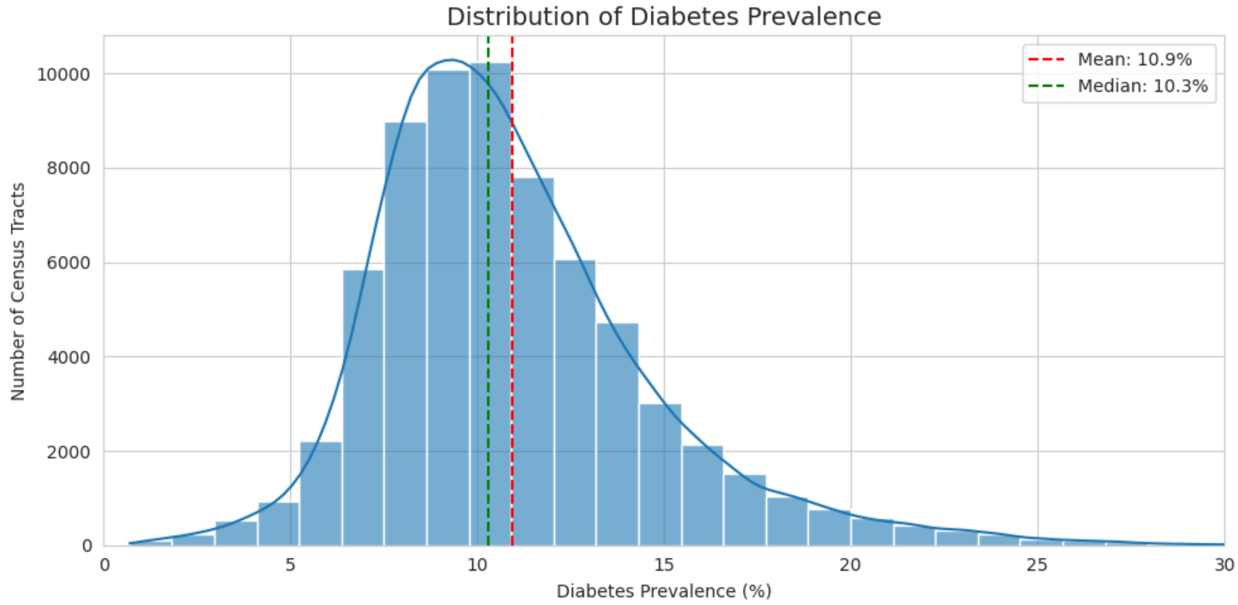
~20%	**~10%**	**Critical Threshold**	**Tipping point**
20–28%	~11.5%	High Risk	Elevated risk
>28%	>12%	Very High Risk	Intervention priority
>38%	>16%	Extreme Risk	Urgent intervention needed

Table A9: Residual Statistics for Random Forest Model

Statistic	**Value**
Mean Residual	0.0403
Standard Deviation	1.3254
Min Residual	−11.2
Max Residual	12.8
% Within ±1 SD	68.3%
% Within ±2 SD	95.1%

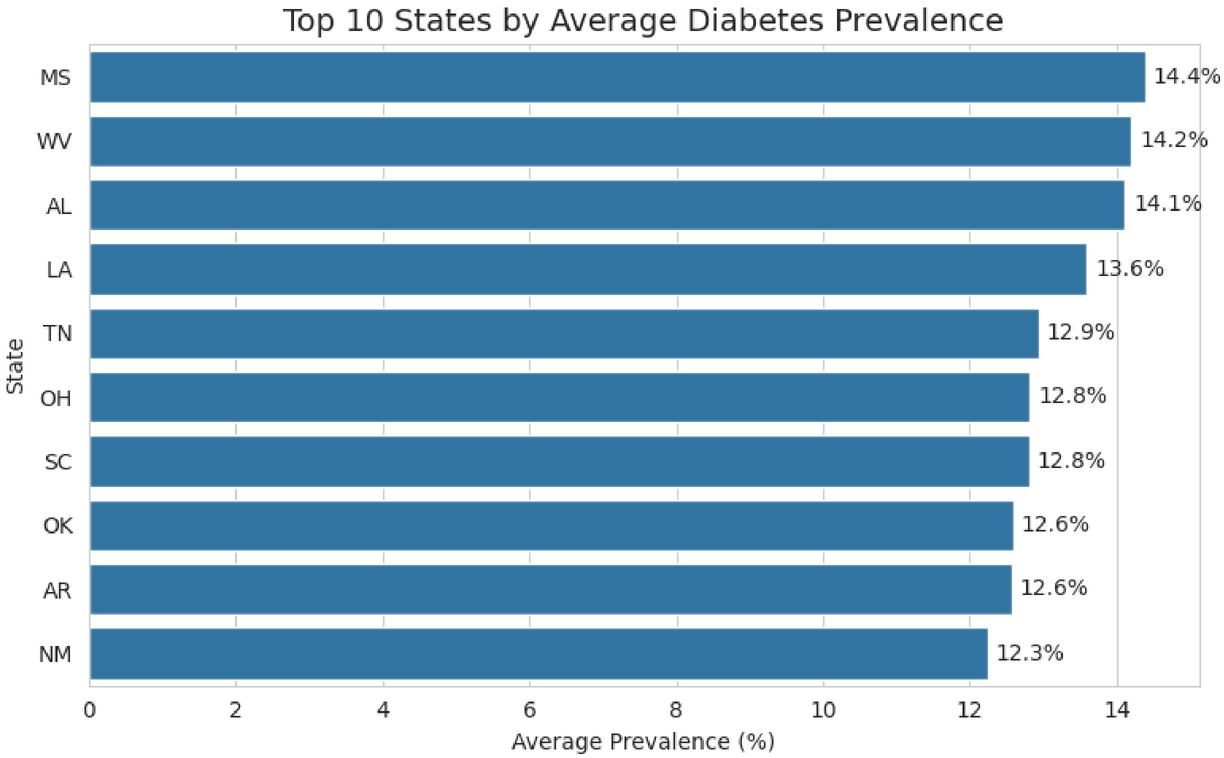
APPENDIX B: FIGURES

Figure B1: Distribution of Diabetes Prevalence



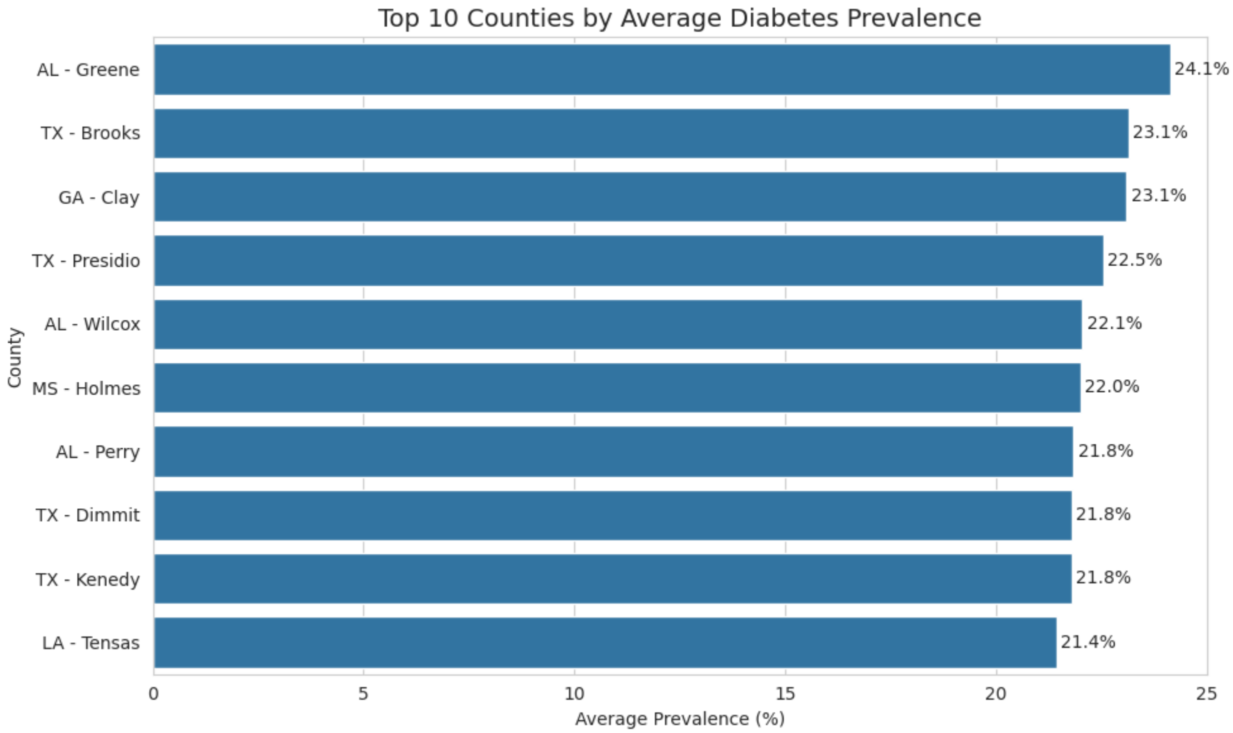
Histogram of the distribution of diabetes prevalence in the census tracts. The data has a right-skewed distribution. The mean = 10.9%, while the median = 10.3%. The red dashed line represents the mean, and the green dashed line represents the median. The x-axis is restricted to the range of 0% to 30% due to the large number of data points that fall into the diabetes category

Figure B2: Top 10 States by Average Diabetes Prevalence



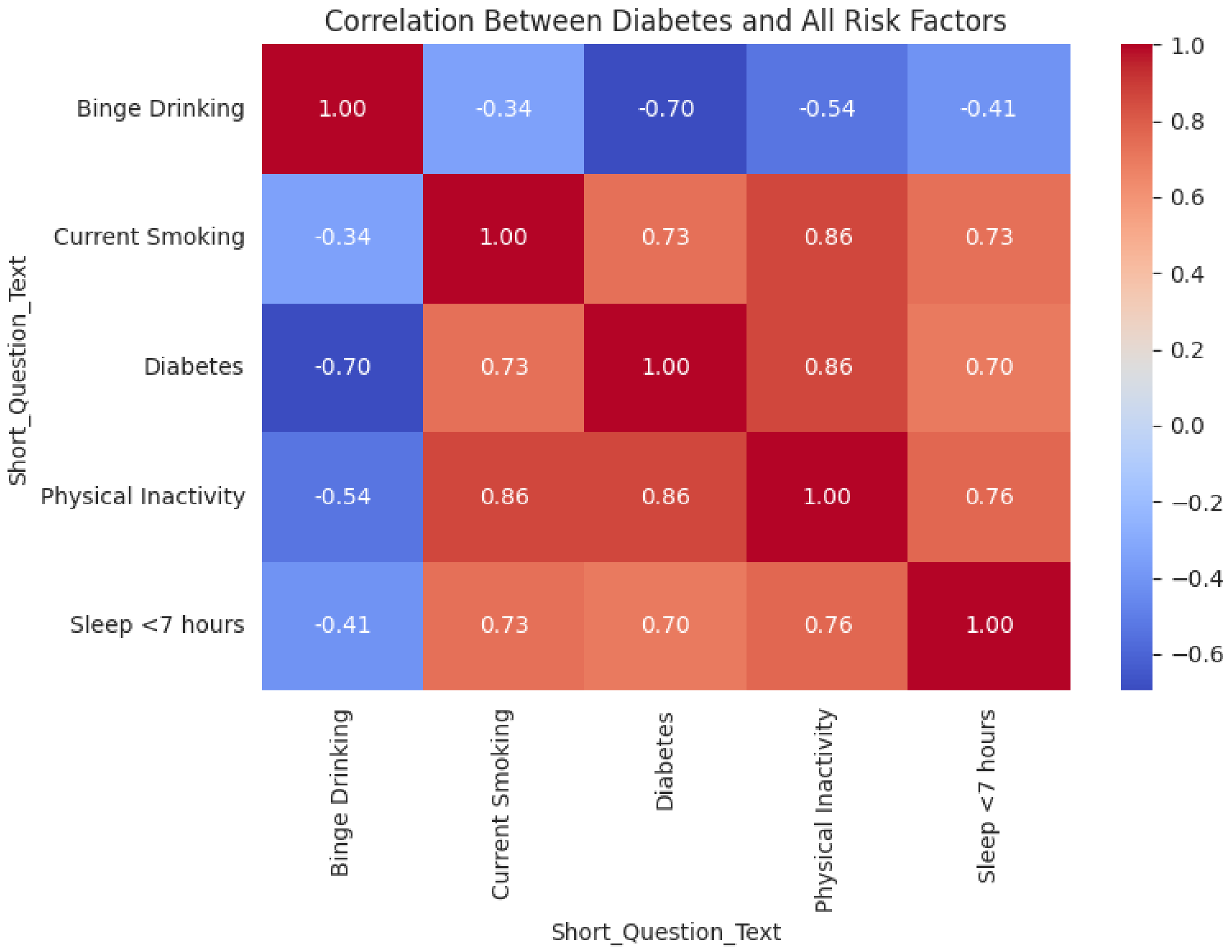
We have compared the average diabetes rate in the top 10 states through a horizontal bar chart, the top state with the highest average diabetes rate is Mississippi, with a rate of 14.4%, followed by West Virginia and Alabama with rates of 14.2% and 14.1%, respectively. The top 10 states are all from the Southern and Appalachian region, showing the "Diabetes Belt" pattern

