**Social Determinants of Health Impact on Quality of Life in US Counties: A Predictive Analysis**

Project Proposal

Team Lead: Steve Uzupis

Recorder: Miranda Gemme

Spokesperson: Udumaga Onyeukwu

Data Science Master’s Program, Merrimack University

DSE6311 Data Science Capstone

Supervisor: Dr. Katherine Geist

July 19, 2024

**Background & Question**

The Centers for Disease Control and Prevention (CDC) has extensively researched social determinants of health (SDOH), demonstrating that factors such as income, education, and access to healthcare significantly influence health outcomes. Their findings suggest that addressing SDOH can improve health and quality of life [1](https://www.cdc.gov/socialdeterminants/index.html.). This project explores the question: “How do social determinants of health affect quality of life in different localities”. We will use available data on SDOH to predict the self-reported health status of the adult population in US counties. Self-reported health status, such as the percentage of adults reporting fair or poor health, is a widely recognized measure of overall health. It encompasses various dimensions of health, including physical, mental, and social well-being.[2](https://www.cdc.gov/socialdeterminants/index.html.) Additionally, the project will rank the relative importance of different social behaviors and environments on our selected health outcome.

This research question is motivated by the CDC's extensive studies, which highlight the significant impact of SDOH on people's lives. Understanding these relationships is crucial for designing targeted public health interventions.

This research addresses a critical gap by providing a detailed analysis of how various SDOH influence health outcomes at the county level. While there is substantial evidence of the impact of SDOH, localized, data-driven insights are needed to inform policy and resource allocation effectively.

Exploring this question is valuable because it can guide policymakers on which SDOH are most influential, leading to more effective public health strategies. It helps prioritize resources and interventions based on the most impactful determinants and reduces health disparities by addressing the root causes of poor health outcomes.

Although the relationship between SDOH and health outcomes is well-documented, this project is novel in its approach because it provides county-level insights rather than broad national or state-level analyses, it ranks the relative importance of different SDOH, offering more actionable insights and it uses data science techniques (predictive modeling) to predict health outcome thereby enhancing the precision and applicability of the findings.

Even though the question is not entirely novel, the methodology and focus on actionable insights make it a valuable and innovative contribution to social research.

**Hypothesis and Prediction:**

**Hypothesis:** Social determinants of health, such as economic stability, social connectedness, access to healthcare, and neighborhood environment, significantly predict self-reported health status in US counties.

**Prediction:** US counties with higher economic security, stronger social support infrastructure, better access to healthcare services, and safer, more accessible neighborhoods will report better overall health status compared to counties lacking these social determinants.

**Data & Analysis**

Data for this project is sourced from two datasets. The details of the contents are presented in the data dictionary (see Appendix 1).

**Predictor Variables**

The predictor variables are sourced from the Social Determinants of Health (SDOH) Database provided by the Agency for Healthcare Research and Quality (AHRQ). This dataset is a robust choice because it includes well-documented, easily linkable SDOH variables across multiple domains. These domains cover critical aspects of health outcomes, such as social context, economic context, education, physical infrastructure, and healthcare context[3](https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html). We selected 80 features distributed across five domains of SDOH[4](https://www.cdc.gov/public-health-gateway/php/about/social-determinants-of-health.html) identified by the CDC, to ensure a comprehensive analysis that captures the multifaceted nature of social determinants of health.

**Response Variable**

The response variable, representing “the percentage of adults reporting fair or poor health” per county, is sourced from the County Health Rankings dataset. This dataset is an excellent choice because it provides detailed health outcomes and factors at the county level, allowing for granular analysis. The data is well-documented and widely used in public health research, ensuring reliability and validity[5](https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation). A response vector will be created from it with each county measurement being better or worse than the national average

The dataset for this project can be reviewed at or our GitHub repo: <https://github.com/suzupis007/capstone_project.git>

**Analysis Plan:**

Our analysis plan is outlined as follows:

**1. Project Overview**

The goal is to predict the quality of life in various counties using social determinants of health (SDOH). We will use data from sources: County Health Rankings and SDOH by the CDC.

**2. Data Collection**

County Health Rankings: This dataset includes the measure of health outcomes that this project intends to study namely “the percentage of adults reporting fair or poor health” per county in the US and some predictors in each of the five categories.

SDOH by the CDC: Provides measures for the Social Determinants of Health in different geographical and administrative areas including counties.

**3. Data Preparation**

Merge Datasets: Combine the County Health Rankings and SDOH datasets using a common identifier; the county FIPS code.

Clean Data: Handle missing values, remove duplicates, and ensure consistency in data formats.

Feature Engineering: Create new features if necessary, such as aggregating related variables or creating interaction terms.

**4. Exploratory Data Analysis (EDA)**

Correlation Analysis: Identify features with strong correlations within the dataset's SDOH that can lead to unwanted biases for removal.

Feature subset selection: use available statistical analysis techniques such as ridge regression, lasso, and partial least squares to identify the best sub-selection predictors in the model.

Descriptive Statistics: Summarize the main characteristics of the data.

Visualization: Use plots (e.g., histograms, scatter plots, heatmaps) to understand the distribution and relationships between variables.

**5. Model Selection**

Define the Response Variable: Percentage of adults reporting fair or poor health.

Split Data: Randomly divide the data into a training and testing set containing 75% and 25% respectively.

Select Models: Consider various models such as linear regression, decision trees, random forests, and gradient boosting machines.

Model Training: Train the models on the training set.

**6. Model Evaluation**

Performance Metrics: For classification models, use accuracy, precision, recall, and F1-score.

**7. Model Interpretation**

Feature Importance: Identify which SDOH features are most influential in predicting our quality-of-life measure.

**8. Reporting and Communication**

Document the entire process, including data sources, methods, and results, and prepare final reports to share insights and recommendations with the team and other stakeholders.

**Conclusion:**

The research question is answered if the analysis provides clear insights into how social determinants of health (SDOH) affect the quality of life in different localities. This can be demonstrated through model accuracy, and the ability to explain the variance in the self-reported health status based on SDOH variables.

Also, the hypothesis is supported if the results show that SDOH variables significantly predict self-reported health status. This can be evaluated using:

Statistical Tests: Look for significant p-values (typically < 0.05) for the predictor variables in the regression models.

Model Performance: High R-squared values in regression models indicate that a large proportion of the variance in the response can be explained by predictors variables.

Feature Importance: In machine learning models such as Random Forest, feature importance scores can indicate which SDOH variables are most influential in predicting health outcomes.

**References:**

1.Centers for Disease Control and Prevention. (2021). Social determinants of health: Know what affects health. Retrieved from <https://www.cdc.gov/socialdeterminants/index.html>.

2.Centers for Disease Control and Prevention. (2021). Social determinants of health: Know what affects health. Retrieved from <https://www.cdc.gov/socialdeterminants/index.html>.

3.Agency for Healthcare Research and Quality. (n.d.). Social Determinants of Health Database. Retrieved from <https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html>

4. Centers for Disease Control and Prevention. (n.d.). Social Determinants of Health. Retrieved from <https://www.cdc.gov/public-health-gateway/php/about/social-determinants-of-health.html>

5. University of Wisconsin Population Health Institute. (2024). County Health Rankings & Roadmaps. Retrieved from <https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation>