Proactive Healthcare Analytics: Early Detection of Diabetes with SDOH Insights and Machine Learning

Jesu Marcus Immanuvel Arockiasamy
(Engineer Lead Sr. and DevOps Expert, Leading Healthcare Company)
Email: jesumarcus@gmail.com
ORCID: 0009-0006-4906-9051

doi: https://doi.org/10.37745/ejcsit.2013/vol13n26474 Published February 22, 2025

Citation: Arockiasamy J.M.I. (2025) Proactive Healthcare Analytics: Early Detection of Diabetes with SDOH Insights and Machine Learning, *European Journal of Computer Science and Information Technology*,13(2),64-74

Abstract: *This white paper presents a proactive healthcare analytics framework for early diabetes detection, combining Social Determinants of Health (SDOH) with machine learning. Traditional models only use clinical biomarkers, ignoring socioeconomic factors like income levels, food access and healthcare availability. By including SDOH data from CDC, County Health Rankings and USDA Food Access Atlas we improve predictive accuracy and get population level insights. Using optimized XGBoost our model has an R² of 0.88 and MAE of 0.63, beating baseline models. The study shows how healthcare analytics can move diabetes prevention from reactive to proactive and support personalized interventions and public health initiatives. We propose integration into healthcare systems via real-time APIs and predictive analytics dashboards. This research highlights the importance of SDOH aware models in addressing health disparities and informing data driven policy decisions.*

Keywords: diabetes, SDOH, machine learning, healthcare analytics, XGB

INTRODCUTION

The Need for Healthcare Analytics in Diabetes Prevention

More than **38.4** million people in the United States have **diabetes**, and it is expected to be the seventh leading cause of death by 2030 [1]. Multiple studies have already shown that diabetes affects public health, with disparities rooted in socioeconomic factors such as income, education, and food access. SDOH—conditions in which people are born, live, and work—account for 80% of health outcomes [2], yet traditional clinical models often overlook these important variables.

Traditional diabetes prediction models focus only on biomarkers such as BMI, HbA1c levels, skin thickness, and genetic predisposition. These models fail to account for environmental and socioeconomic factors contributing to diabetes risk , such as income, access to healthy foods. For example, the availability of supermarkets with nutritious food options is a key determinant of dietary habits, which directly impacts diabetes management and prevention. Communities without access to supermarkets with healthy food options face higher rates of diabetes, even when controlling for individual health behaviors [3]. Indeed, this indicates that a more robust approach is needed for predicting diabetes, integrated with SDOH(Social Determinants of Health) data and clinical insight.

Healthcare analytics involves using data, statistical models, and AI-driven techniques to improve patient care, optimize resource allocation, and enhance disease prevention. Combining the EHR data and SDOH data makes it possible to build a coherent machine-learning model that can reliably predict and intervene to improve public health.

The relationship between social determinants of health (SDoH) and diabetes has been established but is poorly understood. This is mainly because the relationship between SDoH factors and diabetes is not linear and straightforward, as some of the SDOH factors are individual and some are societal-level variables as shown in Figure 1.

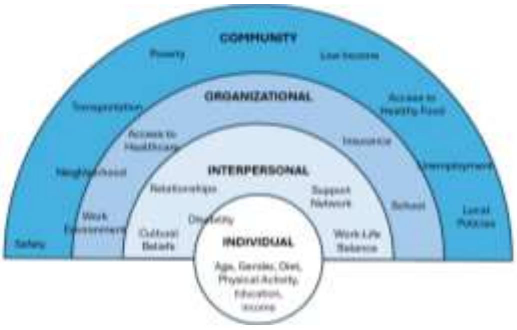


Figure 1 - Four-Dimensional Framework

To address such challenges, machine learning algorithms can be leveraged to combine health information and SDoH factors to address the nuanced relationships between these variables. For this study, we considered SDOH factors like poverty, median home income, low-income tracts, access to nearby grocery stores (for healthy foods), and diabetes prevalence. At the community level, using enhanced data pipelines and optimized XGBoost algorithms, our healthcare analytics framework achieves unprecedented accuracy,

