Table of Contents

HEALTHCARE DATA ANALYSIS – HOSPITALIZATION RISK PREDICTION:

Assessing Healthcare Access and Predicting Hospitalization Risk Using Socioeconomic and Geographic Factors.

Project Overview:

This proposal outlines a data analysis project aimed at examining healthcare accessibility and developing a predictive model for hospitalization risk across U.S. census tracts using social determinants of health (SDOH). By analyzing the relationships between healthcare access, socioeconomic status, and geographic factors, we seek to identify communities at heightened risk and provide actionable insights for public health officials, policymakers, and healthcare providers to address gaps in healthcare accessibility.

1 Introduction:

Healthcare accessibility remains a critical challenge in the United States, with considerable disparities in access and outcomes across different communities. While direct hospitalization data is often difficult to obtain, social determinants of health—including proximity to medical facilities, health insurance coverage, poverty levels, and disability rates—serve as valuable proxies for predicting hospitalization risk.

This project will analyze the Agency for Healthcare Research and Quality (AHRQ) Social Determinants of Health Database to understand these relationships and develop a predictive model for hospitalization risk. We expect our findings to provide actionable insights that can help address health care access disparities.

2 Problem Statement:

Many communities face barriers to healthcare access, which can lead to delayed treatment, increased emergency visits, and higher hospitalization rates. However, hospitalization risk factors are often hidden within socioeconomic and geographic disparities.

This project will:

- Clean and preprocess raw SDOH data to identify missing values and inconsistencies.
- Engineer new features that approximate hospitalization risk based on existing social determinants.

- Perform hypothesis testing to analyze disparities in healthcare access.
- Develop a predictive model to estimate hospitalization risk based on socioeconomic and geographic factors.

3 Data Source and Description:

- **Data Source**: Agency for Healthcare Research and Quality (AHRQ) - SDOH Database at https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html
- **Data Download Source**: Census Tract Data 2009-2020 (2020) https://www.ahrq.gov/downloads/sdoh/sdoh_2020_tract_1_0.xlsx
- **Data Structure**: Census tract level variables, including;
  - **Healthcare Access**: Distance to emergency departments, clinics, and trauma centers.
  - **Socioeconomic Factors**: Poverty levels, median income, food stamp recipients.
  - **Demographics**: Age distribution, disability prevalence, veteran status.
  - **Health Insurance Coverage**: Percentage of uninsured, Medicaid, Medicare recipients.
  - **Environmental Factors**: PM2.5 pollution levels.

4 Research Question and Hypothesis:

This project will analyze a 2020 dataset from the U.S. and its territories to investigate the following research question using statistical analysis and predictive modeling:

Question: Can hospitalization be predicted by distance to care, poverty, health insurance coverage, disability status, and household size, after accounting for other socioeconomic and geographic indicators?

Hypothesis: Socioeconomic and geographical factors contribute to predicting hospitalization risk.

Prediction: A predictive model incorporating distance to healthcare facilities, poverty levels, health insurance coverage, disability status, and household size will effectively estimate hospitalization risk. We expect that greater distance to care, higher poverty, lack of insurance, having a disability, and a smaller household will be associated with increased hospitalization risk.

5 Variable Justification and Intersections:

The chosen predictor variables—distance to care, poverty level, and health insurance coverage—are grounded in public health research and the social determinants of health framework. These factors are not only significant on their own but also interact in meaningful ways. For instance, individuals in high-poverty neighborhoods are more likely to live farther from hospitals and have lower rates of health insurance coverage. This geographic and economic overlap can compound barriers to care, increasing the risk of hospitalization. Understanding and quantifying these intersections will be a key focus of our analysis.

6 Methodology:

This project will be carried out in the following stages:

6.1 Data Cleaning & Preprocessing:

The SDOH dataset has over 85500 observations and 329 features. Based on preprocessing analysis:

- 19 columns contain more than 30% missing values and will be removed.
- 310 columns have less than 30% missing values and will be retained.
- Missing values in retained columns will be imputed using appropriate techniques (mean, median, or mode).
- Normalize numerical variables (e.g., income, distances, pollution levels).
- Convert categorical variables (e.g., insurance type) into numeric features.

6.2 Feature Engineering:

The target variable for modeling will be an engineered hospitalization risk score, derived from the following proxy factors:

- Proximity to hospitals and clinics (higher distance = higher risk).
- Uninsured rate (higher uninsured = higher risk).
- Poverty rate (higher poverty = higher risk).
- Disability and elderly population rates (higher variables = higher risk).
- Air pollution exposure (higher PM2.5 value = higher respiratory risk).
- The risk score may also be converted into categorical levels (Low, Medium, High) for classification tasks and weights (w1-w5) will be determined through correlation analysis with known health outcome.

