Visualizing Type 2 Diabetes Prevalence: Localizing Model Feature Impacts

Youssef Sultan [a,1,*], Mohammad Hammad [a,2], Kelly Lester [a,3]

[a] College of Computing, Georgia Institute of Technology, 801 Atlantic Dr NW, Atlanta, 30332, GA, United States.
[1] ysultan@gatech.edu; [2] mhammad8@gatech.edu; [3] klester@gatech.edu
* corresponding author

ARTICLE INFO

Article history
Received July 27, 2024
Revised November 17, 2024
Accepted December 31, 2024

Keywords
geospatial data analysis
health disparities
predictive modeling in healthcare
spatial epidemiology

ABSTRACT

SHAP values have been a common approach used to understand machine learn- ing model predictions by averaging the marginal contributions of each feature across every possible permutation of the feature set. Our research provides a localized view of SHAP values contributing to Type 2 Diabetes (T2D) prevalence in the United States from 2012 - 2021 covering each year independently. Instead of visualizing SHAP feature importance across an entire geographical dataset using a beeswarm plot, our approach is more granular. We visualize individual SHAP values of Social Determinants of Health (SDOH) features by county on a Choropleth map. Additionally, we found that replacing geographic identifiers such as zipcode with precise latitude and longitude coordinates before applying KNN imputation reduced the MSE by 10%. Our visualization reveals how specific factors influence T2D prevalence at the county level using a non-linear machine learning model. By re-appending the initially preserved geographic identifiers for each record by index, we traced the contribution of each SHAP value back to its locality. Our approach opens up a new geographical vantage point of the mecha- nisms of model predictions, thereby identifying localized key factors influencing Type 2 Diabetes (T2D). This study extends the possibilities for tailored interven- tions and public health policies showing how some factors have varying predictive impact on an outcome at the geographic level.

1. Introduction

Type 2 Diabetes (T2D) remains a leading cause of morbidity and mortality in the United States, particularly affecting historically underserved communities. The prevalence of T2D is intricately linked with various Social Determinants of Health (SDOH), which include the conditions under which people are born, grow, live, work, and age. Leveraging a decade of data from the Diabetes Atlas, this study expands upon traditional factors such as food insecurity to incorporate a broader spectrum of SDOH. The work of Haire-Josh and Hill-Briggs [1] highlights how inequities in living and working conditions significantly impact the biological and behavioral outcomes associated with diabetes prevention and control, substantiating our focus on identifying and analyzing the most influential SDOH factors visually at a more granular level.

Existing visualizations lack granularity, not capturing the nuanced interplay between T2D prevalence and SDOH factors at the state and county levels from a machine learning perspective. Previous works [2] have extracted spatial effects of SHAP values, however this framework does not present a solution

to the issue of missing values within a county-level and longitudinal survey-based dataset [3]. SHAP value aggregation of machine learning models allow for compiling feature impacts in the form of a value for each feature in the model. By training on the social determinants of health data after a notably improved imputation method within the United States with diabetes prevalence being the dependent variable, the focus is on understanding how certain determinants or features have higher or lower impacts to the predicted model output at county level as opposed to overall in the US.

This method of visualization can be compared with real world data or literature to assess its efficacy for other future modalities in the field of health or other domains. Here the use of machine learning is not only used to generalize to unseen data, but to mainly understand how they make predictions for the purpose of insights and analytics. This research provides an interactive map that overlays the most impactful SDOH factors contributing to T2D prevalence in the model. For each county the ordering of top features is based on the highest mean absolute value. Empirical postulations about diabetic outcomes are not the focus of this paper especially given the data is not absolute; the concept of visualizing granular impacts however, with this or similar datasets are of importance. To see changes through out time there is one model for each year with SHAP values and geographic identifiers preserved for each year by index for mapping.

We utilize KNN imputation to enhance visualization by incorporating latitude and longitude to the compiled dataset, filling data gaps for certain counties and years of various features. The Euclidean distance during KNN imputation [4] captures the smallest spatial distance of similar data points for each missing feature within a data point to patch missing gaps. In basic terms, a row within the dataset can be referred to as a data point. Our approach adds in the latitude and longitude as continuous features to incorporate the uniqueness of each data point based on locality before KNN imputation. Filling these gaps before modeling provide a fuller dataset despite the fact that certain counties may not have certain SDOH factors reported in a specific year.

