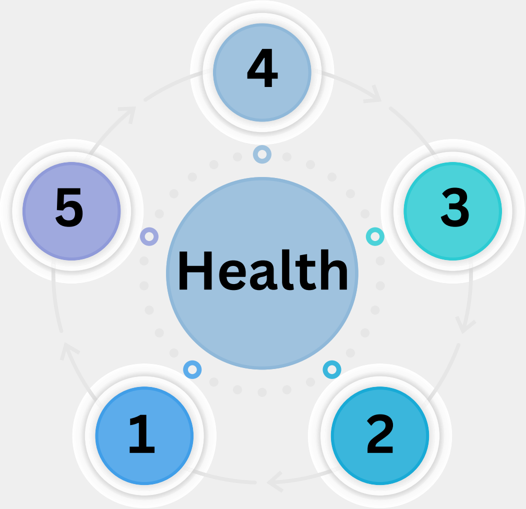
Team : DA24

Table of Contents

SDOH: Social Determinants of Health

Social determinants of health (SDOH), such as where we live, income, education, race, age, and support systems have a profound impact on our health and well-being. These factors contribute to health disparities and are strong predictors of both individual and community-level health outcomes. Understanding an individual's SDOH is crucial, especially when providing care for those with disabilities, chronic conditions, or when supporting aging loved ones.

SDOH are vital for reducing health disparities and achieving equity by recognizing the impact of social and economic factors on health. They influence chronic disease development, treatment outcomes, mental health, and well-being. Addressing SDOH helps reduce healthcare costs by preventing unnecessary emergency visits and readmissions. Additionally, understanding SDOH enables a preventive approach to health, targeting root causes rather than solely treating diseases as they arise.



1: Economic Stability
2: Neighbourhood and Physical Environment
3: Education and Food
4: Community and Social Context
5: Health Care System

How are SDOH currently used?

Patient Navigators (PNs) use standardized inventories or may ask patients to self-report on their social determinants. PNs or physicians then coordinate with internal team members or other organizations to address identified needs and provide referrals. Some physicians customize interventions based on social needs, such as prescribing longer medication supplies for individuals in rural areas.

Challenges faced with using SDOH at present

- **Feasibility**: Time constraints and insufficient reimbursement hinder effective SDOH survey administration.
- **Patient Reluctance**: Reluctance to share personal information and communication barriers limit screening.
- **Binary Evaluation Limitations**: Common SDOH inventories overlook nuanced differences within variables, like varied risk levels among different groups.
- **Ignoring Interdependencies**: Existing tools evaluate SDOH factors in isolation, missing interconnected issues like transportation, work commute, and income.
- **Actionable Guidance Deficit**: Healthcare providers lack clear, actionable steps to address identified social determinants in clinical decisions.
- **Lack of Self-Service Tools:** There's a gap in empowering patients to navigate SDOH challenges independently, beyond generic tools like Google Maps.

Objectives of the 2024 Community Companion Challenge

Our quest is to use the power of data science to provide a new way of understanding social determinants without relying solely on asking the individual about their needs.

- **The Heart of the Challenge - The Social Care Scorecard:** Develop a data-driven tool to assess individual social support needs and health risks, utilizing predictive modeling based on available information. This includes flagging needs like transportation and healthy food options and creating an action-based version to guide healthcare professionals in taking appropriate action.
- **Making a Real-World Difference - Connecting People with Resources:** Pool relevant information and resources to complement the social care scorecard, facilitating the identification of the best course of action and matching individuals with essential support services.
- **India Connection:** Identify innovative data sources to power a predictive system tailored to India's unique healthcare challenges, where social factors play a crucial role in shaping health outcomes. This initiative aims to make a lasting impact on healthcare delivery in India.

Data Sources

These are the provided Data Sources, which we are using for creation of our model:

1. **SDOH and Health Data Codebook:** Provides detailed metadata about variables used in our analysis, including definitions, frequency, and relevance to the study.
2. **Census Tract Data:** Contains files for 2020, 2019, and 2018, offering specific variables across Census Tracts, enabling insights into geographical trends.
3. **Zip Code Data:** Offers files for 2020, 2019, and 2018, with variable-specific information across Zip Codes, providing a detailed understanding of social determinants at this level.
4. **County Data:** Available for 2020, 2019, and 2018, with specific variables across Counties, facilitating analysis of broader geographical trends and comparisons.
5. **Life Expectancy Data:** Provides county-level information on life expectancy values and probabilities, offering insights into health outcomes across regions.
6. **Raw Health Data by Census Tract:** Contains patient and disease information linked to Census Tracts, aiding in understanding health outcomes within specific geographical areas.
7. **Reference File:** Includes three questionnaires (PRAPARE, Health Leads Screening Toolkit, AHMC Health Social Need Screening Tool) for assessing individual health risks based on responses.
8. **Synthea Data Extract:** Comprises dummy data for training purposes, offering simulated patient information to enhance model development and analysis.

Purpose:

- These datasets collectively provide comprehensive information for analyzing social determinants of health, geographical trends, and health outcomes, enabling informed decision-making and interventions in healthcare provision.

The Heart of the Challenge: The Social Care Scorecard

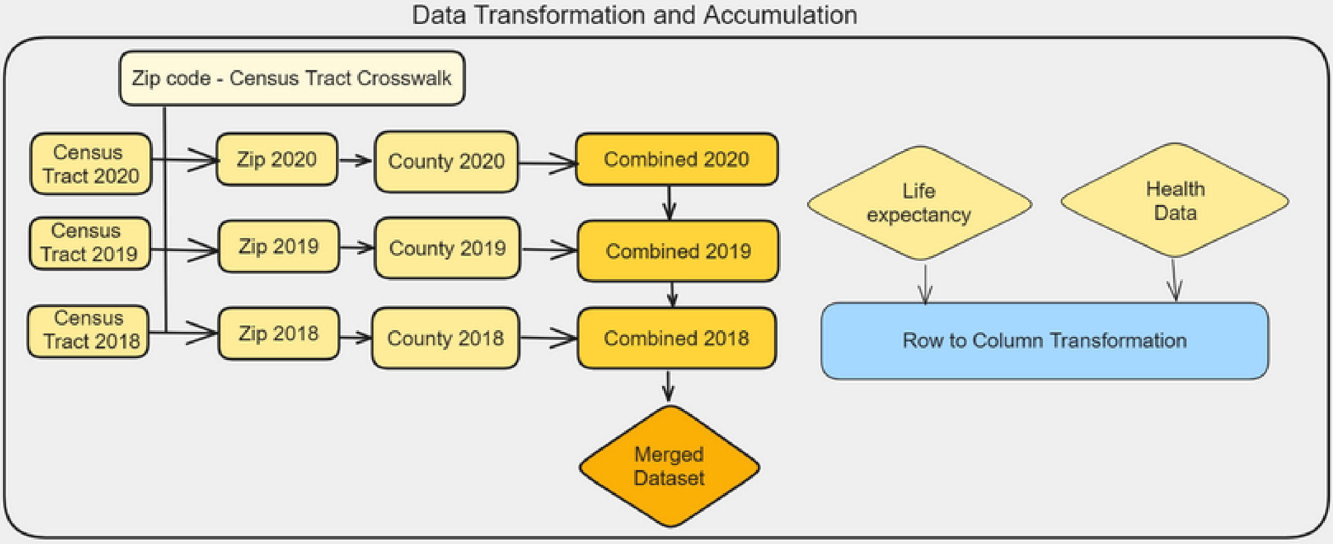
Data Curation

Our initial task was to merge the **zipcode, county,** and **census tract** data books from 2018, 2019, and 2020. However, this process presented **several challenges**:

- **Lack of a Primary Column**: We faced difficulties due to the absence of a primary column for seamless merging of census tract data with zip code data. **To address this**, we utilized a **zip-code census track crosswalk dataset** to establish the relationship between census tract and zip code.
- **Variable Naming Discrepancies:** Another obstacle was the inconsistency in variable names between the zipcode and census tract datasets, even though they denoted the same categories. **To resolve this,** we standardized the variable names to ensure uniformity.
- **Handling Multiple Overlaps**: Some census tracts extended across multiple zip codes, adding complexity to the merging process. **We managed this** by utilizing a **crosswalk Dataset** and creating a unique combination of tractFIPS and zip code.

