Project Report

INDEX

INTRODUCTION

Problem Statement Description

Data analysis is crucial in understanding Social Determinants of Health (SDOH), offering an efficient alternative to relying solely on patient interviews. This saves time for healthcare professionals and empowers patients with information about how they can avail medical care. The primary goal is to employ proactive techniques, emphasizing the adage "prevention is better than cure."

Objectives

Social Care Scorecard

Leveraging data on socio-economic status, demographics, health conditions, and other factors, a scorecard is generated. This helps physician in making well-informed decisions while treating patients.

Connecting People with Resources

The scorecard goes beyond insights, offering customized actions and resources based on individual needs. It not only informs but guides the next steps, connecting people with support services they require.

The India Connection

India's healthcare environment is uniquely influenced by social determinants, which impacts health outcomes. Identifying relevant data sources is crucial for developing a predictive model for India.

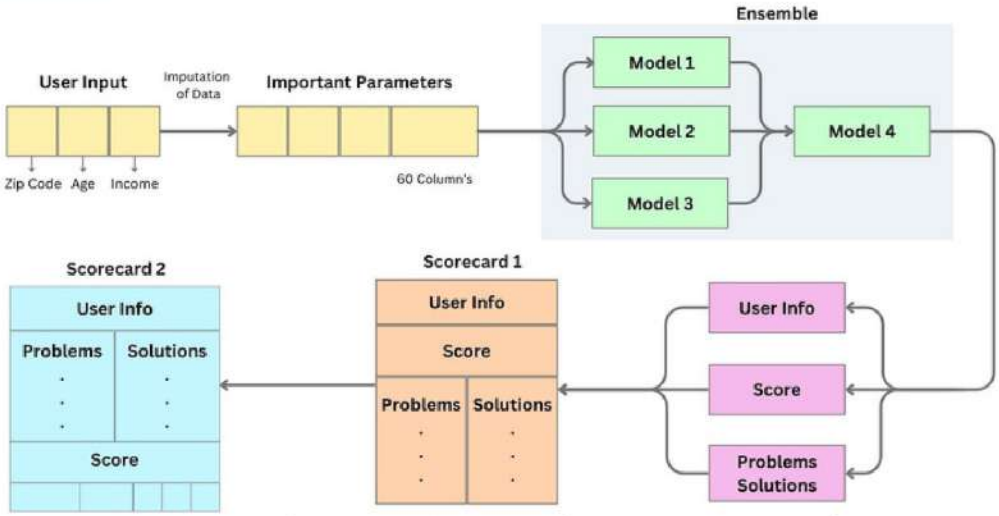


Figure 1: Overview of scorecard generation approach

Dataset Description

Objective 1 involves the analysis of SDOH and health data to understand illness prevalence and localized health outcomes, while secondary data on life expectancy adds context to population health trends. For Objective 2, datasets focus on mapping to support like *The Home Health Provider* dataset, which offers details on home health services, aiding in developing a support network for individuals requiring home-based care. Similarly, datasets with details on other services are provided. A *Census Tract Crosswalk* dataset is used to enhance spatial analysis, allowing for a more thorough understanding of health and socio-economic data at the ZIP code level.

DATA VISUALIZATION

To enhance understanding, raw data is visualized through charts and plots. This allows for clear identification of trends, patterns, and relationships within the data that may be difficult to discern from datasets alone.

SDOH Related Plots

The sunburst plot visually depicts a hierarchical structure of Social Determinants of Health (SDOH) topics categorized within broader SDOH domains. This visual arrangement allows for the quick apprehension of how complex and interrelated the SDOH are, as well as how they might be considered in developing comprehensive healthcare strategies.

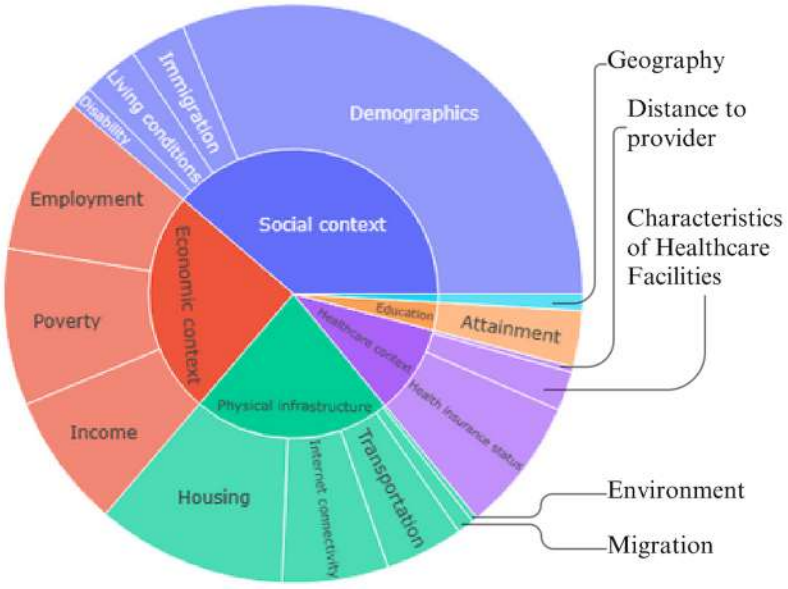


Figure 2:Hierarchical view of SDOH domains

Geospatial Insights

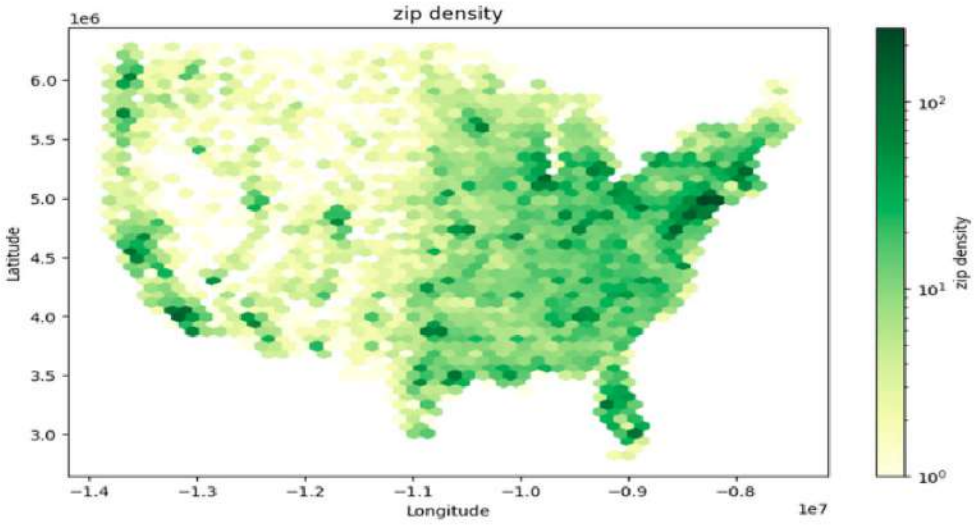
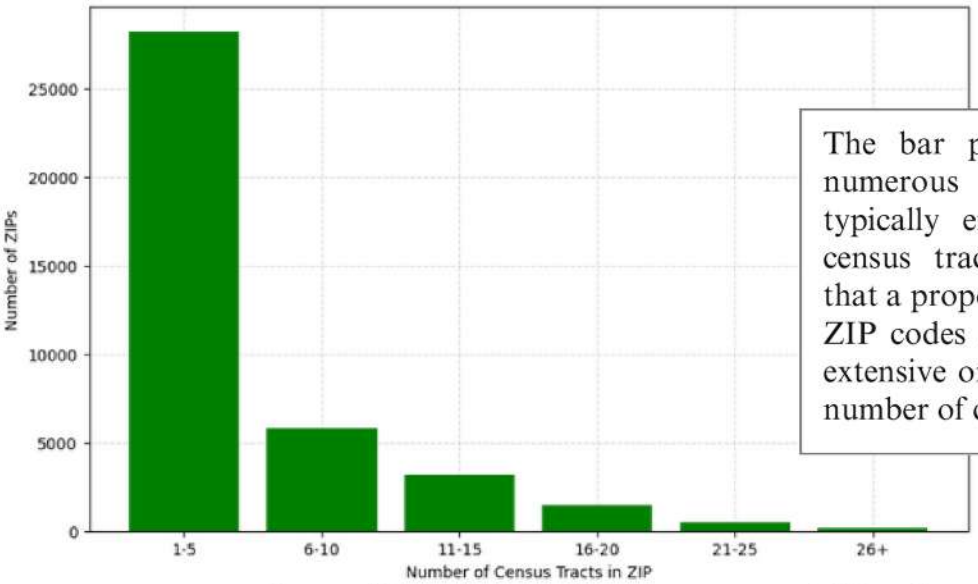