This framework can provide policymakers, public health professionals, and community organizations with a visual understanding of T2D prevalence and its determinants from a machine learning perspective. By helping targeted interventions and resource allocation at the county and state levels, our research can provide other ways of applying machine learning to reduce health disparities related to T2D or other outcomes. The results of the visual can be cross-validated with real world data [5] to see if the feature impacts towards the outcome make sense beyond performance metrics like mean-squared-error and r-squared [6]. Model feature impacts can be used for targeted interventions to improve T2D outcomes at the geographic and temporal level (year by year).

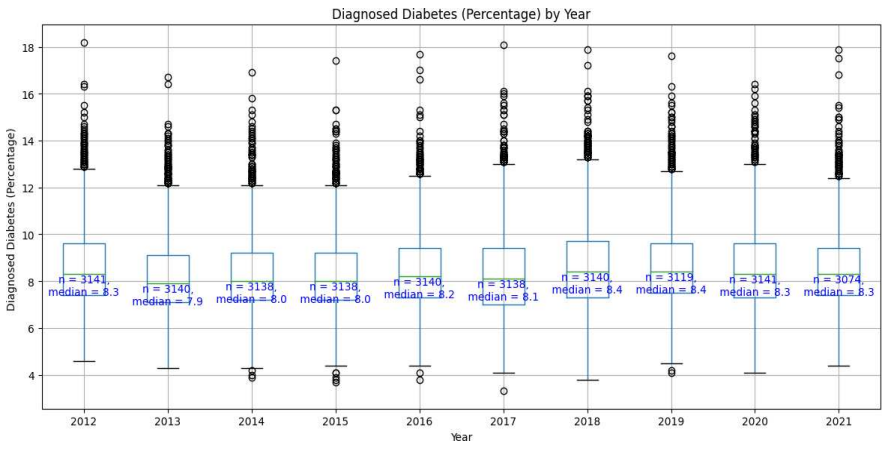

Figure 1. Diabetes prevalence in the United States from 2012-2021 (n = # of counties/records)

2. Materials and Methods

2.1 Data Collection

The first step in finding the most impactful SDOH factors involves comprehensive data collection. This process included merging the NHANES dataset [7], which has food security and demographic data, with the CDC dataset [8] on Diabetes. These datasets are publicly available through government agencies and provide multi-year information at the county level. Due to the complexity of importing all indicators simultaneously, data was collected for each indicator separately from 2012 to 2021.

Diabetes prevalence in the United States by year can be found in Figure 1. A similar plot for obesity prevalence which is used as an independent variable can be viewed in Appendix C. Each indicator dataset was saved as a separate file for each year. Subsequently, the datasets for each indicator and year were joined on county, state and zip to create a dataset stacked per year with all of the features. This process enabled the consolidation of data for various indicators and years into a single dataset, thereby helping analysis. A free API [9] was also used to acquire latitude and longitude as a replacement for the county, state and zip for later imputation.

2.2 Data Cleaning

The merged dataset originally has 53 columns which can be found in Appendix A. Initially, the zip codes were standardized to five digits by padding shorter codes with leading zeros, enhancing consistency across the dataset. Subsequently, records lacking essential information, such as diagnosed diabetes and obesity percentages, were removed, as these are key variables for the analysis and do not contribute to the analysis of diabetes prevalence. Additionally, the dataset underwent type conversions, transforming the diabetes and obesity percentages from string to float, to facilitate numerical operations. We also review the distribution of missing data for each year within the dataset, helping to identify any trends or discrepancies in data availabil- ity over time. For each year the distribution of missingness was found to be similar. As shown in Figure 2 it can be seen that some features have more than half of their records missing, because of this we drop features missing above 48% of their values for the purpose of not inducing additional bias into later imputation.

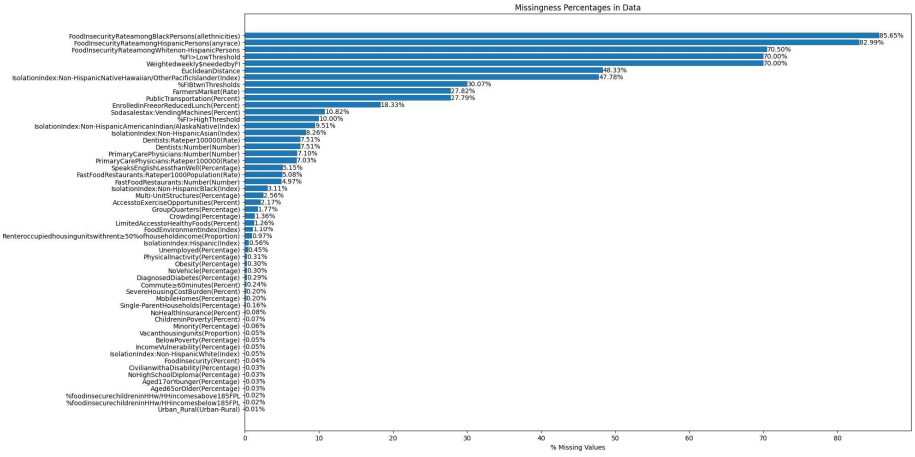

Figure 2. Distribution of missing values overall (all years)

2.3 Data Imputation

Within the compiled dataset, there are cases where certain counties or localities may not have data for a specific feature or determinant. KNN imputation was decided to be used over mean or median imputation based off of previous works [10] as it was shown to outperform other methods. To ensure the most accurate value is imputed within missing data in a data point in the context of locality, we replace zipcode, county and state with latitude and longitude. Now before the euclidean distance is calculated, this allows for distance search to account for the variance of the county specific magnitude

and impute a more accurate value. This process ensures there is no data loss when modeling or visualizing instead of dropping the data points as a whole. Dropping records with nulls would bring the overall dataset record count from 31,000 records to 8,000 which would result in major data loss across various counties.

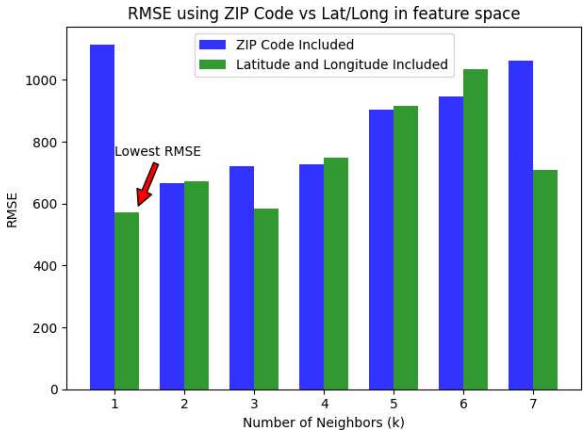

Figure 3. Root mean-squared error of KNN imputation; zip code vs latitude and longitude.

Incorporating the magnitude of the locality (latitude and longitude columns) in the dataset before KNN imputation showed to have lower error as shown in Figure 3. The imputed values when including zipcode as a feature before imputation has a higher error against the true values, even at different levels of k neighbors tested from k=1 to k=7.

2.3.1 Conducting the Experiment

During KNN imputation, we saw how the accuracy of missing value estimation signif- icantly benefits from the precise representation of geographical data. To demonstrate this, we show how to reproduce the comparisons: one using ZIP code as a continuous variable and another using latitude and longitude as continuous variables, which allow for understanding how these results were formulated. Consider a simplified example where we have a feature set from 1 to n with m records, with additional location variables:

Data with ZIP Code:

$$\begin{bmatrix} \text{Feature}_1 & \text{Feature}_2 & \dots & \text{Feature}_n & \text{ZIP Code} \\ 1.94 & 2.32 & \dots & 0.11 & 6003 \\ 0.83 & 0.05 & \dots & 0.34 & 46102 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m & m & \dots & m & 2066 \end{bmatrix}$$

Data with Latitude and Longitude:

$$\begin{bmatrix} \text{Feature}_1 & \text{Feature}_2 & \dots & \text{Feature}_n & \text{Latitude} & \text{Longitude} \\ 1.94 & 2.32 & \dots & 0.11 & 38.5893 & -119.8345 \\ 0.83 & 0.05 & \dots & 0.34 & 43.2977 & -102.5637 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ m & m & \dots & m & 61.8063 & -143.9721 \end{bmatrix}$$

Experiment 1 KNN Imputation for different k values with zipcode Input: Initial data with features and zipcode without dependent variable

1: Initialize list to store RMSE for each k
2: for k = 1 to 7 do
3: Remove all rows containing a null value in any column from data
4: Drop unnecessary columns, retain only ZIP code and other continuous features
5: Introduce random missingness using a binomial distribution (30%) 6: Store original values replaced with NaN in a list in ascending order 7: Apply MinMaxScaler to normalize data
8: Use KNN imputer with parameter k to fill missing values
9: Inverse scale data back to original scale
10: Store imputed values in a list in ascending order
11: Calculate RMSE between original values and imputed values
12: Append RMSE for each k to initialized list
13: end for

Experiment 2 KNN Imputation for different k values with latitude and longitude Input: Initial data with features, latitude and longitude without dependent variable

