PPOL 5204 Data Science Final Report

Assessing the Impact of Social Determinants on Obesity Rates Across U.S. Counties: Machine Learning Analysis and Policy Recommendations

Introduction

Obesity represents a critical public health challenge in the United States, affecting a substantial proportion of the population across various age groups and demographics. The urgency of addressing this issue stems from its extensive impact on health, economic costs, and broader societal implications. According to CDC (2022), approximately 20% of children and 33% of adults are struggling with obesity. It also causes an economic burden, with $147 billion spent annually on obesity-related healthcare. Moreover, a significant portion of Americans, both young and adult, do not meet the recommended levels of physical activity and dietary guidelines. The accessibility of healthy food and recreational spaces is also problematic, as 40% of U.S. households are not within one mile of a healthier food retailer, and more than half of the population does not have easy access to parks. These factors point to the multifaceted nature of obesity and the need for comprehensive strategies to address it.

Literature Review

The literature on obesity and machine learning encompasses a variety of factors and methods for predicting and understanding the condition.

Maternal habits and conditions, such as smoking during pregnancy and high BMI, have been implicated in the predisposition of offspring to obesity, suggesting the importance of maternal health in obesity prevention (Cheng, 2020). Neurological functioning and even personality traits, possibly affecting dietary and physical activity behaviors, are also considered significant factors (Choukem et al., 2020). Health factors such as psychological distress have been correlated with obesity, suggesting a link between mental health and weight management (Keramat et al., 2021). Sociodemographic characteristics like age, sex, marital status, and ethnicity offer a demographic lens through which obesity can be studied, pointing to the tailored public health strategies required for different population segments (Sun et al., 2020). Singh & Tawfik (2019) utilized multi-layer perceptron feedforward neural networks, focusing on Body Mass Index (BMI) as a primary risk factor. Some scholars employed a combination of machine learning techniques, including Support Vector Machine (SVM) and Decision Trees, to assess body fat percentage as a crucial indicator of obesity (Uçar et al.,2021). Some scholars adopted Artificial Neural Networks, reinforcing the relevance of BMI in obesity prediction models (Fergus et al., 2015). Pleuss et al.(2019) employed deep convolutional neural networks to analyze food and energy intake from images, highlighting innovative approaches in combating obesity. Logistic Regression and KNN methods were used by Farran et al(2019). to associate BMI and type 2 diabetes with obesity, showing a pattern of linking obesity to other health conditions.

My research aims to make a novel contribution to the field by utilizing the CDC's comprehensive social determinants dataset to predict county-level obesity rates. This approach acknowledges the multifaceted nature of obesity, as established by the diverse range of factors identified in prior studies. Unlike previous research which often focused on individual-level predictors or single aspects of the social environment, my work will leverage a broader set of community-level indicators provided by the CDC, which may include access to healthcare, education, economic stability, neighborhood and built environment, as well as social and community context. This research is poised to fill a significant gap in existing literature by providing insights on how community-level factors influence obesity. It will also evaluate the predictive power of various machine learning models in a public health context, thereby informing effective, data-driven policy-making and intervention strategies.

Research Goal

The World Health Organization defines social determinants of health as the conditions in which people are born, grow, live, work, and age. These circumstances are shaped by the distribution of money, power, and resources at global, national, and local levels. For obesity, social determinants can include factors like income inequality, education level, neighborhood characteristics, access to healthcare, and availability of healthy foods and recreational facilities. Understanding how these determinants influence obesity prevalence is crucial for developing targeted, effective public health interventions. My project aims to use machine learning instruments to predict the obesity rate in the US at the county level. This comprehensive methodology not only seeks to predict and analyze obesity prevalence but also to inform and guide policymakers toward more effec've, sustainable public health strategies.

DATA

1. **Data Resources Updates:**
   https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html

2. **Data Selection**

- **Prediction Target**
  In the United States Diabetes Surveillance System, I selected **Obesity (Percentage) 2021** as my health outcome, which is my prediction target.

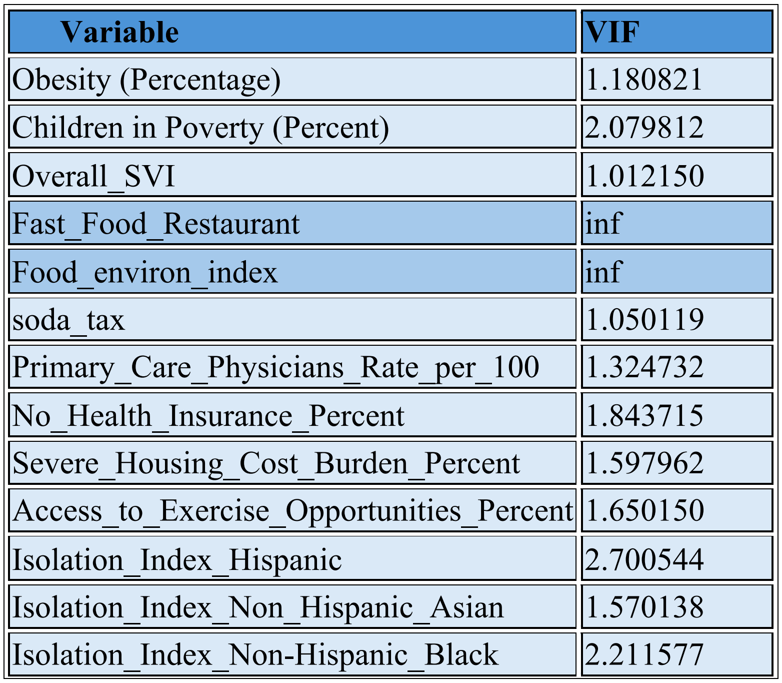
  For the selection of predictors, I considered multiple determinants, encompassing areas of economics, food environment, healthcare, neighborhood dynamics, physical environment, transportation, the Social Vulnerability Index, and rural-urban categories.
- **Economics:**
  - Children in Poverty: Children in poverty are more likely to experience nutritional deficiencies and have limited access to healthy foods, which can lead to obesity.
- **Food Environment:**
  - Fast Food Restaurants Rate: A higher density of fast food restaurants is often correlated with higher obesity rates due to increased consumption of high-calorie, processed foods.