- MAE 0.63 - Predicts diabetes prevalence within 0.63 percentage points of actual rates
- R^2 0.88 - Explains 88% of diabetes variance through SDOH factors
- 10.68% MAPE - Maintains clinical-grade precision across diverse populations

OBJECTIVE OF THIS STUDY

This paper aims to:

- Predict diabetes risk using county/census tract-level diabetes prevalence, SDOH data with following sources:
  - CDC Better Health County Data [4]
  - Behavioural Risk Factor Surveillance System (BRFSS) Prevalence Data (From 2011 to 2025) [5]
  - Food Access Research Atlas (formerly known as Food Desert Locator) from US Department of Agriculture [6]
- Apply multiple ML algorithms to assess and compare predictive accuracy
- Incorporate those findings into healthcare analytics frameworks for proactive disease prevention

LITERATURE REVIEW AND THEORETICAL FOUNDATION

WHAT IS SDOH, AND WHY DOES IT MATTER?

SDOH stands for social determinants of health. According to the World Health Organization, SDOH refers to the conditions in which we are born, grow, live, work, and age. These conditions are shaped by the distribution of money, power, and resources at the global, national, and local levels. They are the roots of health inequities—unfair and avoidable differences in health status within and between countries.

The relationship between SDOH and diabetes has been well studied, and research consistently shows that socioeconomic factors such as income, education, and food access are strong predictors of diabetes [7]. For example, a study in Diabetes Care found that people living in food deserts are 1.5 times more likely to develop type 2 diabetes than those with access to healthy food options [8]. Low-income communities have higher diabetes rates because of limited access to healthcare and preventive services [8].

But the relationships between SDOH and diabetes are complex, non-linear and multi-pronged. For example, poverty and low wages are established risk factors but their impact varies by region and community [9]. This complexity demands advanced analytical approaches to model these relationships.

MACHINE LEARNING IN HEALTHCARE ANALYTICS

The rise of machine learning has revolutionized healthcare analytics by allowing the integration of diverse data sources—from EHRs to SDOH metrics—into a single predictive model. Recent studies have shown the effectiveness of ML algorithms in diabetes prediction, with XGBoost and Neural Networks having R^2 values of 0.60-0.75 in similar contexts [10].

However, the challenge lies in the SDOH data, which is heterogeneous and fragmented across multiple sources. That is a huge stumbling block when it comes to building comprehensive predictive models; This study addresses this challenge by using enhanced data pipelines that unify EHR and SDOH datasets for a holistic analysis.

MODEL PERFORMANCE METRICS

Here is a quick cheat sheet about metrics used to compare each machine learning model,

Metric	Meaning	Description	Interpretation
MAE	Mean Absolute Error	Average absolute error in predictions.	Lower values mean predictions are more accurate.
MSE	Mean Squared Error	Average of squared prediction errors.	Lower values indicate fewer large errors.
RMSE	Root Mean Squared Error	Square root of the average squared errors.	Lower values show better accuracy, with units matching the target variable.
R²	Coefficient of Determination	Proportion of variance explained by the model.	Values closer to 1 indicate a better fit to the

Table 1 Metrics Cheat Sheet

GAPS IN EXISTING RESEARCH

Despite these advances, several gaps remain:

- Limited SDOH Integration: Numerous existing models focus just on clinical data, disregarding the critical role of socio-economic factors.
- Lack of Community-Level Insights: Most studies analyse individual-level data, missing opportunities for population-level interventions.
- Inadequate Validation: Only a few models are validated across different geographic and demographic contexts, limiting generalizability.

This study addresses these gaps by:

- Developing a community-level prediction model that integrates SDOH data.
- Validating the model across multiple datasets and regions.
- Providing actionable insights for healthcare systems and policymakers.