Figure B3: Top 10 Counties by Average Diabetes Prevalence



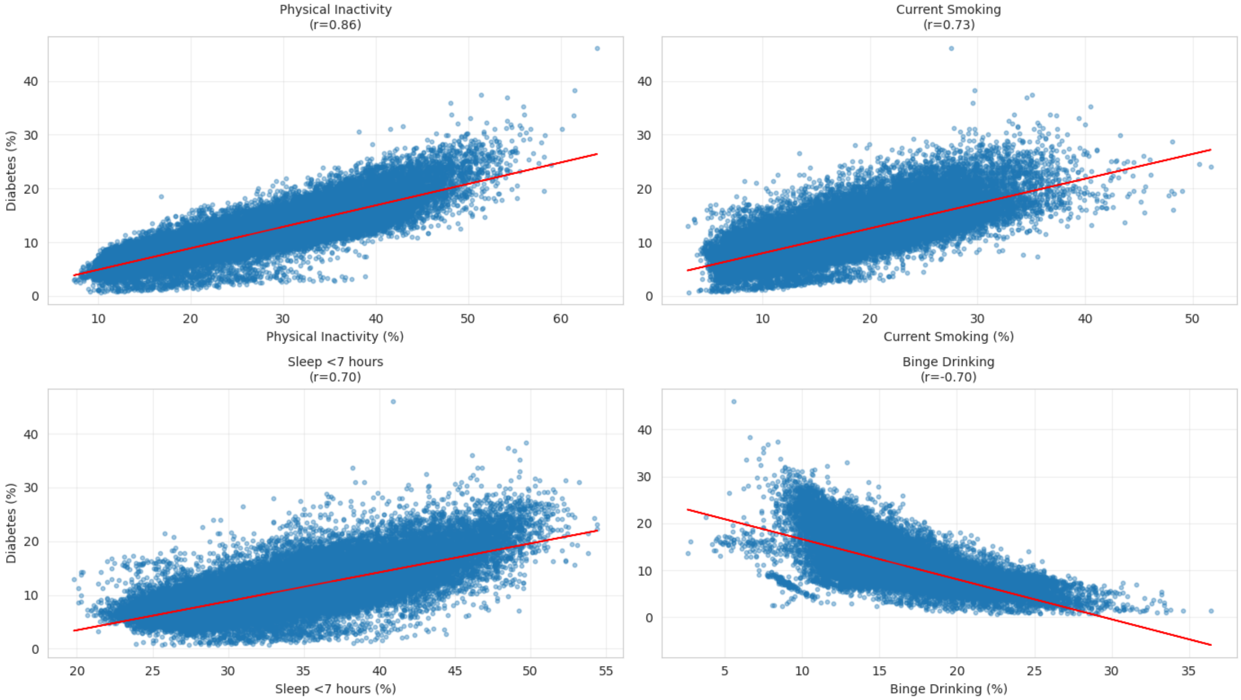
Horizontal bar chart showing the hotspots at county-level. The highest rates are in Greene County, AL (24.1%), Brooks County, TX (23.2%), and Clay County, GA (23.1%). The population in the above counties is relatively low and thus the high percentages.