    - Soda Sales Tax: Taxes on sodas can decrease the consumption of sugary drinks, which are linked to higher obesity rates.
- **Healthcare:**
    - Primary Care Physicians Rate: Better access to healthcare can facilitate preventive care and management of obesity.
    - No Health Insurance: Lack of insurance can lead to underutilization of healthcare services, hindering obesity prevention and treatment efforts.
- **Physical Environment:**
    - Access to Exercise Opportunities: Higher access usually correlates with lower obesity rates as it facilitates physical activity.
- **Housing-Neighborhood:**
    - Severe Housing Cost Burden: Similar to the above, it can force families to prioritize shelter costs over health-related expenses, affecting obesity rates.
    - Isolation_Index_Hispanic: This index measures the likelihood that a Hispanic individual will interact or come into contact with another Hispanic individual within the community. A high isolation index for Hispanics indicates that Hispanics are more likely to be living among other Hispanics, rather than in a diverse community.
    - Isolation_Index_Non_Hispanic_Asian: This index indicates the extent to which non-Hispanic Asians are isolated from other groups, that is, how often they are likely to only interact with individuals from their own ethnic group in their community.
    - Isolation_Index_Non-Hispanic_Black: Similarly, this index measures the degree to which non-Hispanic Black individuals are likely to live among and interact with other non-Hispanic Black individuals, as opposed to a more integrated environment.
    - Isolation_Index_Non-Hispanic_White: This index calculates the probability that non-Hispanic White individuals will only be around other non-Hispanic White individuals in their daily lives within their community.
- **Urban-Rural Categories:**
    - The urban-rural divide often affects obesity rates through differences in lifestyle, food access, and physical activity opportunities.
- **Social Vulnerability Index (SVI):** The SVI reflects the resilience of communities when faced with external stresses. Communities with higher vulnerability may have more difficulty accessing health-promoting resources, thus potentially having higher obesity rates.

3. Data Preprocessing

1) **Data Merging:** Initially, it's imperative to merge 15 distinct datasets that were directly downloaded from the website. Each dataset is parsed into separate files, organized by county name and state. This step is crucial for facilitating subsequent analyses.
2) **Addressing "No Data" Entries**: There are some missing values stored in the data set as "string". I transformed all instances of these textual representations of missing values into NaN (Not a Number) values.

3) **Soda Tax Imputing**: A significant number of missing values are present in the 'soda_tax' feature, which varies by state, not missed randomly. To address this, I manually filled in the missing values by referencing soda tax information for each state from the Tax Foundation.
4) **Isolation Index**: The 'isolation index' measures the extent to which members of a minority or specific demographic group are exposed exclusively to their group, indicating the degree of social, spatial, or economic isolation from other groups. To address missing values in this crucial measure, I utilized the K-Nearest Neighbors (KNN) imputation technique. KNN imputation is particularly appropriate for the isolation index because it preserves the intrinsic relationships within the data by imputing values based on the similarity of the nearest neighbors in the multidimensional feature space.
5) **Encoding the "Urban-Rural" feature**: This is a categorical varaible. Converting this variable into dummy variables is a strategic approach to simplify the analysis, enhancing the interpretability and processing of categorical data within the analytical model.
6) **Drop some features**: In the process of preparing the dataset for analysis, I decided to streamline the data by removing certain features. Specifically, I dropped the 'Year', 'County_FIPS', 'County', and 'State' columns. The rationale behind dropping 'Year' is that all records pertain to the same year, making this attribute redundant for comparative or time-series analysis. As for the geographical identifiers such as 'County_FIPS', 'County', and 'State', these were removed to simplify the initial phases of the analysis focused on non-geographical factors. However, these geographical attributes will be reintegrated into the dataset later.
7) **Screening features by using VIF**:



Variable	VIF
Obesity (Percentage)	1.180821
Children in Poverty (Percent)	2.079812
Overall_SVI	1.012150
Fast_Food_Restaurant	inf
Food_environ_index	inf
soda_tax	1.050119
Primary_Care_Physicians_Rate_per_100	1.324732
No_Health_Insurance_Percent	1.843715
Severe_Housing_Cost_Burden_Percent	1.597962
Access_to_Exercise_Opportunities_Percent	1.650150
Isolation_Index_Hispanic	2.700544
Isolation_Index_Non_Hispanic_Asian	1.570138
Isolation_Index_Non-Hispanic_Black	2.211577

Variable	VIF
Isolation_Index_Non-Hispanic_White	3.559198
rural_urban_encoded	1.432577

In the course of analyzing multicollinearity within our dataset, I employed the Variance Inflation Factor (VIF) as a diagnostic tool. Notably, both the `Fast_Food_Restaurant` and `Food_environ_index` variables exhibited infinite VIF values, indicating perfect multicollinearity. This suggests that these two variables share a linear dependency with other explanatory variables in the model, potentially distorting the analysis. To address this issue and improve the model's reliability, I decided to remove the `Food_environ_index` variable from model.