METHODOLOGY

DATA COLLECTION AND PREPROCESSING

We integrated three primary datasets into a unified healthcare analytics framework:

1. County Health Rankings: Poverty rate, median household income, diabetes prevalence index and urban-rural classification [4] .
2. CDC BRFSS data (2011–2025): Diabetes prevalence (self-reported) and behavioral risk factors at the county level [5].
3. USDA Food Access Research Atlas: Food desert status, low-income tract designation, and grocery store proximity (for this study, beyond-10-miles-radius was selected) [6] .

Preprocessing Steps:

- Geospatial Alignment: County names are standardized and matched with their respective states and FIPS codes across datasets.
- Missing Data: Imputed missing SDOH values using k-nearest neighbors (k=5) based on geographic proximity [11].

Feature Engineering:

- Created a composite SDOH Risk Index combining poverty rate (40%), food desert score (30%), and income-to-poverty ratio (30%).
- Normalized features using z-score standardization.

We compared three ML architectures optimized for healthcare analytics: Baseline XGBoost, XGBoost with Hyperparameter Tuning, Neural Networks

MODEL TRAINING AND VALIDATION FRAMEWORK

There are all the features used in the training model and the data split used for training vs testing for all models,

Data Splitting	Feature Selection
Training: 80% Testing: 20% with spatial cross-validation to prevent geographic leakage.	- Poverty rate, median income, food desert status (beyond 10-mile radius), urban/rural classification - Against diabetes prevalences

Table 2 Data Selection and Training Strategy

For validation framework,
Spatial Cross-Validation: Used to ensure the generalizability of the model across different geographic regions. By partitioning the data based on spatial regions rather than randomly, this method helps assess the model's performance in predicting diabetes prevalence in unseen locations, accounting for geographic variability in SDOH factors.

Statistical Tests: Paired t-tests are employed to statistically compare the performance of different models, with a significance level (α) set at 0.05. This test determines whether there is a statistically significant difference in performance metrics between models across repeated cross-validation folds, ensuring robustness in model comparisons.

BASELINE XGBOOST MODELING

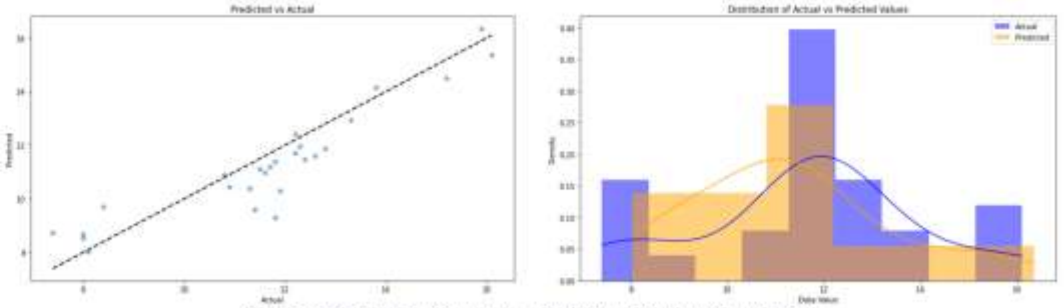
XGBoost or eXtreme Gradient Boosting is a powerful machine learning algorithm based on gradient boosting. It is known for its performance and efficiency in various predictive modeling tasks. XGBoost builds an ensemble of decision trees sequentially to reduce the errors from previous models. XGBoost works by training multiple decision trees, each tree is trained on a subset of the data, and the predictions from each tree are combined to form the final prediction.