Figure 3: ZIP code density in the United States

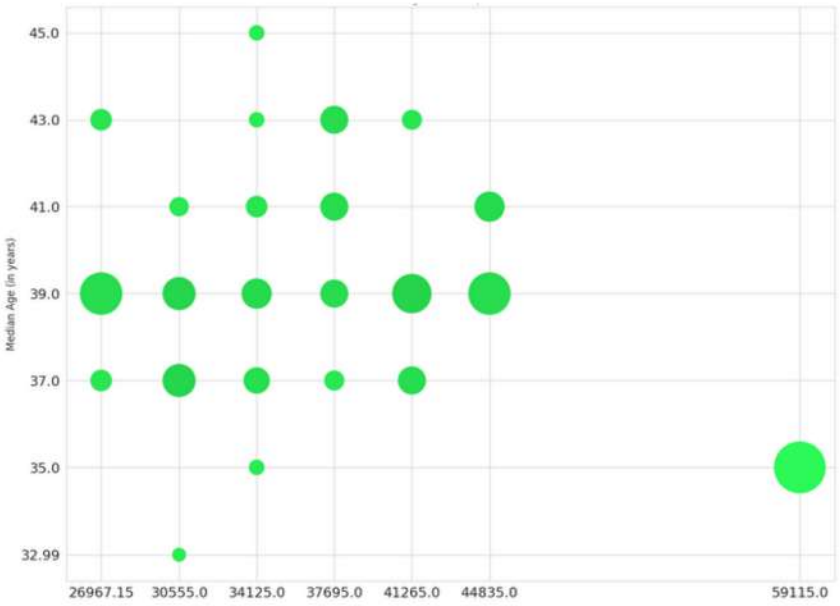
The plot suggests that ZIP code density can be critical for location-based analysis for health outcomes across the United States. The concentration of darker greens, particularly in areas corresponding to the East Coast, Midwest, and West Coast, suggests a higher ZIP code density in these regions.



The bar plot illustrates numerous ZIP codes typically encompass 1-5 census tracts, indicating that a proportion of given ZIP codes are either less extensive or have a lower number of census tracts.

Figure 4: Distribution of census tracts per ZIP code

Socioeconomic Snapshots



The bubble plot highlights a trend where median age and per capita income seem to increase concurrently. Notably, the spread of income is broader than the range of median age, indicating a more extensive variation in wealth than what might be inferred solely from age gaps.

Figure 5: Relationship between median age and per capita income

This choropleth map exposes a pronounced digital divide. Darker-shaded states exhibit a higher percentage of households lacking internet access. The scattered distribution indicates that the divide isn't solely geographic; it extends beyond potentially rural areas to include disparities across all regions, including potentially urbanized states.

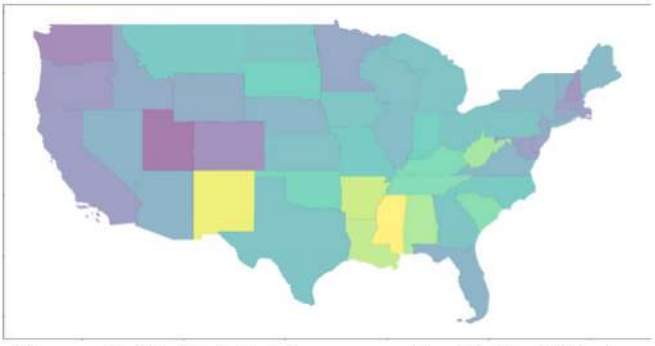
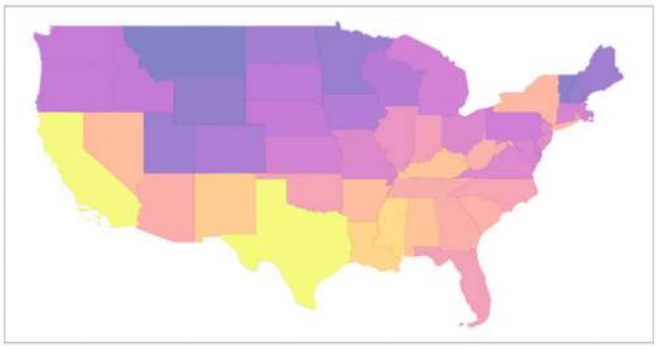


Figure 6: Digital divide across the United States

Demographic Dynamics



The choropleth map unveils an uneven educational landscape, particularly in states like California and Texas, where significant population lack a high school diploma. Immigration rates may influence these figures, given potential differences in educational attainment levels between immigrants and the native population.

Figure 7: Educational disparities across the US

Plotting Public Health

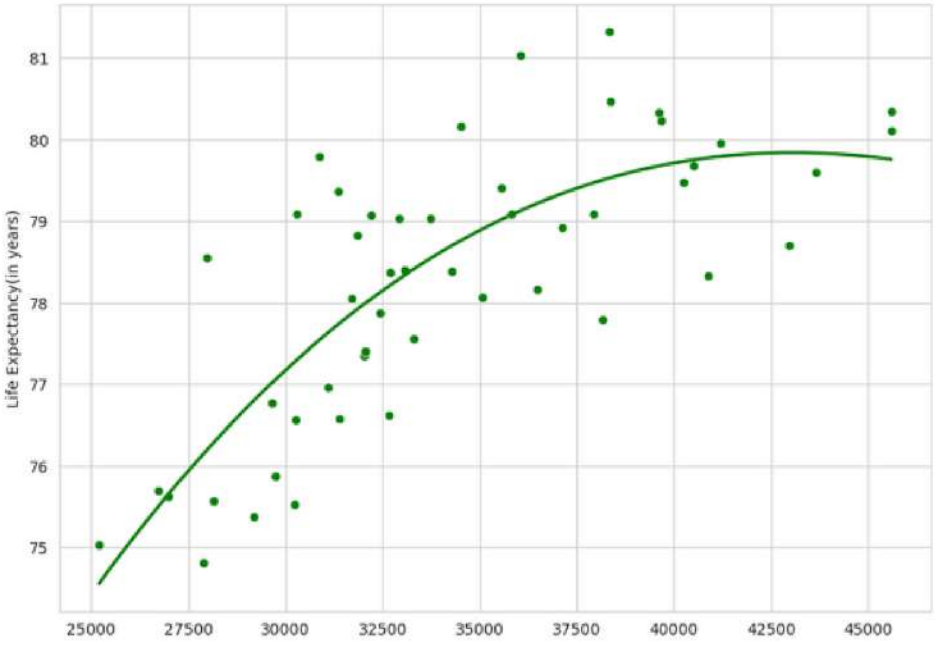
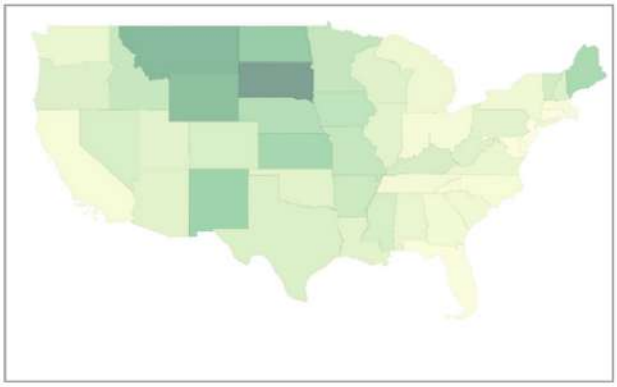


Figure 8: Relationship between income and life expectancy

The scatter plot indicates that higher income generally corresponds to longer life spans. However, this correlation appears stronger at lower income levels but not at higher income levels. This pattern may reflect income's influence on healthcare access.



This choropleth map reveals a notable disparity in urgent care facility access across the U.S. Darker-shaded areas, typically representing less densely populated states, exhibit a demonstrably greater distance to the nearest urgent care center. This implies potential challenges for residents in these regions to receive timely medical attention.

Figure 9: Disparities in urgent care access in the US

FEATURE ENGINEERING

DATA PREPROCESSING

Data preprocessing was performed to enhance the quality of datasets used for training, ensuring they are relevant to the analysis. Normalizing and transforming the data can ensure comparability and compatibility across different datasets. This process was crucial for deriving insights and supporting the integrity of the subsequent analysis.

• Dropping Columns

In the initial step, columns containing more than 40% missing values were identified and subsequently eliminated. This criterion was established to guarantee that the dataset retains columns with a satisfactory level of complete data, thus ensuring the overall data quality while having a meaningful feature column.

• Dropping Rows

Exclusion criteria based on data completeness was applied to the census tracts of the territories. It was discovered that missing data encompassed 319 out of the total 321 variables in 132 census tracts, indicating a significant absence of information. Furthermore, missing data was identified in 311 out of the 321 variables for an additional 981 census tracts. Due to the extensive amount of missing data, which undermines the analytical validity of the dataset, these records were dropped to preserve the integrity and reliability of future analyses

• Null Value Imputation:

Three methods were explored for filling in missing values: K-Nearest Neighbors (KNN) Imputation, Correlation-Based Imputation with DataWig, and Multiple Imputation by Chained Equations (MICE). Each method was designed to enhance data completeness and accuracy by addressing specific aspects of the dataset.

After a thorough evaluation and comparison of the three imputation methods described, it was observed that Method 2, which employs a correlation-based imputation strategy augmented with DataWig and further refined with KNN imputation, yielded the best results. The efficiency of this method was judged by the better formation of clusters and higher silhouette score for clustering.

• Data Normalization:

Various data normalization techniques were applied to process the dataset derived from census data. Techniques such as Min-Max Scaling, Z-Score Normalization, Robust Scaling, Log Transformation, and Quantile Transform were employed. The decision to use Quantile Transform was driven by the dataset's significant skewness, and presence of outliers. Metrics including Standard Deviation, Skewness, and Kurtosis exceeded predefined thresholds, validating the necessity of this normalization method. Quantile Transform emerged as the most suitable technique, reducing outlier influence and achieving a more uniform distribution for further analysis.

CORRELATION ANALYSIS

Correlation analysis explores the strength and direction of relationships between variables, providing valuable insights into their interconnectedness. Utilizing correlation analysis, certain parameters can be identified and removed that exhibit high correlation, streamlining datasets and enhancing the precision of subsequent analysis.

• Correlation Matrices

Correlation matrices efficiently identify linear relationships and redundant features in data, streamlining models by eliminating highly correlated variables. The model can, thus, potentially improve interpretability of the final results. Correlation matrices are also computationally efficient to calculate, making them a valuable first step in feature selection tasks.

• Variance Inflation Factor (VIF)

While correlation coefficients assess pairwise relationships, VIF goes beyond this by quantifying the influence of multicollinearity on a particular feature. This can lead to inflated standard errors for the model's coefficients, making it difficult to assess the true contribution of individual features. It then calculates the extent to which the variance of a specific feature is inflated due to correlation with other features in the model.This information can then be used to refine the feature set by removing or transforming problematic features, leading to a more robust and interpretable model.

FINAL SELECTED FEATURES AND DATASET

After the removal of correlated features and features with more than 40% null values, Random Forest was used to select important features.

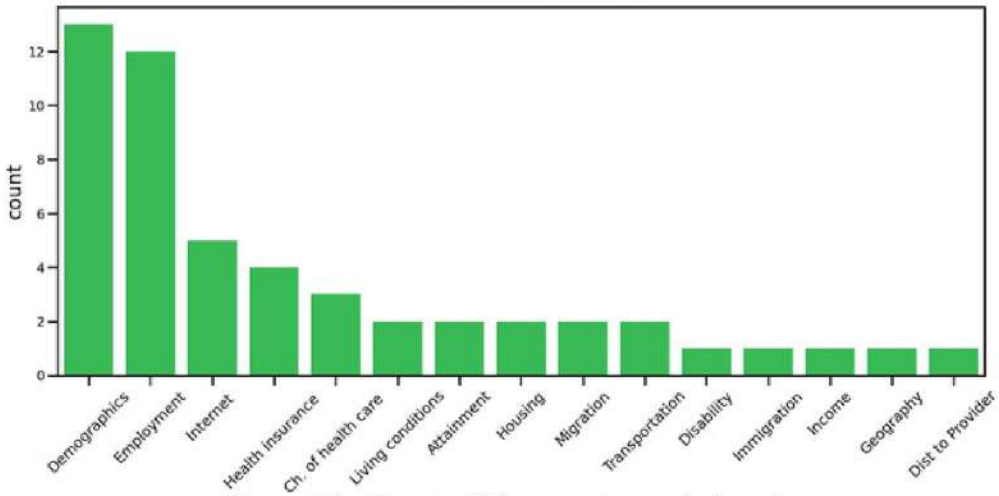


Figure 10: Count of features in each domain

In the chosen feature set, a diverse array of 14 broad areas was included, encompassing demography, employment, connectivity, and more. The aim was to incorporate variables from every relevant category. It was revealed by the analysis that factors such as demography, employment status, internet connectivity, access to healthcare insurance, and the quality of healthcare facilities exert the greatest influence on an individual's health crisis within a given area.[*Refer annexure for full list of features*]

CLUSTERING

Clustering was done to divide the dataset into supersets of problems faced in various census tracts. Clustering was done on the dataset formed after selecting 60 parameters.

GROUPING PROBLEM SUPERSETS

The dataset comprising 60 parameters and 84,414 rows was categorized into 7 distinct problem supersets, manually assigning variables to each domain. Through this process, it was identified that 52 columns exhibited significant correlations within their respective domains. Additionally, clustering was conducted on the dataset to determine the most suitable algorithm for analysis, assessing various clustering techniques and their respective scores.

Clustering Algorithm	Silhouette Score	Davies-Bouldin Score	Calinski - Harabasz Score	No. of Clusters
K-Means	0.052	4.012	4775.547	3
GMM	0.059	4.621	1517.420	3
HDBSCAN	0.118	4.765	619.992	3
DBSCAN	0.069	0.895	100.585	3

Table 1: Comparison of various clustering algorithms

Although HDBSCAN, DBSCAN, and GMM exhibited higher scores compared to the K-Means algorithm, the clusters generated by these methods displayed highly non-uniform sizes, with a majority of rows concentrated in just one cluster. As a result, these algorithms were discarded. Conversely, K-Means consistently produced clusters of uniform size, thus being the most suitable algorithm for our analysis.

Domain Name	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score	No. of Clusters
Socioeconomic Factors	0.143553	2.3456	23456.789123	3
Demographic Factors	0.166667	1.824019	18596.719598	4
Environmental Factors	0.352902	0.996131	65444.853019	4
Healthcare Factors	0.188421	1.832069	18096.281798	3
Health Behaviour Factors	0.241552	1.358535	26351.892943	3
Internet Factors	0.263791	1.222000	27185.534545	4
Housing Factors	0.148638	2.215627	11754.831766	3

Table 2: Optimal number of sub-clusters for each problem superset

Since K-Means was the best algorithm that fits this dataset, clustering was done on the 7 distinct domains dataset using the K-Means clustering algorithm, and the scores obtained are described in the table above. As the scores obtained by K-Means on these domains were greater than the threshold, these clusters were taken for further processes.

PROBLEM JUSTIFICATION & CLUSTER LINKING

Problems were identified based on cluster characteristics, and each cluster was mapped with its most relevant problem and the best possible solution by manual assignment.

- **Problem Identification:** Analysis revealed socioeconomic challenges, transportation, healthcare, education, employment issues, and housing disparities as problem sets.
- **Solution Proposal:** Tailored solutions were crafted for each cluster to address these challenges and improve community well-being.
- **Probability Calculation:** By applying the established model, probability of each test census tract was calculated aligning with the identified clusters.
- **Cluster Assignment:** Based on the calculated probabilities, the test census tracts were assigned to the top three clusters deemed most fitting by the analysis.
- **Assigning solutions:** The corresponding solutions were assigned to each test census tract to address the challenges.
- **Validation:** Synthea dataset and Gen AI techniques validated cluster problems, affirming accuracy in identifying SDOH-related clusters and associated problems.

STRUCTURING MODEL QUERY

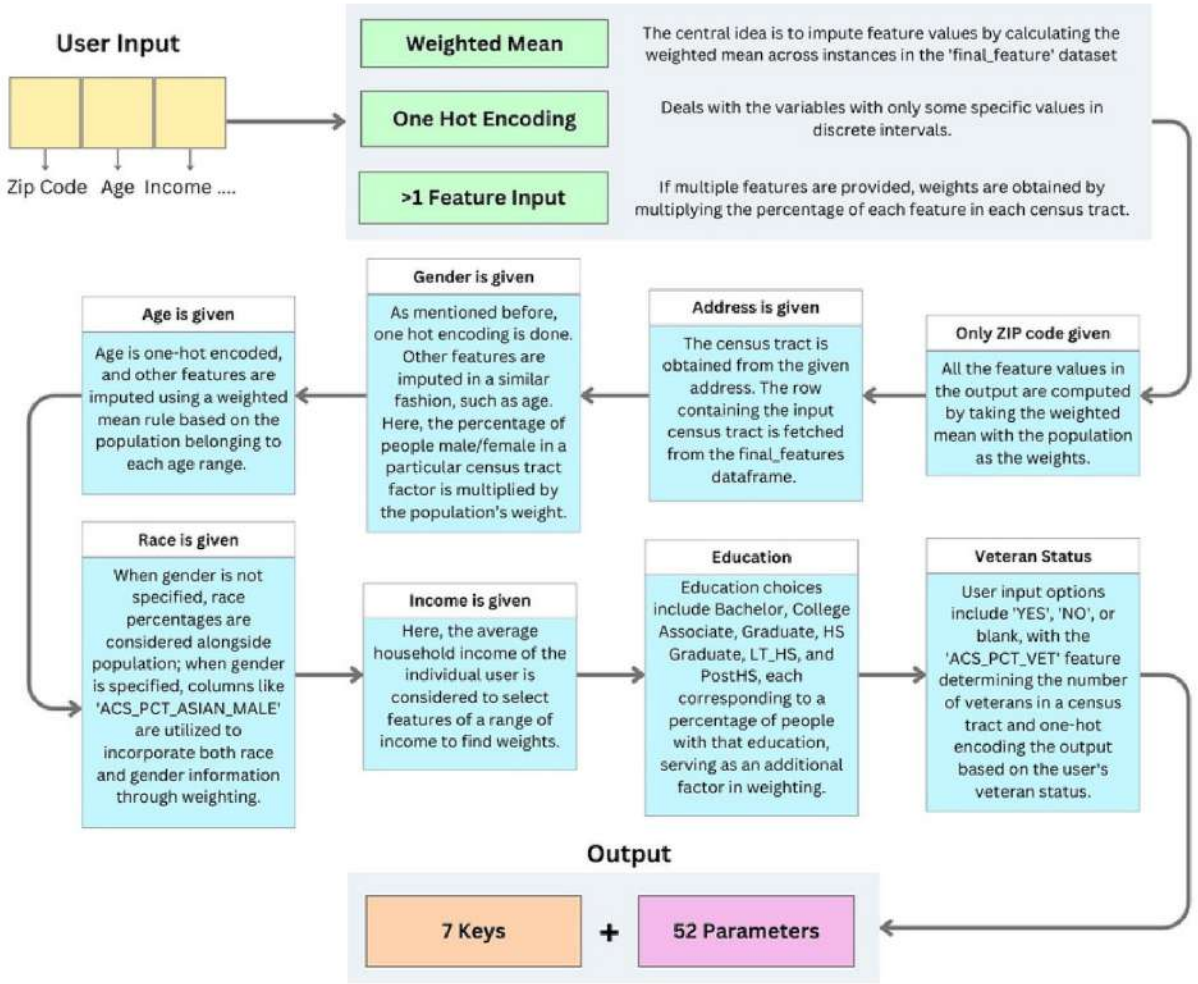


Figure 11: Flowchart depicting how a model query is structured

EXPERIMENTATION

The following section delves into various experimentation and modeling methods employed to identify challenges and solutions for residents while accessing medical facilities. The discussion provides insights into the strategies used to address and understand the healthcare-related concerns faced by the local community.

1. FAR-MEAN APPROACH

This approach assumes that the average of important parameters represents its optimal value for a ZIP code or census tract. This is based on the idea that if the value of a parameter is far from mean, it indicates better/worse conditions than the optimal value.

Domains / Problem Supersets

1. Socioeconomic Factors
2. Internet and Communication Access
3. Demographic Information
4. Healthcare Access and Utilization
5. Health Behavior and Outcomes
6. Environmental Factors
7. Housing and Transportation

After compiling the final dataset, comprising 60 identified important parameters, the data was categorized into seven distinct domains or problem supersets. For a specific ZIP code, the value of these seven domain parameters, as well as the mean value of all parameters for the ZIP code overall, were computed. Then each parameter's value, along with its average, was given with prompting to assign the fine-tuned LLaMA-2-7B chat to generate tailored problems and solutions.

2. PATIENT PERSONA APPROACH

Patient information is captured to derive problems and solutions. A dataset of about 10k rows was created with 2 string columns: one for patient details and zip codes, and the other for problems and solutions from the Synthea dataset. This dataset was used to fine-tune the Llama2-7B model for patient simulation. The clustered dataset was used by dividing the zip code dataset into 7 domains and then clustering each domain into approximately 3-4 clusters. A problems and solutions dictionary was assigned to each cluster. For disjointed zip codes, problems and solutions were extracted from the dictionary and fed into the fine-tuned Llama2-7B model for output.