1: Initialize list to store RMSE for each k
2: for k = 1 to 7 do
3: Remove all rows containing a null value in any column from data
4: Drop unnecessary columns, retain only latitude and longitude and other continuous features
5: Introduce random missingness using a binomial distribution (30%) 6: Store original values replaced with NaN in a list in ascending order 7: Apply MinMaxScaler to normalize data
8: Use KNN imputer with parameter k to fill missing values
9: Inverse scale data back to original scale
10: Store imputed values in a list in ascending order
11: Calculate RMSE between original values and imputed values
12: Append RMSE for each k to initialized list
13: end for

After conducting both experiments on full data with true values and inducing the random missingness, for both methods the RMSE's can be compared. The records removed to have full data are just to show results on a full dataset with true values; to assess the estimation error inducing random missingness. By understanding that KNN imputer performs better with latitude and longitude, in this case where k=1, we can conclude that using KNN imputer on our full dataset with latitude and longitude would be the most accurate over zip code.

2.4 Feature Selection

After imputation one Linear Regression model per year is fit to the data with diabetes prevalence as the target outcome. This is done to understand whether the data fits the assumptions of a linear model, to see if any potential inference from coefficients would be adequate for visualization before nonlinear models. It was found that for all years, the residuals vs fitted values were not randomly scattered and residuals were not normal. A correlation analysis was also conducted via a heatmap as shown in Figure 4.

Comparisons of all correlations for each year showed similar correlations between features. Then a VIF calculation for all features for each year was conducted to analyze features with a VIF greater than 10. For those features with a high VIF having correlation with others $\geq |0.5|$ the feature with the lesser variance out of the pair is dropped and the next was kept. This left us with a reduced feature set of 33, not including the latitude and longitude as they would be dropped during model training.

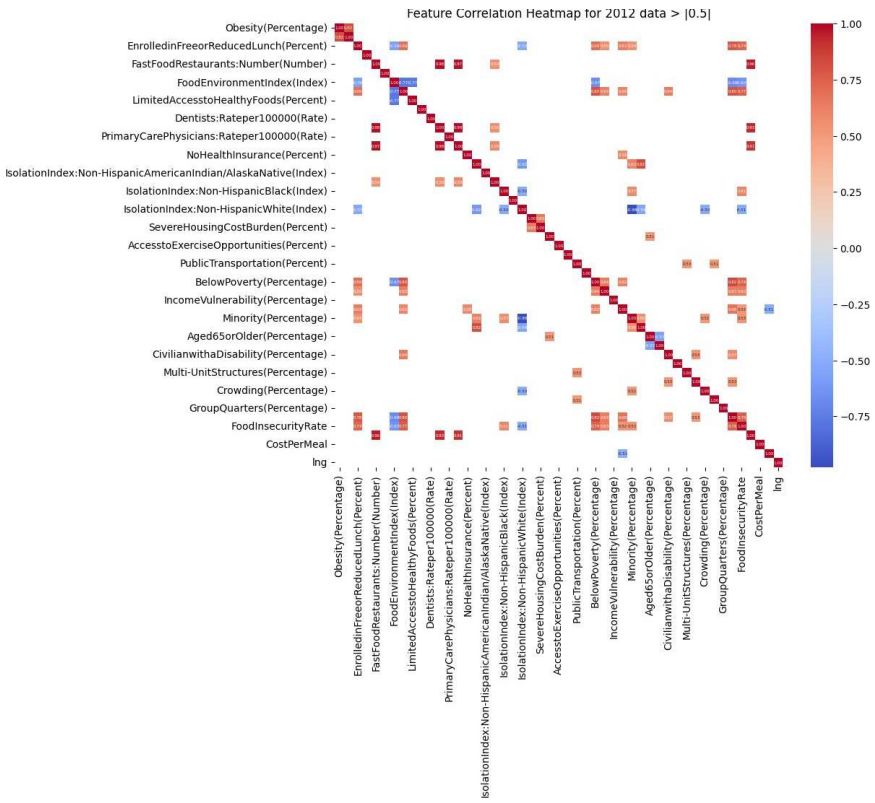

Figure 4. 2012 Data: Features with correlation ≥ |0.5|

3. Results and Discussion

3.1 Finding The Optimal Model

Multiple non-linear models were evaluated to achieve the best r-squared and lowest rmse predicting diabetes prevalence. The main goal here is not only to have the best performing model, but to fundamentally conceptualize how one can interpret SHAP values geographically. The exploratory data analysis (EDA) revealed that the data did not adhere to the required linear relationship between the factors and the response variable before. Consequently, alternative methods were employed. The data was split into 80% train and 20% test datasets. Grid search with cross validation was used to find optimal hyperparameters for different models. XGBoost was selected as the main model of interest to calculate SHAP values from; model metrics for each can be viewed in Table 1. For each year, each model was ran on training sets of that specific year since the visual will be visualizing impacts at the yearly level. This yields for a total of 50 runs, not accounting for cross-validation folds and hyper parameter tuning.