Mathematical Foundation	How It Fits SDOH and Diabetes Prediction
$obj(\theta) = \sum_i^n \overset{\text{Squared error sum of Diabetes Prevalence Samples in the training tree}}{l}(y_i, \hat{y}_i) + \overset{\text{Sum over all trees to minimize the variance}}{\sum_{k=1}^K} \Omega(f_k)$ where $\underset{\text{To Minimize overfitting}}{\Omega(f)} = \underbrace{\gamma T}_{\text{Number of leaves}} + \frac{1}{2}\underbrace{\lambda\|w\|^2}_{\text{L2 norm of leaf scores}} \leftarrow \text{Prediction at each leaf}$	- Handles mixed data types (e.g., continuous poverty rates, binary food desert status). - Captures non-linear relationships between SDOH factors and diabetes prevalence. - Efficiently processes large datasets, making it suitable for county-level analysis.

Table 3 Baseline XGBoost Model

Overall Performance of the model

To understand the following metrics, it's important to note that diabetes prevalence is typically in the single digits or low teens. For instance, in the United States, data from the Centers for Disease Control and Prevention (CDC) indicate that the prevalence of diagnosed diabetes among adults is usually around 10-13%. Therefore, even small variations in predictions can result in significant deviations from actual values.



Baseline XGBoost Performance – Diabetes Prevalence Prediction

Context: Diabetes prevalence usually in single digits or in early teens in population

Figure 1 Baseline XGBoost - Scatter Plot and Histogram

Model	MAE	MSE	RMSE	R²
Baseline XGBoost	0.76	0.88	0.94	0.83

Table 4 Baseline XGBoost Model Metrics

LIGHTGBM (LIGHT GRADIENT BOOSTING MACHINE) MODELING

LightGBM is an ensemble learning framework, precisely another gradient boosting method, which constructs a strong learner by sequentially adding weak learners in a gradient descent manner. It optimizes memory usage and training time with Gradient-based One-Side Sampling (GOSS) techniques. LightGBM employs histogram-based algorithms for efficient tree construction. These techniques and optimizations, like leaf-wise tree growth and efficient data storage formats, contribute to LightGBM's efficiency. They make LightGBM faster and have lower memory usage than XGBoost frameworks.

Mathematical Foundation	How It Fits SDOH and Diabetes Prediction
$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \underbrace{\frac{1}{2}\lambda\|w\|^2}_{\text{Regularization Parameters}} \text{Weight of each leaf } k$	- Handles large datasets faster than XGBoost. - Automatically handles categorical features (e.g., urban/rural classification). - Effective for high-dimensional data, such as multiple SDOH indicators.