3. PROBABILISTIC SCORE CALCULATION

The model's methodology incorporates an unsupervised K-Means clustering algorithm. Following feature selection, the data was split into training and testing with a split ratio of 0.15. Subsequently, SDOH variables were categorized into distinct groups, encompassing Housing and Transportation, Socioeconomic factors, Internet Access, etc. Each category was then subjected to separate K-Means clustering algorithm. Employing exploratory data analysis techniques, unique issues associated with each cluster were identified, resulting in a total of 24 sub-clusters, each indicative of a distinct problem. Following this, the probability score of the test data belonging to each sub-cluster was computed, and the top 3 probabilities, representing the top 3 problems, were selected.

4. FINE-TUNING GEMMA-2B

In this approach, the dataset comprising 84,414 rows is initially partitioned into two datasets, one containing 71,751 rows and the other 12,663 rows. Subsequently, the clustering technique is applied to the larger dataset, creating 24 clusters, each representing approximately three problems.

The top three problems for the smaller dataset were obtained for each data point based on the previously derived clusters as per confidence scores(CF). Solutions were then assigned to each data point accordingly. Consequently, a new dataset comprising approximately 12,000 data points and their respective sets of three problems and solutions is generated. This dataset is then utilized to fine-tune the Gemma-2B model.

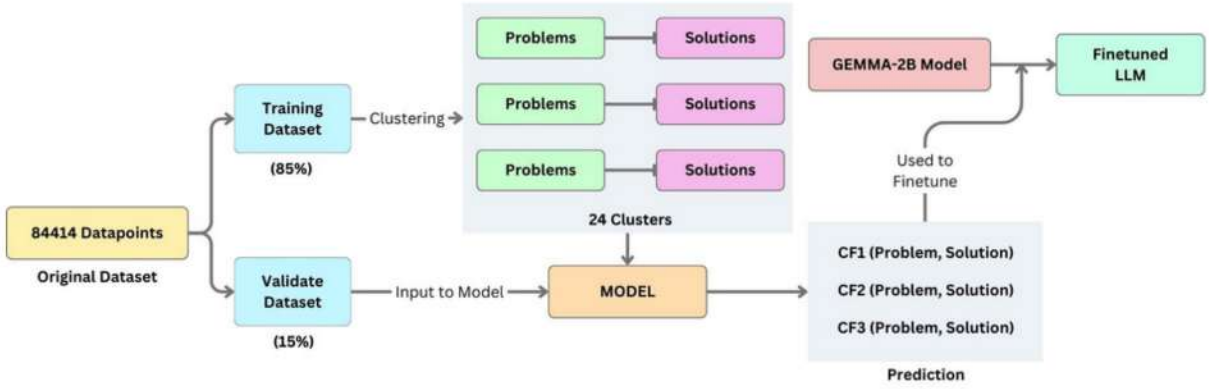


Figure 12: Model pipeline for fine-tuning Gemma-2B

5. MULTI-LLM APPROACH

The approach aims to train various models on 330 parameters categorized into age, gender, race, veteran status, education level, and income and perform real-time ensembles based on the input. For instance, when specific attributes such as age, gender, or income are provided, the model selects the right clusters comprising relevant features and is trained exclusively on those. Consequently, distinct clustering is performed based on the varying feature selections corresponding to different input specifics provided.

This approach allows for targeted model training based on the input data, enabling more accurate and efficient predictions. By focusing on relevant clusters of features, the models can better capture the nuances of the data and provide more meaningful insights for decision-making.

6. RAG-BASED APPROACH

A dataset containing detailed information on important parameters, such as the effects of age and gender on these parameters, was utilized. When a ZIP code input is received, parameter values are retrieved from the ZIP code dataset, and relevant information about parameters is retrieved from the created database. The entire context is then passed to Llama2 for the problems and solutions of that respective ZIP code.

ENSEMBLING

An ensemble methodology was designed to combine problem-solving insights from three different fine-tuned large language models into a single, cohesive output. Querying the Gemma-7B model hosted on Hugging Face's API aims to integrate nine sets of problems and solutions—three from each model—into three unified sets. The process involves sending a payload that includes combined outputs from these models and then reformatting them to create a streamlined presentation of problems and their corresponding solutions, filtering out extraneous information.

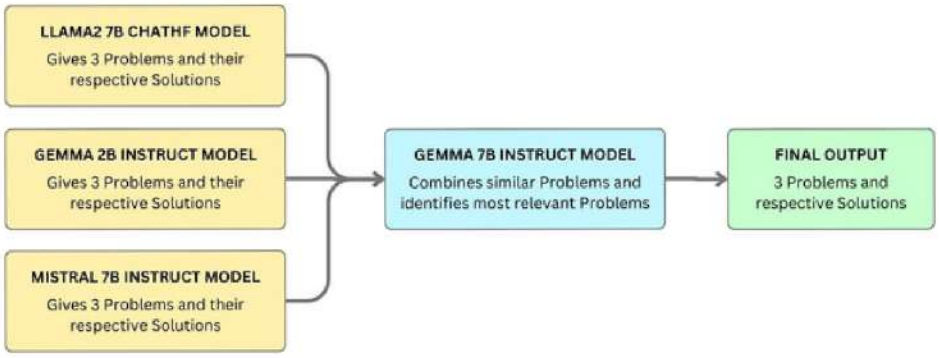


Figure 13: Flowchart depicting the ensembling approach

CONNECTING PEOPLE WITH RESOURCES

The 3 problems obtained from ensembling the models are passed through a BERT model finetuned using a synthetic dataset. The synthetic dataset consists of two columns and 1000 rows, where the first column contains a set of 3 problems, and the second column contains the categories of these problems. This BERT model gives various categories as output (like transportation, age, income, disability, etc…), and each of these categories is mapped to different categories of CSV. A database of CSV files is created for multiple data types(like geriatric care, elder law attorneys, home care, etc…) for each zip code using web scraping. These cases are analyzed, and the nearest favorable hospitals, home care,..(from the user's address) are found from these CSVs and their contacts, and addresses are displayed in the action based scorecard.

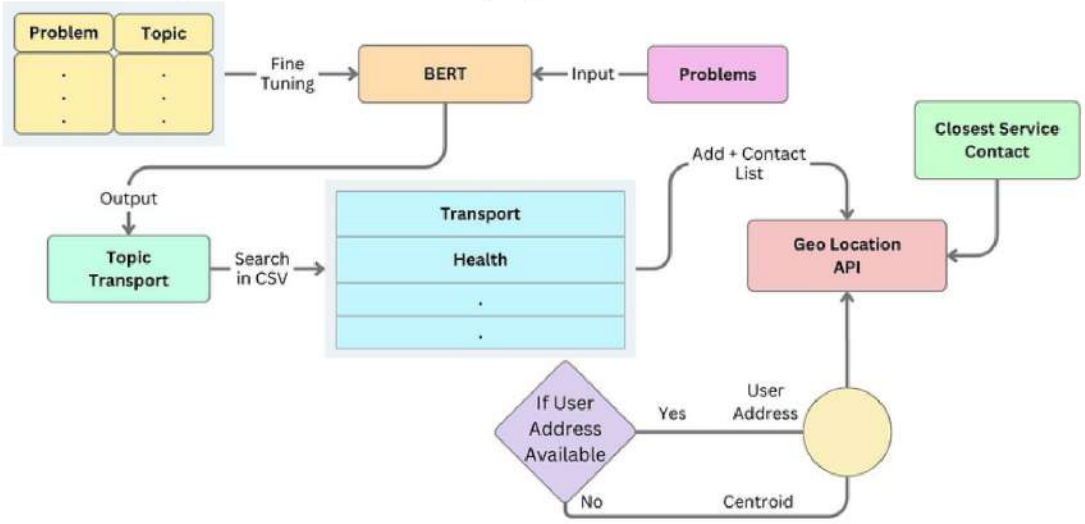


Figure 14: Pipeline for generating role-based action scorecard

SCORECARD

The social care scorecard goes beyond traditional medical data to consider an individual's social context. It uses this information to identify potential health risks and social needs, prompting proactive interventions. This shift from reactive to preventive care ensures people receive the right support before complications arise.

CORE SOCIAL SCORECARD

A social scorecard is a tool used to identify *High-Risk Social Needs(HRSNs)* of individuals and communities based on various social factors that can impact their health. It is considered valuable for informed decision-making and allocating resources to address the root causes of health disparities and improve the overall health and well-being of communities. This scorecard was structured into three main sections, each focused on different aspects of SDOH.

1. Identifying Community Needs:
In this section, the three problems that impact the health of individuals are being addressed, and three action-based suggestions for these problems are being provided.

2. Analyzing SDOH Domains:
Seven domains of social determinants, as mentioned earlier, are analyzed, utilizing the screening tool and several research papers such as PRAPARE, given AHCM's first evaluation report, and socioeconomic health scores by Lexis Nexis Risk Solution.

3. Scoring Methodology:
In assessing Social Determinants of Health (SDOH), 52 parameters across 7 domains are identified. Optimal scores for each parameter are determined based on research papers. These scores are standardized by dividing the optimal value by the median of each parameter, ensuring no single parameter significantly impacts the overall score.

The standardized values are used as multiplying factors to scale the difference between the optimal and actual parameter values. The final SDOH score for each parameter is calculated by subtracting the actual value from the optimal value, applying the multiplying factor, and scaling to 10. Negative values are handled by multiplying by -1 if the optimal value is lower than the actual value.

For each domain, the average of all parameter scores is taken to arrive at the domain's SDOH score. Finally, the overall SDOH score is computed by averaging the scores of all domains, providing a comprehensive assessment of the social determinants impacting health.

ROLE-BASED ACTION SCORECARD

The role-based action scorecard serves as an extension to the core social scorecard by delineating solutions for each identified problem and providing contact information specific to the problem and area, facilitating further assistance and collaboration on individualized solutions. This feature allows for a more detailed understanding of the actions needed to address the issues at hand.

MODEL TESTING & EVALUATION

A web application has been developed to test and evaluate the effectiveness of the models and ideas presented in the solution. The application provides a detailed analysis of the problems faced by the individual and forecasts risk scores across various domains.

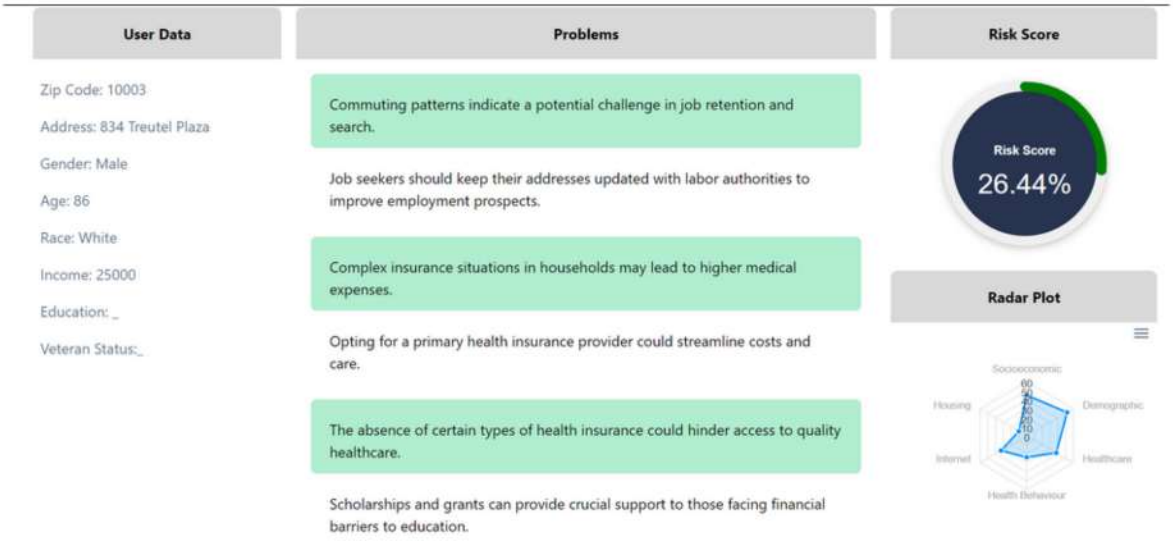


Figure 15: Core-social scorecard

Each individual is assigned a risk score, calculated using the models' probabilistic output. Additionally, a radar plot is used to visually represent the relative importance of various factors within the problem superset (including demographic, socioeconomic, and housing factors) in determining tailored solutions to meet the needs of the patients, ensuring that interventions are targeted and effective, addressing the specific challenges faced by each individual.

Contacts

Problem Specific Contacts		
Alone Elder Care LLC	(856) 497-3512	125 Cedar St Colonia NY
R.A.I.N. Alzheimer's Caregiver	(347) 346-9676	White Plains Rd Bronx NY
New York City Department for Aging	(212) 639-9675	2 Lafayette St New York

Figure 16: Role-based action scorecard

In this system, patients receive targeted services tailored to their needs. Additionally, they are provided with the service provider's contact details. This approach aims to remove any barriers to connectivity and facilitate a faster resolution of their issues. For each input, three contact options are provided, enhancing accessibility and ensuring patients can easily reach out for assistance.

THE INDIA CONNECTION

Concerted efforts have been undertaken to extract detailed information from public census abstracts through scraping, aiming to cultivate a thorough understanding of the subject matter. In this approach data-driven insights were used for the betterment of public health in India, laying the foundation for a healthier future through comprehensive and nuanced analyses.

• Challenges in Building a Predictive Healthcare System in India

Comprehensive Data Accessibility Issues	Inconsistencies in Data Collection
Outdated Census Information	Limited Detail Hinders Analysis

• Innovative Data Sources for Improved Predictions

By exploring these innovative data sources and addressing the challenges outlined below, we can move towards building a more robust and comprehensive predictive healthcare system for India.

Primary Health Centers
Strengthening data collection efforts at PHCs can provide valuable insights into disease prevalence and the effectiveness of healthcare interventions at the grassroots level.

Satellite Data for Health Parameters
Using satellite data to determine health parameters indirectly, including monitoring environmental factors, disease vectors, and population density for more accurate predictions.

Cultural and Traditional Medicine Data
Including data on traditional medicine practices allows us to understand their prevalence and effectiveness within cultural contexts, informing future healthcare interventions.

NGO & Community Organization Data
Collaboration with NGOs and community-based groups allows access to data on grassroots health initiatives, community health education programs, and local health challenges.

mHealth for Urban Areas
Leveraging data from mHealth applications, wearable devices, and remote monitoring devices offers insights into health behaviors, disease patterns, and treatment adherence.

Social Media Sentiment Analytics
Analyzing public sentiment on social media reveals public perceptions regarding healthcare that provide real-time insights into emerging health issues and public awareness.