Table 1. Model mean-squared error across different years

Model	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
RandomForest	1.0771	0.8653	0.8779	0.9991	0.8919	1.4400	1.0453	1.0720	0.9687	0.9034
GradientBoost	0.9445	0.8598	0.8096	1.0082	0.8512	1.3516	1.0166	1.0431	0.9390	0.8974
SVR	2.6191	2.3357	2.0961	2.0453	2.2005	3.3289	2.5296	2.5070	2.3537	2.2493
XGBoost	0.9454	0.8582	0.8001	1.0120	0.8620	1.3533	1.0070	1.0364	0.9323	0.9040

The r-squared of the best model had a median of 0.7, indicating that the fea- tures potentially explained 70% of the variance represented in the outcome. Given the foundation of the data itself, we would not receive better results given that the distribution of diabetes prevalence is quite similar year over year, with outliers being actual collected values and less prominent. This can be seen in

Figure 5, where the other unknown 30% of variability the features could not explain could be those data points outside of the highest frequency of prevalence.

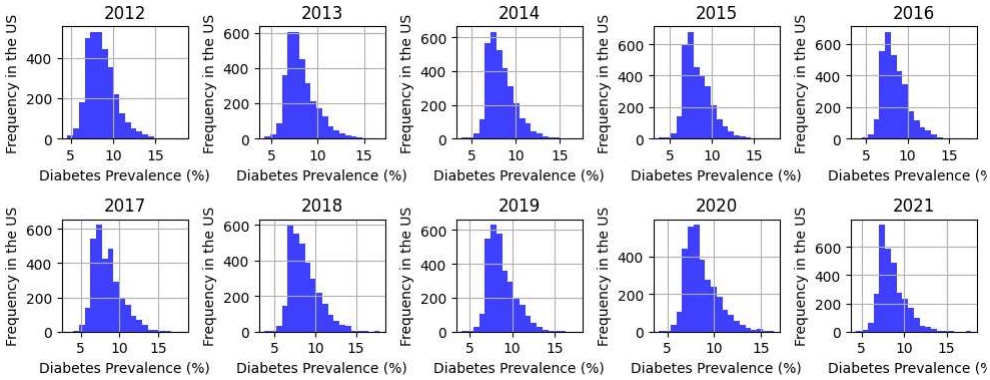

Figure 5. Distribution of diabetes prevalence percentages by year showing skewness > 0

3.2 Localizing SHAP Values

After finding the optimal model parameters, the model is trained on the full dataset and passed through the SHAP explainer to calculate SHAP values on the full seen data. The reason for this is because we want predictions for each record in the dataset, to then have a matrix of the same size as the dataset but with SHAP values for each feature. The idea here is to understand how the model creates predictions on the all of the data so it can be visualized. If we were to only predict on the test set, we would only be able to visualize the test set, which would not show impacts for all counties. Since we preserved the index of the original dataset with the county, zip and state, we can concatenate these columns back to the SHAP value dataset, to then plot the SHAP values by county.

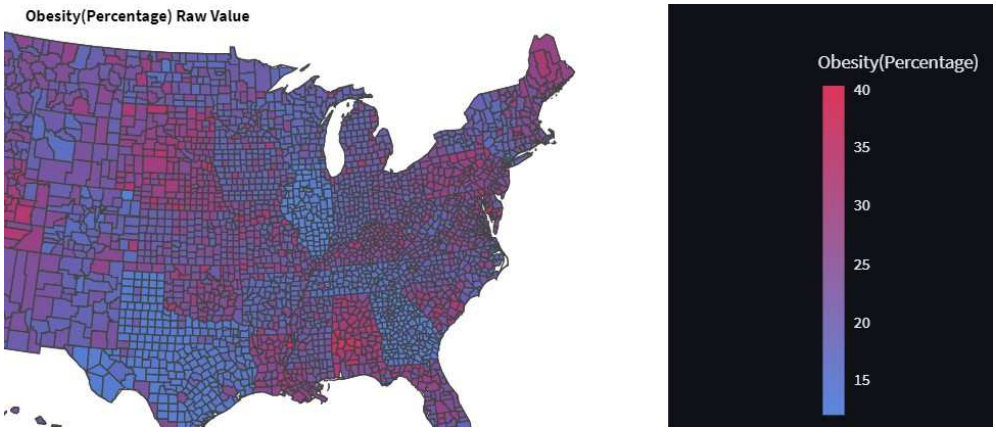

Figure 6. 2012: Obesity Percentage prevalence plotted without any modeling

In Figure 6 we can view the raw values of the Obesity prevalence percentage before any modeling by county to view the overall density via a choropleth map. This is how the data is generally visualized on the CDC website. This allows us to compare to Figure 7; this shows the same feature however its SHAP value impacts towards diabetes prevalence by county. For each instance in each data point there is a SHAP value and this is plotted with the reattached geographical information known for that index position in the matrix.