Exploratory Data Analysis

1. **Descriptive Statistics**
   The dataset encompasses 3,065 observations for each variable, with the 'const' acting as an intercept with consistent values of 1 across the board. 'Obesity (Percentage)' has a mean of 28.17% with a standard deviation of 6.52, indicating variability, while 'Children in Poverty (Percent)' averages at 20% with an 8.43 standard deviation. The 'Overall_SVI' is skewed by an outlier, suggested by a -999 minimum, diverging from its median of 0.49. Both 'Fast_Food_Restaurant' and 'Food_environ_index' show an identical distribution with a mean of 0.62 and a median of 0.60. 'Soda_tax' has an average value of 4.71, with half of the counties levying a tax greater than 5.50. 'Primary_Care_Physicians_Rate_per_100' and 'No_Health_Insurance_Percent' demonstrate significant spread in their values, with the former potentially influenced by an outlier at 576.43. 'Severe_Housing_Cost_Burden_Percent' and 'Access_to_Exercise_Opportunities_Percent' suggest a moderate burden and good access to exercise opportunities, respectively. The 'Isolation_Index' for various ethnic groups indicates distinct degrees of isolation, with 'Isolation_Index_Non-Hispanic_White' having the highest average of 0.79. Finally, the 'rural_urban_encoded' variable, ranging from 0 to 1, suggests a binary nature, with a mean close to being evenly split at 0.43. This descriptive analysis provides a quantitative synopsis of the data's distribution, central tendencies, and variability.

variable	count	mean	std	min	50%	max
Obesity (Percentage)	3065	28.17	6.52	12.20	28.60	46.50
Children in Poverty (Percent)	3065	20.00	8.43	2.80	19.00	58.50
Overall_SVI	3065	0.17	18.06	-999.00	0.49	1.00
Fast_Food_Restaurant	3065	0.62	0.28	0.04	0.60	5.81
soda_tax	3065	4.71	2.00	0.00	5.50	7.00
Primary_Care_Physicians_Rate_per_100	3065	55.24	34.82	3.52	50.20	576.43
No_Health_Insurance_Percent	3065	11.41	5.23	2.39	10.20	38.71
Severe_Housing_Cost_Burden_Percent	3065	10.63	3.50	0.69	10.15	31.25
Access_to_Exercise_Opportunities_Percent	3065	60.82	23.30	0.03	62.70	100.00

Isolation_Index_Hispanic	3065	0.13	0.15	0.00	0.07	0.99
Isolation_Index_Non_Hispanic_Asian	3065	0.03	0.05	0.00	0.02	0.52
Isolation_Index_Non-Hispanic_Black	3065	0.14	0.18	0.00	0.06	0.88
Isolation_Index_Non-Hispanic_White	3065	0.79	0.17	0.03	0.85	1.00
rural_urban_encoded	3065	0.43	0.49	0	0	1

2. OLS Results

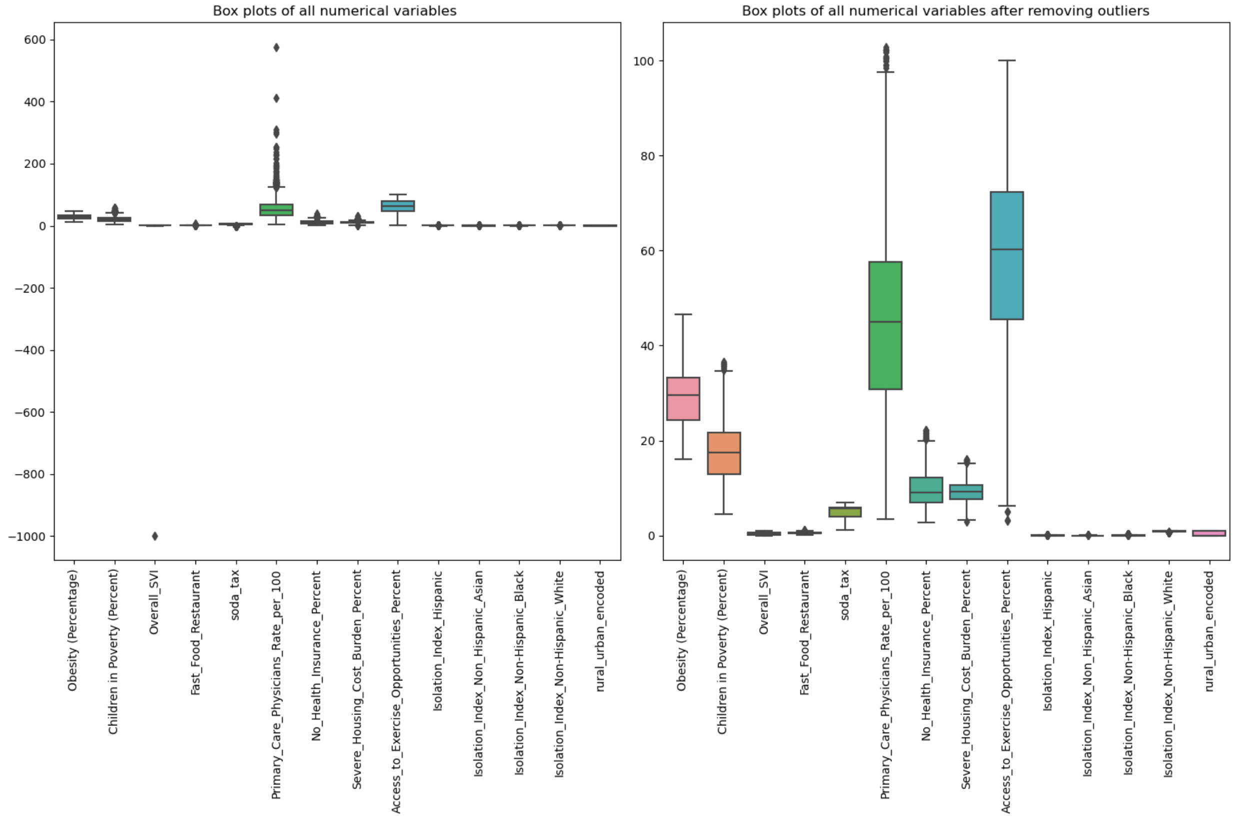
The significant negative coefficients for 'Fast Food Restaurant', 'No Health InsurancePercent', 'Isolation Index Hispanic', and 'Isolation Index Non Hispanic Asian' indicate inverse relationships with obesity rates. Conversely, 'Isolation Index Non-Hispanic Black' and 'Access to Exercise Opportunities Percent' have positive relationships with obesity rates, pointing to potential areas for targeted public health interventions. The variable 'rural_urban_encoded' shows a strong negative association, suggesting a marked difference between rural and urban areas. Variables such as 'Children in Poverty (Percent)'. 'Overall_SVI', and others are not statistically significant, which might be not strong factors contributing to the obesity rate.