Telemedicine Consultation Records
Analyzing data from telemedicine consultations allows us to understand healthcare-seeking behavior, prevalent health concerns, and the effectiveness of remote healthcare services.

Weather and Environmental Data
Integrating weather and environmental data can help identify correlations between climatic conditions, pollution levels, and health outcomes.

CONCLUSION AND DISCUSSION

CONCLUSION

This report systematically addresses complex healthcare access challenges faced by the people through a strategic & comprehensive approach. Here's a conclusion of the work:

Our approach involved the utilization of robust data handling techniques such as Correlation-Based Imputation with DataWig and KNN imputation, followed by data normalization and quartile transformation to ensure uniformity and preserve statistical properties. Advanced methodologies like Random Forest and feature reduction based on multicollinearity were employed to identify key features impacting health parameters. By categorizing features into domains and employing K-Means clustering, granular insights into problem areas were gained, allowing for tailored solutions based on the highest probability subclusters. This methodology offers a systematic and effective approach to addressing health-related challenges.

Different techniques were employed for the problem and solution generation task on the refined dataset. The power of various LLMs (Llama2 7b, Mistral 7b, Gemma 2b) was leveraged to analyze potential problems and solutions. The model was fine-tuned on the training dataset values of zip code parameters to enable predictions of possible problems and solutions for different parameters. A sensitivity analysis of parameter values and model output was conducted, and it was observed that our ensemble model outperformed others.

In the task of connecting people with resources, the BERT model was fine-tuned to recognize problem topics in different sections, such as healthcare, transportation, etc. Subsequently, the geolocation API was utilized to extract the exact distance from the user's address to the nearest appropriate resources.

FUTURE NEEDS FORECASTING

Analysis of data from 2018 to 2020 across census tracts highlights key societal challenges and intervention strategies:

1. **Health Insurance:** The decrease in health insurance coverage in 10,483 census tracts signals rising health disparities and healthcare system strain. Solutions include initiating public awareness campaigns and offering financial assistance to improve insurance accessibility.

2. **Transport Facilities:** The increase in workers with commutes over 60 minutes, noted in 10,826 census tracts, points to issues like traffic congestion and commuter dissatisfaction. Addressing this requires Transit-Oriented Development (TOD) and more investment in public transportation.

3. **Elder Population:** The growth of the population aged 65+ in 13,632 census tracts highlights upcoming healthcare demands and economic burdens. Aging-in-place initiatives and expanding elderly healthcare services are crucial responses.

These insights inform targeted policy and intervention needs in health insurance, transportation, and elder care to address emerging societal issues.

ANNEXURE

SOFTWARE STACK

Library	Version	Purpose
Python	3.7.16	Language of choice
Numpy	1.26.4	To perform mathematical operations on arrays
Pandas	2.2.1	Data handling, manipulation and analysis
Matplotlib	5.19.0	Plotting visualizations
Seaborn	0.13.2	For making statistical graphics
Geopandas	0.14.3	For plotting data on geographic maps
Ploty	5.19.0	For Data Visualization
Datawig	0.2.0	For imputing null values
Sklearn	1.4.0	For Machine Learning and statistical modelling
Selenium	4.18.1	Used for automating web browsers for scraping purposes
Beautifulsoup4	4.12.3	For pulling data out of HTML code from a website
Requests	2.31.0	For making HTTP requests and working with APIs
Huggingface	-	For importing LLM models

MORE ON FEATURE ENGINEERING

K Nearest Neighbours

KNN imputation is designed to find k nearest neighbors for a missing datum (incomplete instance) from all complete instances (without missing values) in a given dataset, and then fill in the missing datum with the most frequent one occurring in the neighbors if the target feature (or attribute) is categorical, referred to as majority rule, or with the mean of the neighbors if the target feature is numerical, referred to mean rule.

Datawig Imputation

Datawig is an open-source machine learning library designed for imputing missing values in tabular data through deep learning techniques. It automates the process of predicting missing entries by learning patterns within the data, handling various data types including numerical, categorical, and text. Datawig stands out for its ability to capture complex relationships between columns in a dataset, making it a versatile tool for a wide range of applications, from healthcare to customer analytics. Despite its advantages in flexibility and performance, it requires considerable computational resources and a substantial dataset for effective training, posing challenges in some scenarios.

Multiple Imputation by Chained Equations (MICE)

It is a statistical technique for addressing missing data by iteratively imputing missing values across multiple variables, thereby creating several complete datasets from an incomplete one. This method assumes that the missing data are Missing At Random (MAR) and employs a chained, or sequential, approach where each variable with missing data is imputed based on observed and previously imputed values.

The process incorporates randomness to reflect the uncertainty associated with imputing missing values. After generating multiple imputed datasets, standard statistical analyses are applied to each, and the results are combined, providing estimates that account for the variability due to missing data. MICE enables more accurate statistical inferences by acknowledging and addressing the uncertainty inherent in the imputation of missing data.

The techniques used for correlation are:

Correlation Matrix

A correlation matrix is a statistical table that shows the correlation coefficients between sets of variables. Each cell in the matrix shows the correlation between two variables. The values range from -1 to +1, indicating the strength and direction of the relationship: values close to +1 or -1 signify strong positive or negative correlations, respectively, while values near 0 indicate little to no linear relationship.

Variance Inflation Factor (VIF)

This is a measure used to detect the presence and severity of multicollinearity, where multiple independent variables in a regression model are correlated. VIF assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated.

If no multicollinearity exists, VIF is 1; values above 1 indicate multicollinearity, with higher values signifying greater inflation. VIF values exceeding 5 or 10 suggest a concerning level of multicollinearity that may necessitate corrective measures, such as removing correlated predictors from the model.

MORE ON DATA NORMALIZATION

The techniques used for Data Normalization are

Min Max Scalar

This technique scales the data to a fixed range, usually between 0 and 1.The formula for min-max scaling is: Xnorm = (X-Xmin)/(Xmax-Xmin). This was done but the Results of the metrics such as Standard Deviation, Skewness and Kurtosis are less than the threshold.

Z Score Normalization

This technique scales the data to have a mean of 0 and a standard deviation of 1.The formula for the Z-Score Standardization is: Xstd = (X - u)/std where X is the original value, u is the mean of the feature, std is the Standard Deviation of the feature. This method has solved the issue of Standard Deviation threshold but Results of the metrics such as Skewness and Kurtosis are less than the threshold.

Robust Scaling

This technique is similar to min-max scaling but uses the interquartile range (IQR) instead of the minimum and maximum values to scale the data. The formula for the robust scaling is: Xrobust = (X-Q1)/(Q3-Q1) where X is the robust value, Q1 is the first quartile and Q3 is the third quartile. This was done but the Results of the metrics such as Standard Deviation, Skewness and Kurtosis are slightly less than the threshold.

Log Transformation

This technique applies a logarithmic function to the data to make it less skewed. The formula for the log transform is: Xlog = log(X). This method has solved the issue of Skewness and Kurtosis threshold but Results of the metrics such as Standard Deviation are less than the threshold.

Quantile Transform

This technique is used to normalize the distribution of data across different samples or datasets. It ensures that the distributions of the variables are similar across samples, making them comparable. This technique is commonly used in microarray analysis and other high-throughput genomic data analysis to remove systematic variations between samples.