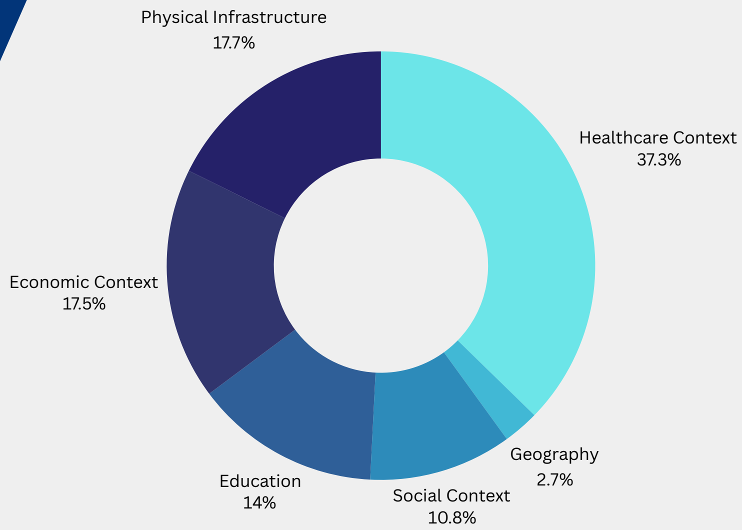
Additionally, during data curation, we prioritized the most recent values. If a variable was missing in 2020, we sourced it from the previous year's dataset. If the variable was absent in all three years, we omitted it from the analysis.

Additionally, we undertook transformations on the **life expectancy** and **raw health data**. This involved creating a condensed dataset illustrating the distribution of specific health indicators across census tracts.

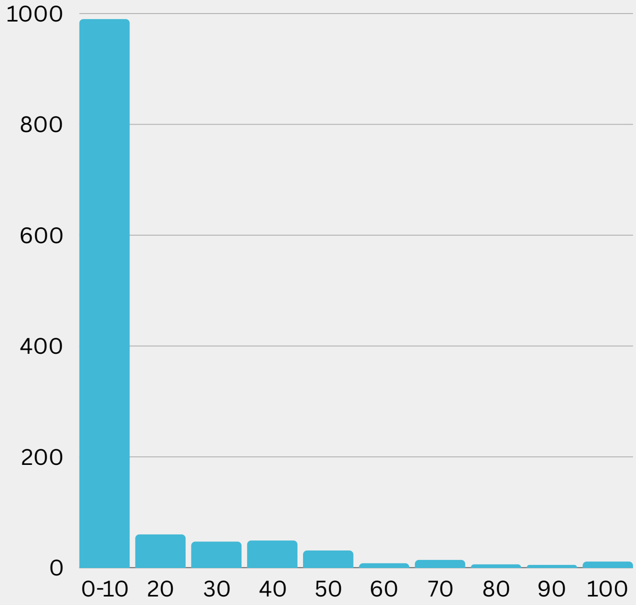


Roadmap of Data Curation

EDA: Exploratory Data Analysis



Distribution of available data for the different domains of SDOH.



Distribution of variables by their null value percentages.

Input Parameter Mapping

Our subsequent task involves mapping the labels of Social Determinants of Health (SDOH) variables to the input parameters obtained from the user, namely zip code, age, gender, income, education, race, veteran status, and address. We employ advanced natural language processing techniques such as few-shot Large Language Models (LLMs) prompting, Sentence-BERT, and cosine similarity calculations.