Example formula:
Risk Score = w1 × Distance to Hospital + w2 × Uninsured Rate + w3 × Poverty Rate + w4 × Disability Rate + w5 × PM2.5 Level

6.3 Exploratory Data Analysis:

Following multi stage analytic approach can be implemented.

- Statistical analysis of healthcare access variables.
- Correlation analysis between socioeconomic factors and healthcare access metrics.
- Geographic distribution of key predictors.

6.4 Hypothesis Testing:

- T-tests/ANOVA to compare hospitalization risk proxies across different demographic groups.
- Correlation analysis between poverty, insurance coverage, and hospital access.
- Statistical tests examining relationships between all key variables.

6.5 Predictive Modeling:

To assess the model's ability to predict hospitalization risk based on social and geographic factors, the following modeling strategy will be adopted:

6.5.1 Define Modeling Objective:

- If using the continuous risk score as the target, we will apply **regression** models.
- If using risk score categories (e.g., Low/Medium/High), we will use **classification** models.

6.5.2 Model Selection:

For classification tasks:

- Logistic Regression for baseline interpretability.
- Decision Trees and Random Forest Classifier to capture non-linear relationships.
- XGBoost Classifier for advanced performance and feature importance.

For regression tasks:

- Linear Regression as a baseline model.

- Random Forest Regressor and XGBoost Regressor to model non-linearity and interactions to leverage ensemble methods of decision trees to capture complex relationships.
- Ridge/Lasso Regression to manage multicollinearity and perform regularization.

6.5.3 Model Evaluation:

- Classification models be evaluated using:
  - Accuracy.
  - Precision.
  - Recall.
  - F1-Score
  - ROC – AUC.
- Regression models be evaluated using:
  - RMSE.
  - MAE.
  - R-Square.
- Cross-validation to ensure model generalizability to reduce the risk of overfitting and providing a more reliable estimate of its performance.

7 Planned Visualization:

- Histogram and density plots of distance to care and income.
- Correlation heatmap for predictor variables.
- Risk score distribution by region.
- Bar charts comparing hospitalization risk by demographic groups.
- Feature importance plots from Random Forest/XGBoost.
- ROC curves for classification models.
- Actual vs. predicted plots for regression models.

8 Expected Key Insights:

Based on the analysis, the following key insights are expected:

- Identify regions with high estimated hospitalization risk based on social determinants.
- Provide policy recommendations to improve healthcare access in underserved areas.

- Demonstrate the feasibility of using public SDOH data to estimate hospitalization risk.

9 Target Audience:

Following audience may be intended for this project:

- Public Health Agencies in the United States and its Territories

10 Conclusion:

This preliminary proposal outlines a comprehensive approach to analyzing healthcare accessibility and predicting hospitalization risk using socioeconomic and geographic factors. By leveraging the AHRQ Social Determinants of Health Database and applying robust data science methodologies, we aim to generate valuable insights that can inform healthcare policy and resource allocation decisions. The project aligns with current efforts to address healthcare disparities and improve access to care across diverse communities.

References:

1. Artiga, S., & Orgera, K. (2020). Disparities in health and health care: Five key questions and answers. *KFF (Kaiser Family Foundation)*. This source provides broad information on health disparities. https://www.kff.org/racial-equity-and-health-policy/issue-brief/disparities-in-health-and-health-care-5-key-question-and-answers/
2. Am J Manag Care. 2021;27(3):e89-e96. https://doi.org/10.37765/ajmc.2021.88603
3. Agency for Healthcare Research and Quality (AHRQ). (n.d.). Social Determinants of Health (SDOH) Data & Analytics. https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html
4. Agency for Healthcare Research and Quality – Data Source Documentation https://www.ahrq.gov/sites/default/files/wysiwyg/sdoh/SDOH-Data-Sources-Documentation-v1-Final.pdf
5. National Academies of Sciences, Engineering, and Medicine. (2021). Implementing high-quality primary care: Rebuilding the foundation of health care. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK578537/
6. Bhuiyan, A. R., Fennie, K. P., & Ahmed, S. M. (2020). Exploring the relationship between social determinants of health and healthcare utilization among adults with multimorbidity: a cross-sectional study. *BMC Health Services Research*, *20*(1), 593. https://pmc.ncbi.nlm.nih.gov/articles/PMC7314918/
7. National Academies of Sciences, Engineering, and Medicine. (2012). Living well with chronic conditions: A self-management approach. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK218212/
8. Decision Tree, Random Forest, and XGBoost: An exploration into the heart of Machine Learning. Medium. https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948
9. Cross-validation in machine learning: How to do it right. Neptune AI Blog. Retrieved from https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right