Figure B4: Correlation Heatmap



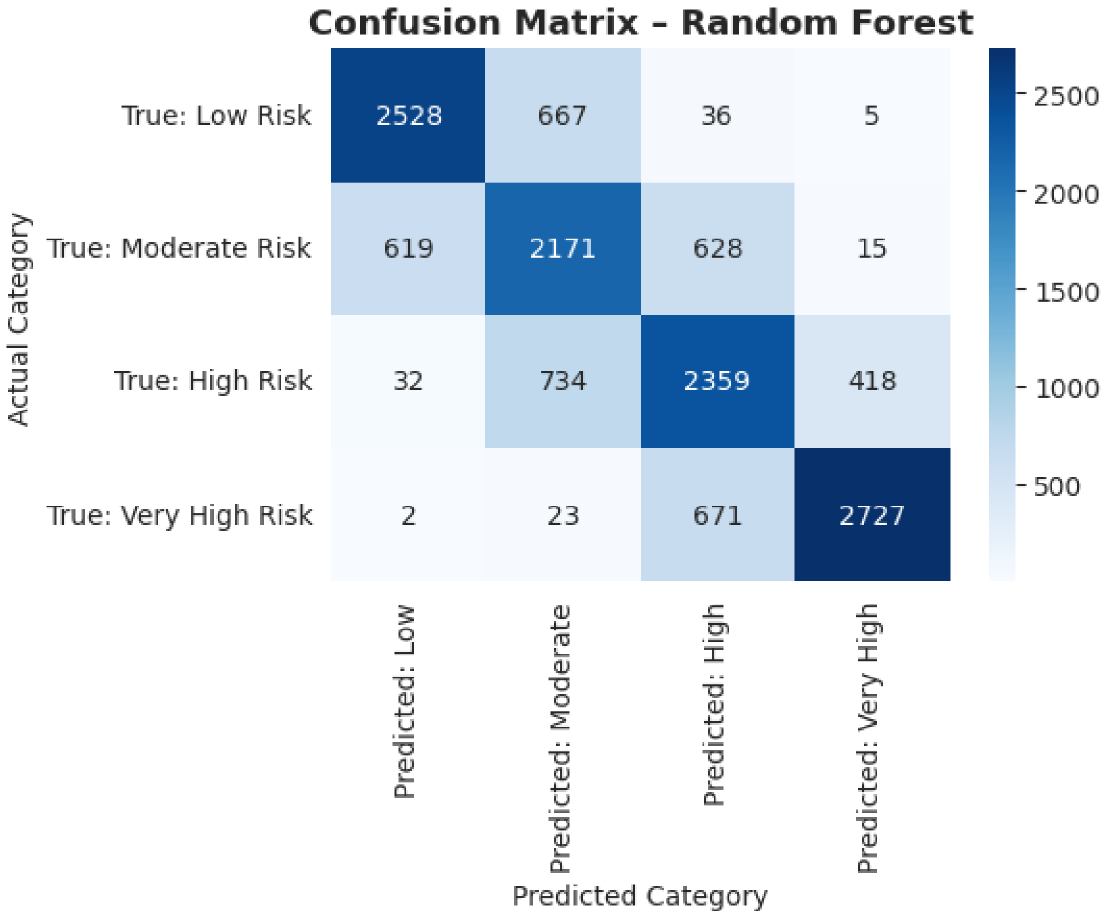
Correlation matrix illustrating relationships between diabetes prevalence and four behavioral risk factors. Physical inactivity has the highest positive correlation with diabetes (r=0.86), followed by smoking (r=0.73) and sleep <7 hours (r=0.70). Binge drinking has a high negative correlation (r=−0.70). The highest inter-predictor correlation occurred between physical inactivity and smoking (r=0.76).

Figure B5: Scatter Plot Matrix (Diabetes vs Risk Factors)



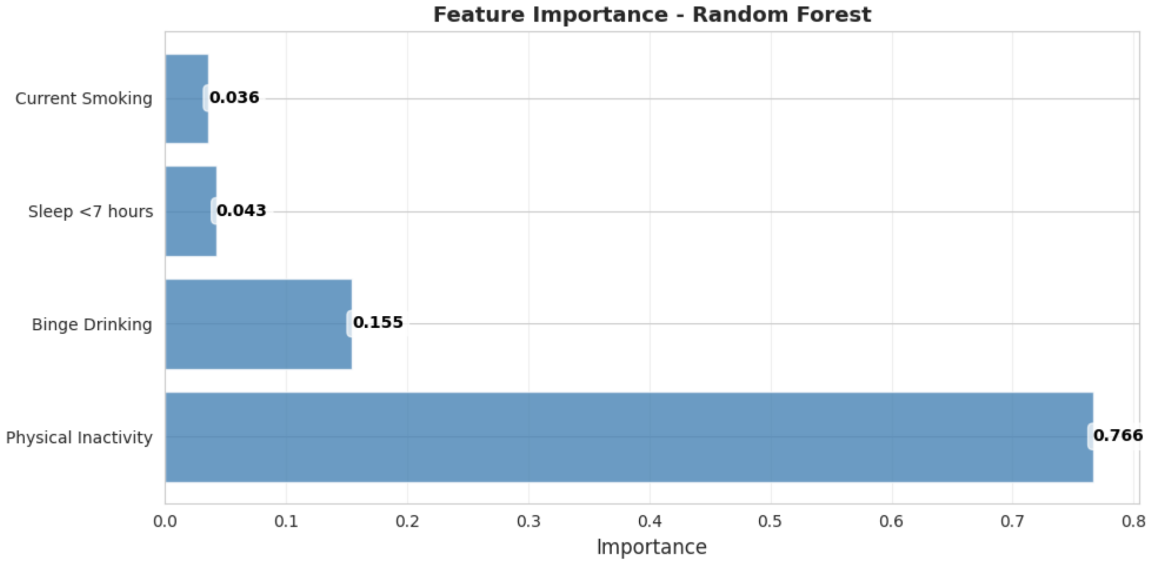
Four scatter plots with regression lines illustrating the relationship of each behavioral risk factor to the prevalence of diabetes. The strongest linear relationship with the least scatter is for physical inactivity. The relationship for binge drinking has a negative slope. There are no important non-linear associations.

Figure B6: Confusion Matrix for Random Forest Risk Classification



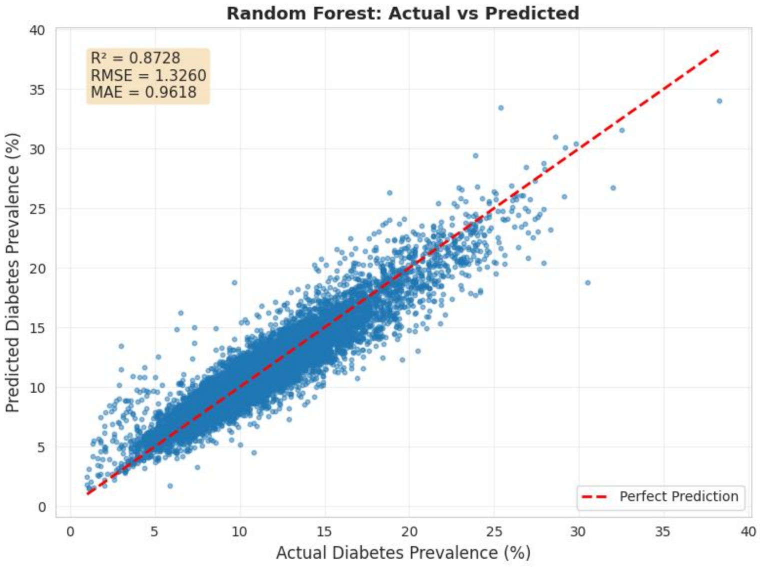
Confusion matrix showing classification performance across four risk categories. Diagonal cells represent correct classifications. The model performs best at extremes: Very High Risk (2,727 correct of 3,423) and Low Risk (2,528 correct of 3,236). Misclassifications occur primarily between adjacent categories.

Figure B7: Feature Importance from Random Forest Model



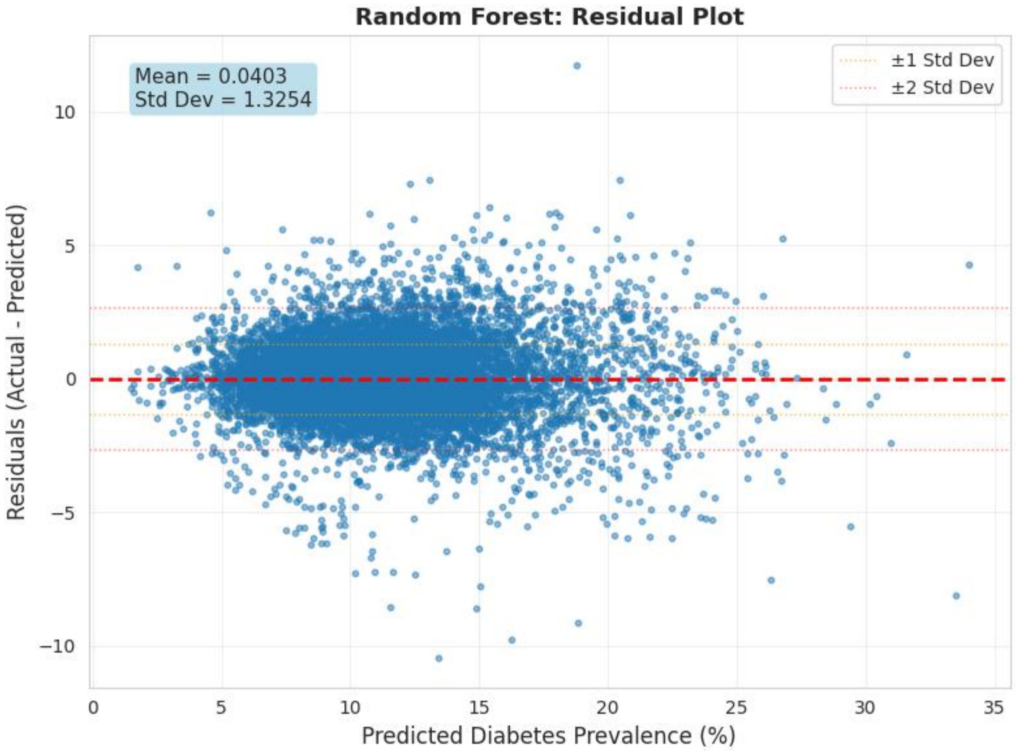
Feature importance scores for the Random Forest model shows that the physical inactivity has the highest importance score at 76.6%, followed by binge drinking at 15.5%, sleep <7 hours at 4.3%, and current smoking at 3.6%.