	coef	**std err**	**t**	**P>|t|**
const	32.3365	1.434	22.555	0.000
Children in Poverty (Percent)	0.0282	0.019	1.518	0.129
Overall_SVI	-0.0057	0.006	-0.940	0.347
Fast_Food_Restaurant	-1.4501	0.436	-3.329	0.001
soda_tax	0.0494	0.055	0.891	0.373
Primary_Care_Physicians_Rate_per_100	-0.0064	0.004	-1.776	0.076
No_Health_Insurance_Percent	-0.2168	0.028	-7.763	0.000
Severe_Housing_Cost_Burden_Percent	-0.0565	0.039	-1.442	0.150
Access_to_Exercise_Opportunities_Percent	0.0302	0.006	5.062	0.000
Isolation_Index_Hispanic	-4.2964	1.152	-3.731	0.000
Isolation_Index_Non_Hispanic_Asian	-5.9029	2.799	-2.109	0.035
Isolation_Index_Non-Hispanic_Black	2.7965	0.900	3.109	0.002
Isolation_Index_Non-Hispanic_White	-0.7318	1.173	-0.624	0.533
rural_urban_encoded	-3.5601	0.255	-13.984	0.000

3. Numerical Variable Analysis and Removing Outliers

The comparative box plots depict distributions of numerical variables before and after outlier removal. Prior to cleansing, the data exhibit pronounced extremes, particularly in 'Overall_SVI', where values plummet below -800, indicating potential data entry issues or exceptional cases needing scrutiny. Post-cleansing, the plots reflect a more standardized range with diminished variability, enhancing interpretability. Notably, 'Obesity (Percentage)' and 'Children in Poverty (Percent)' exhibit wider interquartile ranges, suggesting significant variations across the dataset's entities. These cleaned distributions will likely yield more reliable insights in subsequent analyses, as extreme outlier influence is mitigated.

Nevertheless, the presence of remaining outliers in 'Children in Poverty (Percent)' after cleansing signals persistent heterogeneity that may still warrant investigation. These refined visualizations underscore the diversity of socioeconomic factors encapsulated within the dataset, where the emphasis can now shift towards understanding the underlying patterns and causes within the data's normal operating range.



Methodology

1. Model

In the pursuit of predicting county-level obesity rates, due to the complex and hidden mechanism by which social determinants impact social determinants, I will employ a multi-model approach to discern the most effective predictive model, instead of assigning a specific approach.

Firstly, I will utilize a **Linear Regression** model, the bedrock of statistical modeling, which assumes a linear relationship between the dependent and independent variables. Its simplicity and interpretability make it an essential starting point. However, recognizing the potential for overfitting and the need for regularization, I will also implement **Lasso Regression** and **Ridge Regression**. Lasso, with its L1 regularization, has the added advantage of feature selection by reducing coefficients for less important features to zero. Ridge Regression, on the other hand, employs L2 regularization, which does not eliminate coefficients but reduces their magnitude, thus mitigating overfitting while maintaining the model complexity.

To capture non-linear relationships and interactions between features, I will **utilize Random Forest Regression**, an ensemble method that combines the output of multiple decision trees to improve prediction accuracy and control overfitting. **Support Vector Regression (SVR),** which excels in high-dimensional spaces, is also one of my choice. It operates on the principle of finding a hyperplane that best fits the data within a certain margin, leveraging kernel tricks to manage non-linear relationships. Lastly, **XGBoost,** a gradient boosting algorithm, is known for its performance and speed. It builds sequential trees that learn from the errors of the predecessors, constantly improving upon them, which can be particularly effective when the relationship between social determinants and obesity rates is complex and non-linear.

The choice of these models is governed by the desire to balance simplicity with predictive power, the need for regularization to prevent overfitting, and the capability to capture complex, non-linear relationships inherent in social health data. By **comparing these models**, I aim to uncover the most robust and insightful predictors of obesity rates, providing a valuable tool for public health strategy and policy-making.

2. Endogeneity

Endogeneity is a concern in my model due to the potential two-way causation between obesity rates and the selected predictors. For instance, while the density of fast food restaurants may contribute to higher obesity rates, it is also plausible that areas with higher obesity rates attract more fast food establishments. Similarly, the relationship between healthcare accessibility and obesity is bidirectional; areas with better healthcare might have lower obesity due to more effective weight management services, but conversely, areas with higher obesity might have improved healthcare services in response to the population's needs.