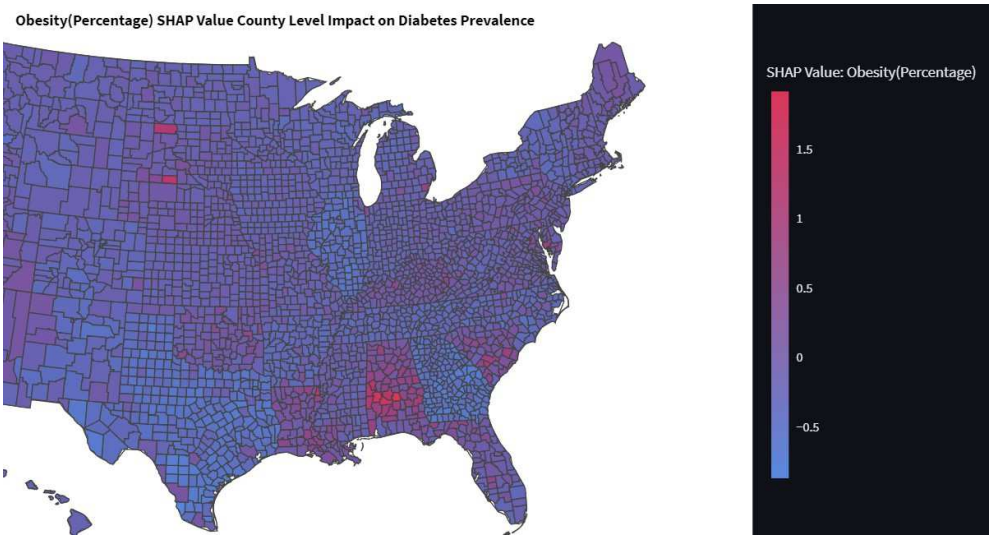

Figure 7. 2012: Obesity Percentage prevalence impact on diabetes prevalence prediction

In Figure 7 we can see that the Obesity percentage has a higher impact on the predicted model output in Alabama, with certain counties having a higher SHAP value (red) than others. This highlights counties where Obesity prevalence rates are predictors of higher diabetes prevalence as opposed to others.

This allows us to understand the features that contribute to the model prediction at the geographic level; to understand whether a specific determinant has a higher contribution to diabetes prevalence in some counties rather than others. This approach allows to see the granular contributions, positive or negative towards the outcome. Positive SHAP values indicate that the feature positively impacts the outcome, while negative SHAP values indicate the opposite. This can be identified through the color scheme in the legend of the visual, this is the same color scheme which is used by the original SHAP beeswarm plot visualization.