Figure B8: Random Forest Actual vs Predicted Plot



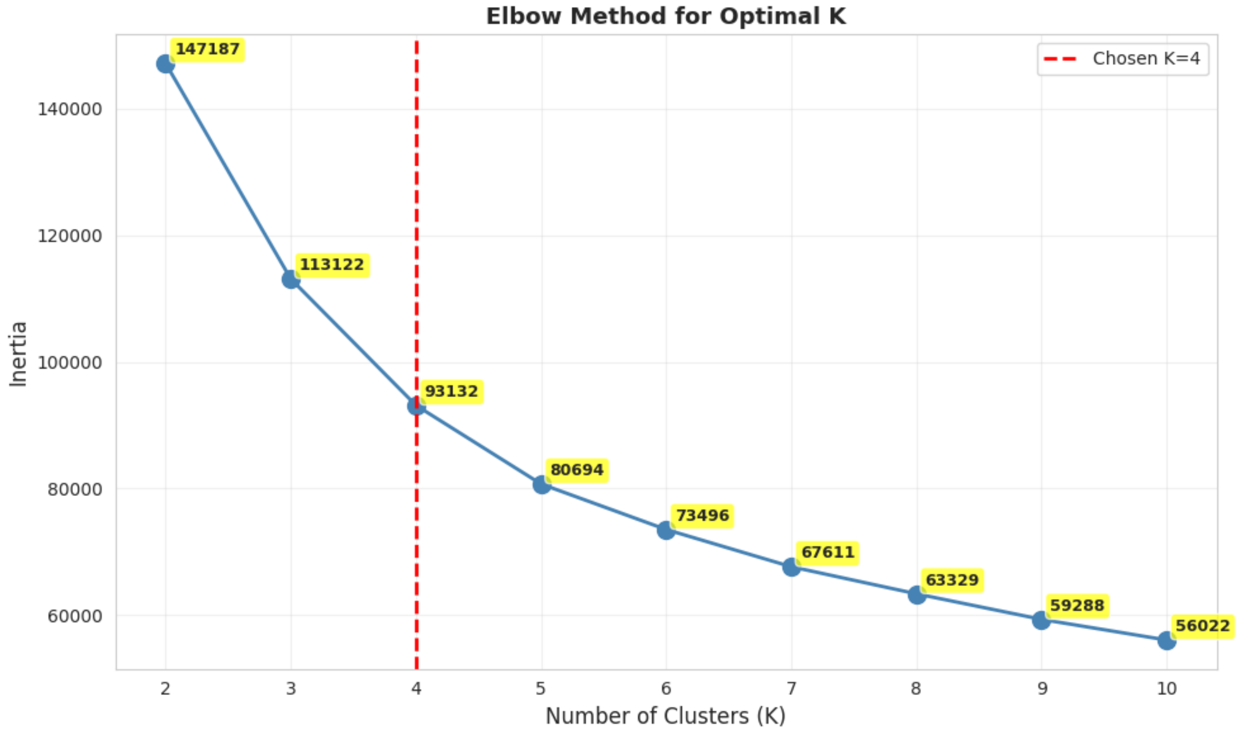
Scatter plot comparing the actual diabetes prevalence represented in x-axis, while the y-axis presents the predicted values. The red dotted line shows the perfect fit. The points are tightly packed along the diagonal line, and the metrics include R^2=0.8728, RMSE=1.326, and MAE=0.962.

Figure B9: Residual Plot for Random Forest Model



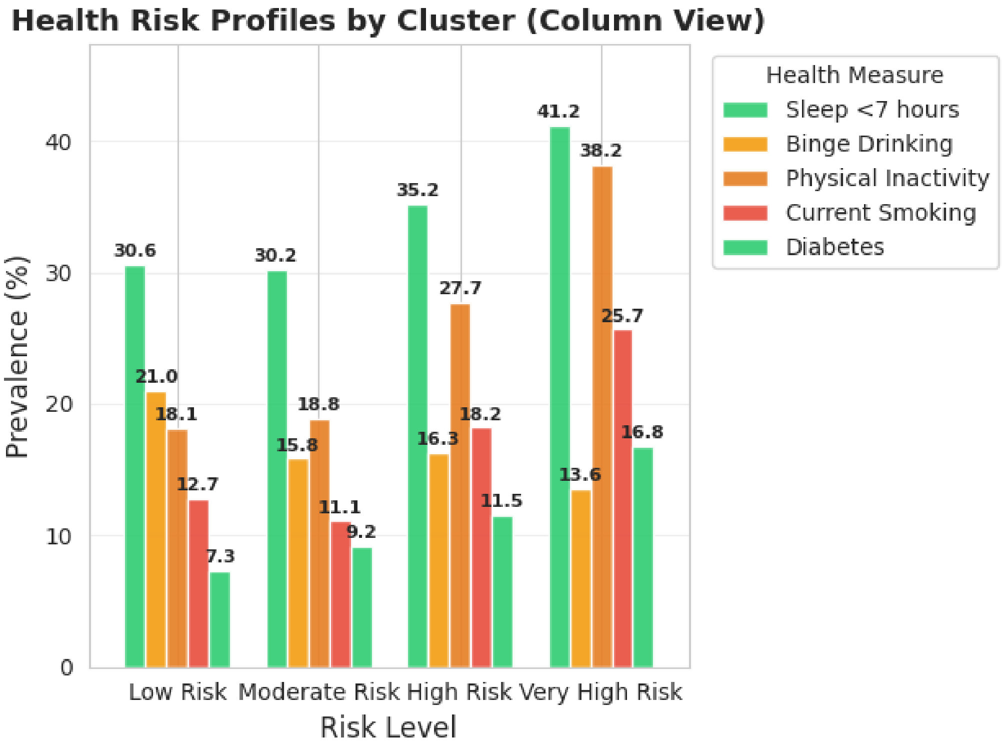
Residuals vs Predicted Values plot shows the residuals are scattered randomly and evenly around the center line corresponding to a value of 0.04. The orange dotted lines represent values of +1 and -1 standard deviation, while the red lines represent +2 and -2 standard deviations. Small amounts of heteroscedasticity occur in the higher predicted values.

Figure B10: Elbow Plot for K-means Clustering



Elbow plot with values of within-cluster inertia (y) and number of clusters (x), where k ranges from 2 to 10. k=2: (147,187), k=3: (113,122), k=4: (93,132), k=5: (80,694). Red line to highlight the value of k=4 where the elbow point occurs (23% reduction from k=3 to k=4 vs. 13% from k=4 to k=5).

Figure B11: Health Risk Profiles by Cluster



Bar chart comparing the mean prevalence rate of each behavioral factor to diabetes for the four risk groups. Physical inactivity has the sharpest rise in prevalence from Low Risk (18.1%) to Very High Risk (38.2%). The highest rates of binge drinking are in the Low Risk group (21.0%) and lowest in Very High Risk groups (13.6%).

APPENDIX C: MODEL PARAMETERS AND CONFIGURATION

C1: Data Preparation Summary

Step	**Description**	**Result**
Original dataset	CDC PLACES 2023 Census Tract Data	2,555,113 rows
Filter health measures	Diabetes + 4 behavioral risk factors	345,025 rows
Pivot to tract-level	One row per LocationID	68,172 rows
Missing data removal	Dropped rows with missing prevalence	1 row removed

Columns dropped	Year, DataSource, footnotes, etc.	10 columns removed
Final dataset	Complete data for analysis	68,172 tracts × 5 variables

C2: Machine Learning Model Parameters

Model	**Parameters**
Linear Regression	Default sklearn parameters
Decision Tree	max_depth=10, random_state=42
Random Forest	n_estimators=100, max_depth=15, random_state=42
Gradient Boosting	n_estimators=100, max_depth=5, random_state=42

C3: K-means Clustering Parameters

Parameter	**Value**
Algorithm	K-means
n_clusters	4
n_init	10
random_state	42
Preprocessing	StandardScaler (z-score normalization)

C4: Train/Test Split Configuration

Parameter	**Value**
Test size	20%

Training samples	54,537 census tracts
Test samples	13,635 census tracts
Random state	42

C5: Risk Category Thresholds (Based on Training Data Quartiles)

Category	**Diabetes Prevalence Range**
Low Risk	< 8.4% (below Q1)
Moderate Risk	8.4% – 10.3% (Q1 to Q2)
High Risk	10.3% – 12.8% (Q2 to Q3)
Very High Risk	> 12.8% (above Q3)

APPENDIX D: DATA SOURCE INFORMATION

Dataset: CDC PLACES: Local Data for Better Health, Census Tract Data 2023 Release

Source: Centers for Disease Control and Prevention (CDC), Division of Population Health

URL: https://catalog.data.gov/dataset/places-local-data-for-better-health-census-tract-data-2023-release

Data Year: 2021 (released 2023)

Geographic Coverage: 50 states + District of Columbia

Unit of Analysis: Census tract (average population ~4,000 residents)

Variables Used:

- Diabetes: Model-based estimate of diagnosed diabetes prevalence among adults ≥18 years

- Physical Inactivity: Adults reporting no leisure-time physical activity in past month
- Current Smoking: Adults who currently smoke cigarettes
- Sleep <7 Hours: Adults reporting fewer than 7 hours of sleep per night
- Binge Drinking: Adults reporting heavy episodic drinking in past 30 days