3. Bias, Variance, Interpretability

When developing predictive models, it's essential to balance bias, variance, flexibility, and interpretability to achieve a model that not only generalizes well to unseen data but also aligns with the project's objectives. Bias refers to the error introduced by approximating real-world problems, which may be complex, by a simpler model. In the context of my project, if I employ a simple linear model to predict obesity rates, my project might suffer from high bias if the true relationship is non-linear or involves interactions that the linear model cannot capture. Variance is the error due to the model's sensitivity to fluctuations in the training set. Models with high flexibility, like Random Forest and XGBoost, can capture subtle patterns in the data, leading to low bias but potentially high variance if not carefully tuned. This could result in overfitting, where the model performs well on the training data but poorly on new test data. Interpretability is a trade-off for flexibility; more complex models tend to be "black boxes". While SVR and XGBoost might offer powerful predictive capabilities, they lack the straightforward interpretability of a Linear Regression model. This makes it challenging to extract policy-relevant insights directly from their predictions.

In my project, the selection of multiple models serves as a strategy to navigate these trade-offs. Linear, Lasso, and Ridge models might offer more interpretability at the cost of potentially higher bias, making them suitable for inference purposes. In contrast, Random Forest, SVR, and XGBoost can model the complex relationships that likely exist in my data, which might reduce bias but also sacrifice some interpretability and potentially increase variance. Through careful cross-validation and hyperparameter tuning, I could mitigate overfitting, balancing bias and variance to improve model performance and generalize the model.

4. Model Selection Criterion

In selecting the most suitable model for predicting county-level obesity rates, my approach will be rooted in a rigorous evaluation of performance metrics, specifically focusing on R-squared, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) across both training and testing datasets. A higher R-squared value will indicate a greater proportion of variance in obesity rates explained by the model, thus reflecting its explanatory power. I will also focus on lower MSE and MAE values, as these metrics will quantify the average magnitude of the model's prediction errors, with MSE giving more weight to larger errors due to its squared term. RMSE quantifies the average magnitude of prediction errors in the same units as the predicted variable, sensitively highlighting even occasional large deviations between observed and predicted values. By closely analyzing these metrics, I intend to find a balance between the model's fit and its predictive accuracy.

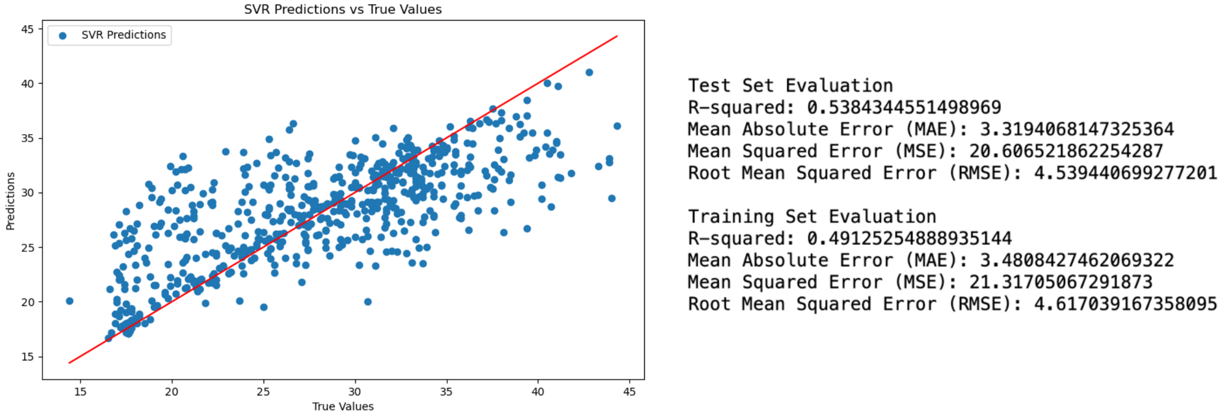
Results Analysis

1. Initiative Results (Removing Feature "States")

Model	Data Set	R-Squared	MSE	MAE	RMSE
Linear	Training	0.1672	34.8958	4.8568	5.9073
	Test	-870.7886	38920.8649	12.9694	197.2837
Lasso (α=0.5)	Training	0.1017	37.6394	5.1134	6.1351
	Test	0.1179	39.3827	5.2067	6.2756
Ridge (α=0.5)	Training	0.1672	34.8969	4.8577	5.9074
	Test	-856.7761	38295.2802	12.9051	195.6918
Random Forest	Training	0.9185	3.4150	1.4548	1.8480
	Test	0.4507	24.5226	3.9204	4.9520
XGBoost	Training	0.9843	0.6578	0.5856	0.8111
	Test	0.4472	24.6790	3.9258	4.9678
SVR	Training	0.4913	21.3171	3.4808	4.6170
	Test	0.5384	20.6065	3.3194	4.5394

Linear Models (Linear, Lasso, Ridge): These models exhibit a dramatic disparity between training and test performance. Particularly, the Linear and Ridge models show extremely negative R-squared values on the test set, suggesting that these models are highly unsuitable. This indicates that the relationship between the predictors and the target variable is likely non-linear or involves complex interactions that these models fail to capture.
Ensemble Models (Decision Tree, XGBoost): Both the Decision Tree Regressor (DTR) and XGBoost show much better training performance, with particularly high R-squared values. However, there's a noticeable drop in performance on the test set, indicating an overfitting to the training data. This drop suggests some overfitting but also confirms that ensemble methods, handle non-linearities and feature interactions more effectively.

The SVR model, as illustrated in the "General Model" rows of the comparative performance table, shows an R-squared of 0.4913 on the training set and an improved R-squared of 0.5384 on the test set. This is supported by the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) which are lower on the test set compared to the training set.



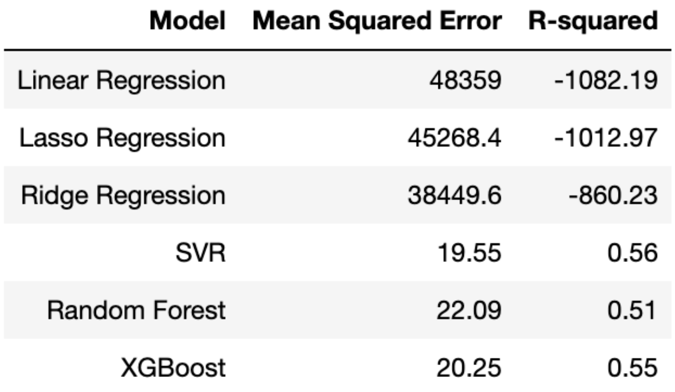
```
Test Set Evaluation
R-squared: 0.5384344551498969
Mean Absolute Error (MAE): 3.3194068147325364
Mean Squared Error (MSE): 20.606521862254287
Root Mean Squared Error (RMSE): 4.539440699277201

Training Set Evaluation
R-squared: 0.49125254888935144
Mean Absolute Error (MAE): 3.4808427462069322
Mean Squared Error (MSE): 21.31705067291873
Root Mean Squared Error (RMSE): 4.617039167358095
```

The relatively stable performance on both the training set and the test set makes the SVR the preferable model in the initiative phase.

2. Second Phrase (Including Feature "States")

I incorporated the "State" feature into the dataset and re-executed all the model analyses. The results indicated that the performance of the Linear, Lasso, and Ridge regression models



Model	Mean Squared Error	R-squared
Linear Regression	48359	-1082.19
Lasso Regression	45268.4	-1012.97
Ridge Regression	38449.6	-860.23
SVR	19.55	0.56
Random Forest	22.09	0.51
XGBoost	20.25	0.55

continued to be suboptimal, as evidenced by their high Mean Squared Errors and negative R-squared values. Conversely, the inclusion of the “State” feature markedly **improved** the R-squared values for the SVR, Random Forest, and XGBoost models, indicating a significantly enhanced predictive accuracy on test data.

This improvement can be attributed to the "State" feature potentially capturing **regional variations** that impact the dependent variable, which was not previously accounted for in the models. This kind of geographical information can often be a proxy for underlying factors that influence the behavior of the dataset, such as economic, demographic, or policy differences across states. The traditional linear models may still be struggling due to their inability to capture complex relationships or interactions between features, which non-linear models like SVR, Random Forest, and XGBoost can model more effectively.

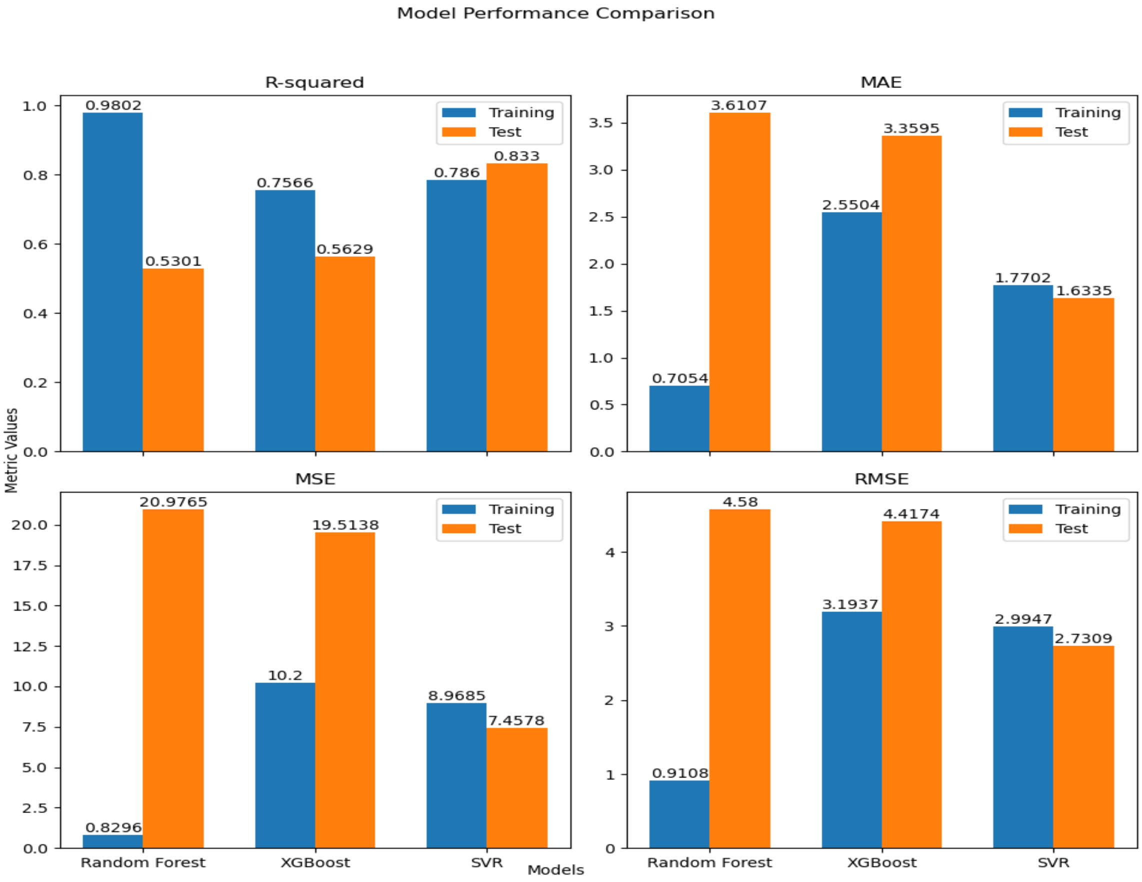
3. Precise Hyperparameter Tuning on SVR and Ensembles

In pursuit of conducting a more in-depth and **precise hyperparameter tuning**, considering the computation and time cost, I elected to implement a **random search strategy**. This approach allowed for a broader and more computationally efficient exploration of the hyperparameter space. Additionally, I expanded the hyperparameter set and incorporated additional layers into the models, further enhancing their complexity and capacity to capture nuanced patterns in the data. Further, I implemented **cross-validation** to ensure that the model's performance is consistent across different subsets of the data.