SELECTED LIST OF FEATURES TABLE

Name of Feature	Short Description
ACS_PCT_HH_SMARTPHONE_ONLY	Percentage of households with a smartphone with no other type of computing device (ZCTA level)
ACS_PCT_AGE_15_17	Percentage of population between ages 15-17
ACS_PCT_WHOLESALE	Percentage of employed working in wholesale trade (ages 16 and over)
ACS_PCT_PVT_EMPL_DRCT	Percentage of population with employer-based and direct-purchase coverage
ACS_PCT_HH_OTHER_COMP	Percentage of households with other type of computer
ACS_PCT_MARRIED_SP_AB_M	Percentage of male population now married and spouse absent (ages 15 and over)
ACS_PCT_MARRIED_SP_AB_F	Percentage of female population now married and spouse absent (ages 15 and over)
ACS_PCT_OTHER	Percentage of employed working in other services, except public administration (ages 16 and over)
ACS_PCT_VET_COLLEGE	Percentage of civilian veterans that have some college education or an associate's degree (ages 25 and over)
ACS_PCT_OTH_EURP	Percentage of population who speak other Indo-European languages (ages 5 and over)

ACS_PCT_OTHER_INS	Percentage of population with other health insurance coverage combinations
ACS_PCT_IN_COUNTY_MOVE	Percentage of population that moved within the same county in the past year (age 1 and over)
ACS_PCT_FINANCE	Percentage of employed working in finance and insurance, real estate, and rental and leasing (ages 16 and over)
75-84_e(x)	Expectation of life between ages 75 and 84
ACS_PCT_ADMIN	Percentage of civilian employed population working in public administration (ages 16 and over)
ACS_PCT_IN_STATE_MOVE	Percentage of population that moved from different county within same state in the past year (age 1 and over)
ACS_PCT_TRANSPORT	Percentage of employed working in transportation and warehousing, and in utilities (ages 16 and over)
ACS_TOT_GRANDCHILDREN_GP	Total grandchildren under 18 living with grandparent householder
ACS_PCT_CONSTRUCT	Percentage of employed working in construction (ages 16 and over)
ACS_PCT_INFORM	Percentage of employed working in information services (ages 16 and over)
ACS_PCT_HH_BROADBAND_ONLY	Percentage of households with broadband such as cable, fiber optic or DSL with no other type of Internet subscription
ACS_PCT_PROFESS	Percentage of employed working in professional, scientific, management, administrative, and waste management services (ages 16 and over)
ACS_TOT_POP_16_19	Total population (between ages 16 and 19)
ACS_PCT_MULT_RACE	Percentage of population reporting multiple races
ACS_PCT_PRIVATE_SELF	Percentage of population with direct-purchase health insurance only
POS_DIST_CLINIC_TRACT	Distance in miles to the nearest health clinic (FQHC, RHC), calculated using population weighted tract centroids
ACS_PCT_HH_SAT_INTERNET	Percentage of households with satellite internet service

ACS_PCT_PVT_NONPROFIT	Percentage of population who are private not-for-profit wage and salary workers (ages 16 and over)
HIFLD_DIST_UC_TRACT	Distance in miles to the nearest urgent care, calculated using population weighted tract centroids
85 and older_e(x)	Expectation of life older than 85 years of age
ACS_PCT_WORK_RES_F	Percentage of female population who worked in place of residence (ages 16 and over)
ACS_PCT_BLACK	Percentage of population reporting Black or African American race alone
ACS_PCT_COMMT_60MINUP	Percentage of workers with at least 60-minute commute time (ages 16 and over)
ACS_PCT_MANUFACT	Percentage of employed working in manufacturing (ages 16 and over)
ACS_PCT_HU_OIL	Percentage of occupied housing units with fuel oil heating
ACS_PCT_CTZ_NATURALIZED	Percentage of population consisting of U.S. citizens by naturalization
OBESITY_Data_Value	Obesity among adults aged >=18 years
MAMMOUSE_Data_Value	Mammography use among women aged 50-74 years
DEPRESSION_Data_Value	Depression among adults aged >=18 years
CHOLSCREEN_Data_Value	Cholesterol screening among adults aged >=18 years
COLON_SCREEN_Data_Value	Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50-75 years
ACS_PCT_TAXICAB_2WORK	Percentage of workers taking taxicab, motorcycle, bicycle, or other means to work (ages 16 and over)
ACS_PCT_HH_INTERNET_NO_SUBS	Percentage of households with internet access without a subscription
ACS_PCT_CHILD_DISAB	Percentage of children with a disability (ages 17 and below)
ACS_PCT_AGE_0_17	Percentage of population between ages 0-17
ACS_PCT_AGE_30_44	Percentage of population between ages 30-44
ACS_PCT_AGE_ABOVE65	Percentage of population ages 65 and over
ACS_PCT_MALE	Percentage of population that is male

ACS_PCT_VET	Percentage of civilian veterans with a disability (between ages 18 and 64)
ACS_PER_CAPITA_INC	Per capita income (dollars, inflation-adjusted to data file year)
ACS_PCT_LT_HS	Percentage of population with less than high school education (ages 25 and over)
ACS_MEDIAN_RENT	Median gross rent (dollars)
ACS_PCT_HH_1PERS	Percentage of households with only one occupant
ACS_PCT_UNINSURED	Percentage of population with no health insurance coverage
POS_DIST_ED_TRACT	Distance in miles to the nearest emergency department, calculated using population weighted tract centroids

MORE ON CLUSTERING

The techniques used for Clustering are:

K-Means Algorithm

K-Means is a widely used clustering algorithm that partitions a dataset into a pre-defined number of clusters. It iteratively assigns data points to the nearest cluster centroid based on Euclidean distance and updates centroids until convergence. While efficient and straightforward, K-Means is sensitive to initial centroid selection and may converge to local minima, affecting clustering quality.

Gaussian Mixture Model (GMM)

GMM is a probabilistic clustering algorithm that assumes data comes from a mixture of Gaussian distributions, each representing a cluster. It employs the Expectation-Maximization (EM) algorithm to iteratively refine parameters such as means, covariances, and mixture weights. However, GMM's performance heavily relies on initial parameterization and may not handle complex data distributions well.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN clusters data based on density, categorizing points as core, border, or noise points. It is robust to outliers and can identify clusters of arbitrary shapes and sizes. DBSCAN does not require specifying the number of clusters beforehand but may struggle with data of varying

D. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

HDBSCAN is an extension of DBSCAN that leverages hierarchical structures to identify clusters of varying shapes and densities. It automatically determines the optimal number of clusters and is robust to varying densities and noise. HDBSCAN excels in handling irregularly shaped clusters but may not provide uniformly sized clusters.

MORE ON IMPUTATION

One hot encoding was used for the following cases:

1) **Gender**
If the gender is 'M,' set the feature 'ACS_PCT_MALE' value as 100. If the gender is 'F,' set the feature 'ACS_PCT_MALE' value as 0.

2) **Age**
When a user has given his age, depending upon in which age interval it lies, we set that feature value as 100 and the rest as zero. As our final_features dataset contained age columns:--

1) ACS_PCT_AGE_0_17 age [0,17]
2) ACS_PCT_AGE_15_17 age [15,17]
3) ACS_PCT_AGE_30_44 age [30,44]
4) ACS_PCT_AGE_45_64 age [45,64]
5) ACS_PCT_AGE_ABOVE65 age[>65]

REFERENCES

- https://www.cms.gov/priorities/innovation/data-and-reports/2020/ahc-first-eval-rpt
- https://patents.google.com/patent/WO2020049404A2/en
- https://www.census.gov/programs-surveys/acs/data.html
- https://link.springer.com/article/10.1007/S40745-015-0040-1
- https://www.biorxiv.org/content/10.1101/012203v1.full