Table 5 LightGBM Modeling

Overall Performance of the model

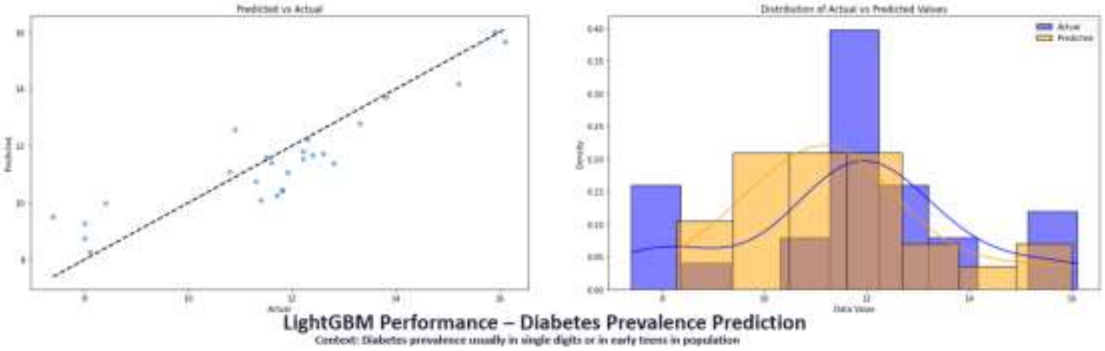


Figure 2 LightGBM - Scatter Plot and Histogram

Model	MAE	MSE	RMSE	R²
LightGBM	0.82	1.00	1.00	0.80

Table 6 LightGBM Modeling Metrics

3.4 OPTIMIZED XGBOOST MODELING

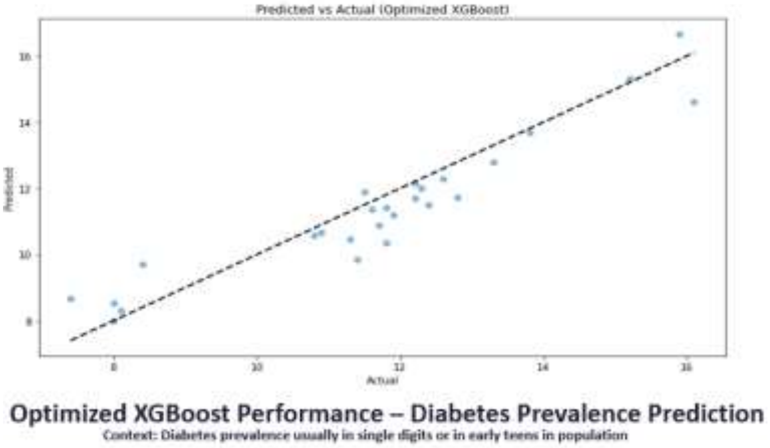
This is an enhanced version of the XGBoost algorithm, fine-tuned with hyperparameters optimized through Bayesian search to boost predictive accuracy. This advanced tuning method efficiently explores the hyperparameter space, improving the model's ability to capture complex relationships within the data. As a result, the model provides more reliable predictions for diabetes prevalence, leveraging both clinical data and Social Determinants of Health (SDOH) for early detection and intervention.

Mathematical Foundation	How It Fits SDOH and Diabetes Prediction
Mathematical foundation remains the same as XGBoost modeling but optimized with following 'hyperparameters' tuned, - learning rate: Controls step size during optimization. - max_depth: Limits tree depth to prevent overfitting.	- Fine-tuned hyperparameters improve generalization to unseen data. - Better captures interactions between SDOH factors (e.g., poverty × food access).

- subsample: Fraction of samples used for each tree. - colsample_bytree: Fraction of features used for each tree.	

Table 7 Optimized XGBoost Modeling

Overall Performance of the model



Optimized XGBoost Performance – Diabetes Prevalence Prediction
Context: Diabetes prevalence usually in single digits or in early teens in population

Figure 3 Optimized XGBoost - Scatter Plot

Model	MAE	MSE	RMSE	R²
Optimized XGBoost	0.63	0.62	0.79	0.88

Table 8 Optimized XGBoost Modeling Metrics

RESULTS/FINDINGS

Key Observations:

- Optimized XGBoost outperformed baseline XGBoost by 17% in MAE and 6% in R², which makes it ideal for batch processing of SDOH data.
- LightGBM was slightly less accurate but 2.5 times faster to train, making it suitable for real-time analytics in healthcare systems.

Model Performance Comparison:

Model	MAE	MSE	RMSE	R²
LightGBM	0.82	1.00	1.00	0.80
Baseline XGBoost	0.76	0.88	0.94	0.83
Optimized XGBoost	0.63	0.62	0.79	0.88

Table 9 Modeling Results Summary

IMPLICATIONS TO RESEARCH AND PRACTICE

PRACTICAL APPLICATIONS IN HEALTHCARE ANALYTICS

The integration of SDOH-aware machine learning models into healthcare analytics workflows enables proactive, data-driven decision-making at multiple levels:

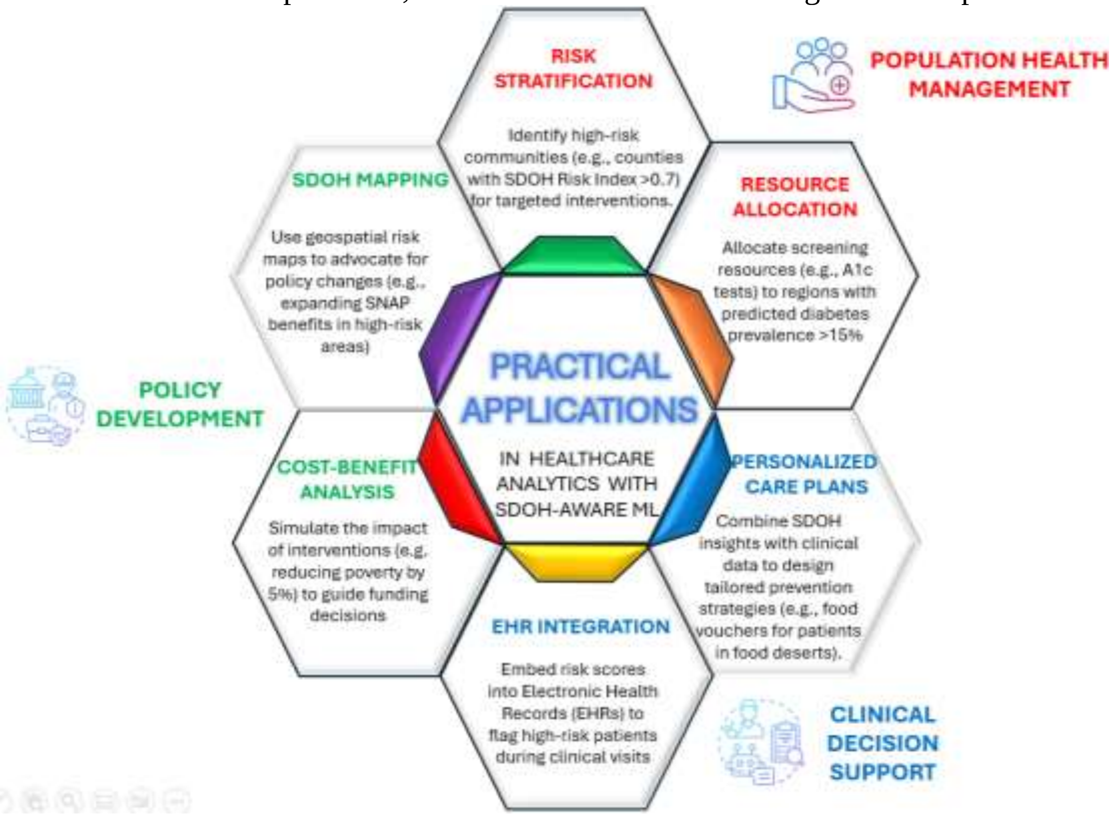

RISK STRATIFICATION
Identify high-risk communities (e.g., counties with SDOH Risk Index >0.7) for targeted interventions.
POPULATION HEALTH MANAGEMENT
SDOH MAPPING
Use geospatial risk maps to advocate for policy changes (e.g., expanding SNAP benefits in high-risk areas)
RESOURCE ALLOCATION
Allocate screening resources (e.g., A1c tests) to regions with predicted diabetes prevalence >15%
PRACTICAL APPLICATIONS
IN HEALTHCARE ANALYTICS WITH SDOH-AWARE ML
POLICY DEVELOPMENT
COST-BENEFIT ANALYSIS
Simulate the impact of interventions (e.g. reducing poverty by 5%) to guide funding decisions
PERSONALIZED CARE PLANS
Combine SDOH insights with clinical data to design tailored prevention strategies (e.g., food vouchers for patients in food deserts).
EHR INTEGRATION
Embed risk scores into Electronic Health Records (EHRs) to flag high-risk patients during clinical visits
CLINICAL DECISION SUPPORT

Figure 4 Practical Applications with SDOH & ML

INTEGRATION WITH EXISTING HEALTHCARE ANALYTICS FRAMEWORKS

We can integrate our model into existing healthcare analytics ecosystems through:

1. Real-Time APIs:
Deploy the model as a REST API for seamless integration with EHRs and population health platforms [12].
   - Example: A hospital system could query the API to generate risk scores for all patients in a region.