To optimize the performance of my predictive models, I've embarked on a detailed hyperparameter tuning process for XGBoost, Random Forest, and Support Vector Regression (SVR). For XGBoost, I am experimenting with parameters like the number of gradient-boosted trees (`n_estimators`), tree depth (`max_depth`), learning rate, subsample ratios, column sampling by tree (`colsample_bytree`), and minimum loss reduction (`gamma`). The Random Forest model's tuning involves adjusting the number of trees (`n_estimators`), the maximum number of features considered at each split (`max_features`), tree depth (`max_depth`), minimum samples per split (`min_samples_split`), leaf node sample requirements (`min_samples_leaf'), and the bootstrap method for sampling data points. For SVR, I am tuning across the regularization parameter (`C`), the epsilon in the loss function (`epsilon`), the kernel type, and the kernel coefficient (`gamma`).

As the Graph below indicates, for Random Forest, there's a significant drop from the training set (around 0.98) to the test set (around 0.53). This large discrepancy suggests that the Random Forest model is overfitting the training data. Overfitting occurs when a model learns the training data too well, including its noise and outliers, resulting in high performance on the training set but poor performance when predicting new data. XGBoost experiences a decrease in R-squared as well, but it’s less pronounced, going from approximately 0.76 on the training set to around 0.56 on the test set. This model appears to generalize better than the Random Forest. However, there is still overfitting in training data to some degree.

The SVR model shows an R-squared of about 0.79 on the training set and about 0.83 on the test set. The results suggest that **SVR is the best** generalizing model among the three, with the highest and most stable R-squared and lowest RMSE, MSE, and MAE on average.



4. Conclusion

In summary, my study deployed six distinct regression machine learning models to forecast county-level obesity rates. After initially removing the "States" feature, I discovered that simple linear regression performed poorly with my dataset, showing significant inadequacies. Ensemble methods such as Random Forest and XGBoost indicated a tendency to overfit the training data. However, Support Vector Regression (SVR) emerged as the most competent, exhibiting consistent and stable performance with an R-squared exceeding 0.5 across both training and test datasets, marking it as the superior model choice.

Upon reintroducing the "States" feature as dummy variables, the ensemble models' test performance notably improved, while linear regression models continued to underperform. Further refinement through meticulous hyperparameter tuning and cross-validation revealed that the overfitting on training data in the ensemble models persisted, whereas SVR not only enhanced its predictive accuracy on both the training and test sets but also remained free from overfitting. Given its robust and reliable performance, SVR stands out as the optimal model for my project, adeptly navigating the intricacies of the data without succumbing to overfitting, making it the model of choice for this research endeavor.

Limitation and Reflection

limitation

1. Data Quantity Limitation: The finite number of U.S. counties constrains the dataset's size, limiting the potential for expanding the sample size horizontally. As the number of counties remains constant, the opportunity to increase the dataset through additional observations is inherently restricted. This could impact the robustness of the machine learning models due to a limited scope of variability and prevent the full capture of the heterogeneity across the country.

2. Data Granularity Limitation: County-level data aggregates individual experiences, masking the nuances of personal circumstances. The risk here is that significant predictors of obesity at an individual level may not be detectable at the aggregate level, and policies derived from such data may not effectively address the needs of specific demographics within each county.

3. Temporal Limitation: The study's cross-sectional nature may not adequately capture the temporal dynamics of obesity rates. Changes in policies, economic conditions, and social behaviors over time are integral to understanding obesity trends, which a single-year snapshot may not fully encapsulate.

4. Model Interpretability: The 'black box' nature of the SVR model limits the interpretability of the results. Unlike more transparent models, SVR does not easily allow for the extraction of the relative importance of each feature, complicating the process of identifying the most influential factors driving obesity rates.

5. Socioeconomic and Cultural Factors: The data may not fully capture the socioeconomic and cultural nuances that significantly influence obesity rates. Variables like dietary habits, cultural attitudes towards exercise and body image, and other localized social determinants might be underrepresented in the dataset.

Future Research

Utilizing time-series data for dynamic modeling provides a potent tool for understanding the temporal dynamics of obesity trends at the county level, enabling researchers to observe how shifts in policy, economic conditions, and social factors impact obesity rates over time. Methods such as interrupted time-series analysis can be particularly useful, allowing for the assessment of specific public health interventions or policy changes by isolating their effects from other simultaneous influences. Additionally, forecasting techniques like ARIMA or Seasonal Decomposition of Time Series (STL) models can provide valuable predictions based on historical trends, aiding policymakers in anticipating future obesity rates and formulating effective interventions. In parallel, integrating technological advancements can significantly enhance data analysis: leveraging mobile health data from wearable devices allows for the real-time monitoring of physical activity and dietary patterns, while the use of artificial intelligence can facilitate the comprehensive modeling of obesity determinants by analyzing diverse data types such as text, images, and numeric data. Furthermore, developing simulation models can predict the outcomes of potential policy changes before their implementation, optimizing public health strategies. This is complemented by conducting rigorous effectiveness studies on past policies, which help refine current models and predictions, ensuring that interventions are both effective and evidence-based. Together, these approaches form a multifaceted strategy to tackle obesity by harnessing the latest in data science and technology, providing a robust framework for public health planning and policy-making.