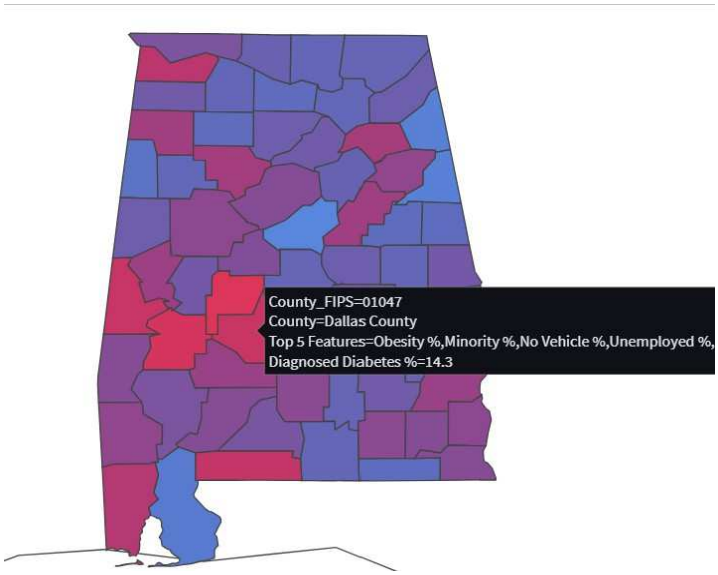

Figure 8. 2014 Data: Dallas County, Alabama Diabetes Prevalence

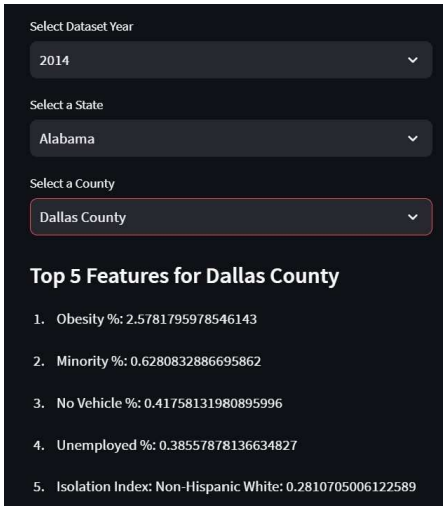

Figure 9. 2014 Data: Top features contributing to Diabetes Prevalence in Dallas County

If we take a look at another example: Dallas County's top features are based on the top absolute values, we can see in Figure 8 (color is based on diabetes prevalence in a legend not shown) the top five features contributing to the outcome. In Figure 9 the SHAP value for the minority percentage feature is 0.62 and second most important, this indicates that as the minority percentage increases, the diabetes prevalence increases in Dallas County, Alabama in 2014. These granular details are more nuanced than an overall holistic view of feature impact from the dataset. An example of the 2014 data's top feature impacts based on the default beeswarm plot can be viewed in Figure 10.

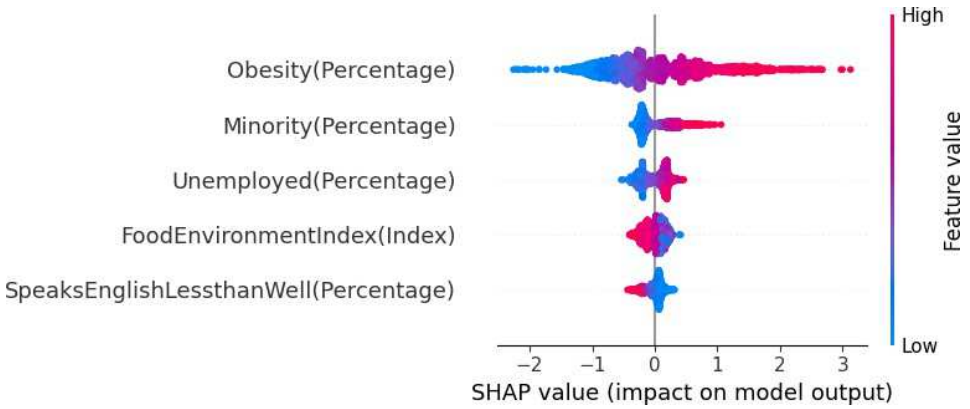

Figure 10. 2014 Data: Beeswarm Plot of Top 5 Features contributing to Diabetes Prevalence in overall data

There are clear differences in the top feature impacts between the overall beeswarm plot and the granular geographic views of the impacts. The beeswarm plot calculates the mean absolute value of the entire feature vector for each feature, while our visual provides the SHAP value for the feature in the specific row which is attributed to the selected county. Understanding the direct impacts at the data point/row level allows us to understand why the model is creating such predictions for those specific counties. It can allow us to understand if more research should be conducted around a specific county for tailored targeting. By retaining each training data point's county, zip and state information we can be able to pinpoint the impact using these details as a label, localizing the SHAP impacts to assess the results more closely.

4. Conclusion

In this study we showed the difference between holistic and localized feature importance in a non-linear machine learning model. In addition, we provided a more accurate way of imputing missing

values of certain determinants by adding their associated Latitude and Longitude for each record before KNN imputation. This method improved on the RMSE against the true values as opposed to using zip code. The impacts of social determinants of health on Type 2 Diabetes (T2D) prevalence at a localized county level were visually different than raw social determinant values plotted in the choropleth map. Mean absolute SHAP values from the overall dataset (year) were also different than county level SHAP values, indicating differing top contributors to diabetes prevalence by locality.

Our visualization can provide a deeper understanding of the data since not all counties have the same SHAP values; some are more affected by other features. This could suggest how specific policies in T2D prevention cannot be ubiquitous. For instance, a national policy on Obesity might have the best chance at lowering T2D overall but could still not be effective in certain regions in the United States. These methods can be performed on state, zip code, city or county-level data.

The interpretation of SHAP values does depend heavily on the model's accuracy and the quality of the data used. Future research should focus on incorporating more comprehensive datasets that include additional variables potentially affecting T2D or other outcomes that may not relate to health. Results should be cross-referenced with real-world data, primary or secondary sources to make empirical claims or confirm certain hypotheses using these methods.

A different vantage point of the underlying mechanisms influencing these predictions is established, which can allow for better targeted and effective public health strategies. By enhancing model accuracy and incorporating a broader range of determinants, we can understand and address the complex interplay of factors contributing to T2D and other health outcomes.

Authors Contribution

Mohammad Hammad worked on curating the dataset collecting individual contribu- tors from NHANES and CDC, exploratory data analysis and software engineering of the visual features. Youssef Sultan worked on the statistical analysis, methodology, experiments, software engineering of the visual and writing of this manuscript. Kelly worked on sourcing the data and acquiring necessary approval for its use for publishing and research purposes.

References

[1] Debra Haire-Joshu and Felicia Hill-Briggs. The next generation of diabetes trans- lation: A path to health equity. Annual Review of Public Health, 40:391–410, 2019.

[2] Ziqi Li. Extracting spatial effects from machine learning model using local inter- pretation method: An example of shap and xgboost. Computers, Environment and Urban Systems, 96:101845, 2022.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.

[4] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: A short walk through theory, algorithms and applications. CoRR, abs/1502.07541, 2015.

[5] Fang Liu and Demosthenes Panagiotakos. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology, 22, 11 2022.

[6] Steven A Hicks, Inga Stru¨mke, Vajira Thambawita, Malek Hammou, Michael A Riegler, P˚al Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. Scientific reports, 12(1):5979, 2022.

[7] Feeding America. Hunger and poverty in the united states — map the meal gap, 2024.

[8] Centers for Disease Control and Prevention. United states diabetes surveillance system, 2024.

[9] OpenStreetMap contributors. Openstreetmap, 2024. Open Database License.

[10] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of per- formance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10):913–933, 2019.

Appendix A

1. Appendix A Columns in dataset before analysis

Features in compiled dataset			
euclidean_distance	iso_idx_native_hawaiian	public_transportation_%	farmers_market_rate
reduced_lunch_%	soda_sales_tax%	iso_idx_american_indian	iso_idx_asian
dentists_rate_per_100k	dentists_number	physicians_number	physicians_per_100k
deficient_in_english_%	fast_food_proportion	fast_food_number	iso_idx_black
multi_unit_structures_%	exercise_access_%	group_quarters_%	crowding_%
no_healthy_foods_%	food_environment_idx	*rent* ≥ 50% *_income*	*iso_idx_hispanic*
unemployed_%	no_vehicle_%	commute_60_minutes_%	housing_cost_burden_%
mobile_homes_%	single_parent_home_%	no_health_insurance_%	iso_idx_white
vacant_housing_units	below_poverty_%	income_vulnerability_%	minority_%
children_in_poverty_%	food_insecurity_%	no_hs_diploma_%	65_or_older_%
17_or_younger_%	disabled_civilian_%	diabetes_prevalence_%	obesity_%
physical_inactivity_%	urban_rural	food_insecurity_rate	# food_insecure_children
cost_per_meal	zipcode	state	year

Table A1 Features compiled from the Diabetes Atlas Data and the Food Insecurity Data. Features in red were not included in the model due to high VIF and multicollinearity. They are the part of the pair that has lower variance.

2. Appendix B Obesity prevalence yearly

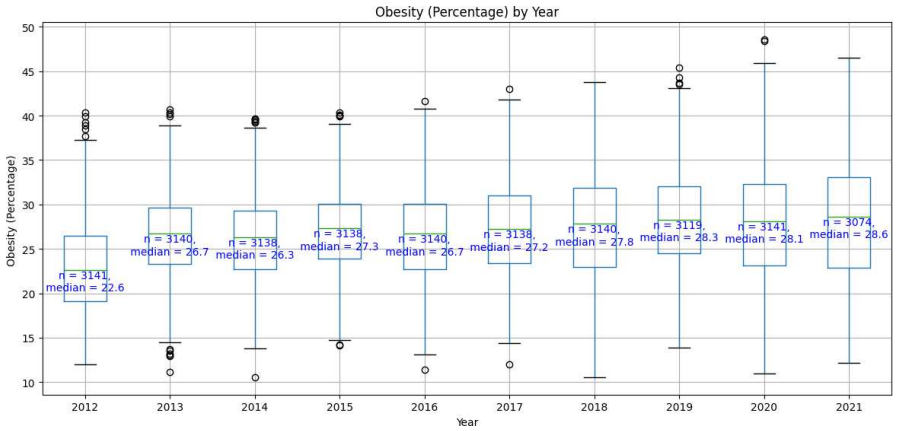

Figure B1. Obesity prevalence in the United States from 2012-2021