2. Interactive Dashboards:
Develop dashboards that visualize diabetes risk by county, SDOH factors, and intervention impact [13].
   - Example: A state health department could use the dashboard to monitor progress toward diabetes reduction goals.

3. Predictive Analytics Modules:

Add the model to existing healthcare analytics platforms (e.g. Epic, Cerner) as a predictive module.

- Example: A primary care provider could use the module to identify at-risk patients during routine visits.

CONCLUSION

This study shows the power of integrating Social Determinants of Health (SDOH) into machine learning driven healthcare analytics for early diabetes detection. By moving from biomarker-based models to a population level approach the framework improves prediction and informs proactive intervention. Optimized XGBoost performed best ($R^2 = 0.88$, MAE = 0.63) and is ready for real world use.

Beyond the technical performance the findings highlight the importance of community level insights for health disparities. Integrating this model into existing healthcare systems—through real-time APIs, dashboards and policy driven interventions—can make a big impact on chronic disease prevention.

Key takeaways:
Actionable Risk Stratification: Identify high risk counties for targeted screenings.
Policy Leverage: Advocate for SNAP expansions in food deserts and reduce diabetes.
Scalability: A transferable framework for other chronic diseases (e.g. hypertension, cardiovascular diseases).

Future work should focus on causal inference to isolate SDOH impacts and real-time data integration via satellite/SDOH insights. Collaboration with providers and policymakers will be key to translating these insights into equitable interventions. By marrying ML innovation with public health needs this work lays the foundation for a future where data driven prevention mitigates health disparities at scale.

To maximize impact, we recommend piloting the framework in state health departments, partnering with providers and continuously refining the model based on real world outcomes. By using data driven prevention we can drive real change in public health and chronic disease management [14].

REFERENCES

[1] American Diabetes Association. (2023) - Statistics About Diabetes. Retrieved from https://diabetes.org
[2] World Health Organization. (2021). Social Determinants of Health. Retrieved from https://www.who.int
[3] Drewnowski, A., & Specter, S. E. (2004). Poverty and obesity: the role of energy density and energy costs. The American Journal of Clinical Nutrition, 79(1), 6-16
[4] Centers for Disease Control and Prevention. (2023). County Health Rankings & Roadmaps. Retrieved from https://www.countyhealthrankings.org

[5] Centers for Disease Control and Prevention. (2023). Behavioral Risk Factor Surveillance System (BRFSS). Retrieved from https://www.cdc.gov/brfss

[6] USDA Economic Research Service (2023) - Food Access Research Atlas- Retrieved from https://www.ers.usda.gov

[7] Hill, J., Nielsen, M., & Fox, M. H. (2021). Understanding the Social Factors That Contribute to Diabetes: A Review of the Literature. American Journal of Preventive Medicine, 60(2), S12-S21.

[8] Gucciardi E, Vahabi M, Norris N, Del Monte JP, Farnum C. The Intersection between Food Insecurity and Diabetes: A Review. Curr Nutr Rep. 2014;3(4):324-332. doi: 10.1007/s13668-014-0104-4. PMID: 25383254; PMCID: PMC4218969.

[9] Braveman, P., & Gottlieb, L. (2014). The Social Determinants of Health: It's Time to Consider the Causes of the Causes. Public Health Reports, 129(Suppl 2), 19-31.

[10] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104-116.

[11] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001) Missing Value Estimation for DNA Microarrays. Bioinformatics, 17(6), 520–525. DOI: [10.1093/bioinformatics/17.6.520}

[12] Mandl, K. D., & Kohane, I. S. (2012). Escaping the EHR trap—the future of health IT. New England Journal of Medicine, 366(24), 2240-2242

[13] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the IEEE Symposium on Visual Languages

[14] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453