What I learn

Throughout this project, my learning curve has been steep and enriching, marked by the gradual mastery of complex machine learning models and a deeper understanding of data analytics. I began with basic regression models, which quickly proved inadequate for capturing the nuanced influences on obesity rates, prompting a shift to more sophisticated ensemble methods like Random Forest and XGBoost. The challenges of overfitting and underfitting provided practical lessons in model tuning and the importance of cross-validation. As I delved deeper, the use of Support Vector Regression (SVR) illuminated the balance between model complexity and performance stability, teaching me the critical value of hyperparameter optimization.

Embarking on this individual project presented a unique learning trajectory, where the autonomy of decision-making came with its own set of challenges and revelations. Each step, from the initial research proposal to the intricate fine-tuning of advanced models, required personal initiative and the willingness to dive deep into problem-solving. Facing difficulties head-on, such as model overfitting or grappling with the 'black box' nature of certain algorithms, I was propelled to rely on my own resourcefulness to seek solutions and make critical trade-offs. This self-reliance cultivated a comprehensive understanding of the data and the modeling process, as each choice, be it methodological or strategic, was mine to make and justify. This solo journey not only augmented my technical acumen but also strengthened my confidence in independently managing complex projects and making informed, autonomous decisions.

Several policy suggestions can be made:

- **Health Education and Promotion**: Develop community-specific health education programs that address local social determinants identified as risk factors for obesity.

These programs could focus on nutritional education, the importance of physical activity, and how to access local health resources.

- **Urban Planning**: Advocate for policies that encourage the development of walkable neighborhoods, access to parks, and recreational facilities. Such infrastructure can promote physical activity and reduce obesity rates.
- **Food Accessibility**: Implement policies that increase access to affordable, healthy food options, especially in areas identified as 'food deserts'. Support for local markets and incentives for retailers to offer healthier food choices could be part of this initiative.
- **Socioeconomic Support Programs**: Strengthen economic support programs that indirectly affect obesity rates, such as housing assistance, education funding, and job training. Financial stability can lead to better health outcomes.
- **Healthcare Access**: Work towards expanding access to healthcare services, including preventive care that can address obesity before it leads to more serious health issues. This may involve supporting community health centers and subsidized health programs.
- **Data-Driven Interventions**: Utilize machine learning models to continue assessing the effectiveness of interventions. Policymaking should be informed by data analytics to ensure resources are directed where they are most needed.
- **Intersectoral Collaboration**: Encourage collaboration between different sectors such as health, education, and urban planning to address obesity in a more holistic manner.
- **Monitoring and Evaluation**: Establish a framework for ongoing monitoring and evaluation of obesity-related policies, ensuring they remain responsive to the changing dynamics of community health.

Reference:

1. Centers for Disease Control and Prevention. (n.d.). Diabetes Atlas. Retrieved from https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html

2. Centers for Disease Control and Prevention. (n.d.). Why Obesity Matters. Retrieved from https://www.cdc.gov/obesity/about-obesity/why-it-matters.html

3. Tax Foundation. (2018). State Sales Taxes on Soda, Candy, and Other Groceries, 2018. Retrieved from https://taxfoundation.org/data/all/state/sales-taxes-on-soda-candy-and-other-groceries-2018/

4. Ivanescu, A. E., et al. (2016). The importance of prediction model validation and assessment in obesity and nutrition research. International Journal of Obesity, 40(6), 887–894. https://doi.org/10.1038/ijo.2015.214

5. Singh, B., & Tawfik, H. (2019). A machine learning approach for predicting weight gain risks in young adults. In 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT) (pp. 231-234).

6. Uçar, M. K., Uçar, Z., Köksal, F., Daldal, N. (2021). Estimation of body fat percentage using hybrid machine learning algorithms. Measurement, 167, Article 108173.

7. Fergus, P., Hussain, A., Hearty, J., Fairclough, S., Boddy, L., Mackintosh, K. A., Stratton, G., Ridgers, N. D., & Radi, N. (2015). A machine learning approach to measure and monitor physical activity in children to help fight overweight and obesity. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9226(pp. 676-688). Springer Verlag.

8. Pleuss, J. D., Talty, K., Morse, S., Kuiper, P., Scioletti, M., Heymsfield, S. B., & Thomas, D. M. (2019). A machine learning approach relating 3D body scans to body composition in humans. European Journal of Clinical Nutrition, 73(2), 200-208.

9. Cheng, H., Scott, M., Green, A., & Furnham, A. (2020). Biomedical, psychological, environmental and behavioural factors associated with adult obesity in a nationally representative sample. Journal of Public Health, 42(3), 570-578.

10. Choukem, S.-P., Tochie, J. N., Sibetcheu, A. T., Nansseu, J. R., Julian, P., & Hamilton-Shield, J. P. (2020). Overweight/obesity and associated cardiovascular risk factors in sub-Saharan African children and adolescents: a scoping review. International Journal of Pediatric Endocrinology, 2020(1), Article 6.

11. Keramat, S. A., Alam, K., Gow, J., Stuart, J., & Biddle, H. (2021). Impact of disadvantaged neighborhoods and lifestyle factors on adult obesity: Evidence from a 5-year cohort study in Australia. American Journal of Health Promotion, 35(1), 28-37.

12. Sun, S., He, J., Shen, B., Fan, X., Chen, Y., Yang, X. (2020). Obesity as a "self-regulated epidemic": Coverage of obesity in Chinese newspapers. Eating and Weight Disorders, 26(2), 569-584.

13. Farran, B., AlWotayan, R., Alkandari, H., Al-Abdulrazzaq, D., Channanath, A., & Thanaraj, T. A. (2019). Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: A retrospective cohort study of health data from Kuwait. Frontiers in Endocrinology, 10, Article 624.