Furthermore, we delve deeper into the sub-details of each input parameter, such as race, by utilizing Regular Expressions (RegEx). This enables us to map SDOH variables to specific demographic categories, such as "Blacks," "Hispanics," "Whites," and others, within the input parameters. By employing this comprehensive approach, we aim to establish a robust framework for associating SDOH variables with user-provided demographic information, facilitating a more nuanced understanding of the social determinants impacting individual health outcomes.

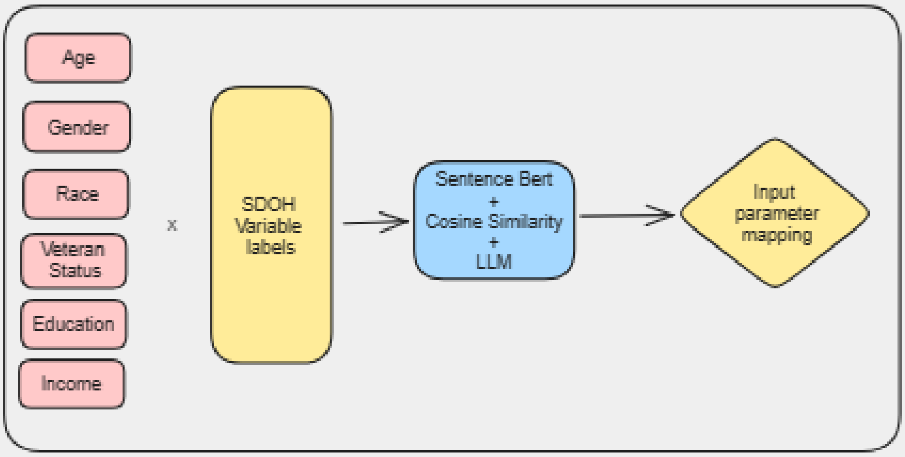
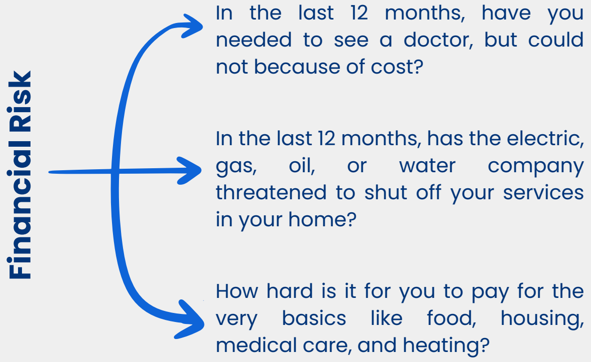


Fig. Schematic representation of input parameter mapping to SDOH variables

Risk Identification

Moving forward, our next objective is to identify potential risks that users may encounter in their area. To achieve this, we initiated a clustering process using data from three questionnaires that are **PRAPARE ,Health leads Screening_toolkit, AHMC-screening tool**. Through this clustering analysis, we identified 11 distinct risk categories:

- Educational Challenges
- Food Security
- Lifestyle
- Transportation Risk
- Financial Risk
- Technology Access Risk
- Social Environment Risks
- Disease Risk
- Housing Challenges
- Climate Risk
- Healthcare Access Risk



In the last 12 months, have you needed to see a doctor, but could not because of cost?

In the last 12 months, has the electric, gas, oil, or water company threatened to shut off your services in your home?

How hard is it for you to pay for the very basics like food, housing, medical care, and heating?

Feature Classification based on the Risk

Our subsequent task involves mapping the labels of Social Determinants of Health (SDOH) variables to the 11 identified risks. To accomplish this, we utilize a Sentence Transformer named "**multi-qa-MiniLM-L6-cos-v1**" along with **tf-idf label encoding** techniques. This enables us to create encodings for the variable labels, facilitating the establishment of semantic embeddings. Subsequently, we sort these embeddings based on their cosine similarity scores with each risk category.

Clustering:

Once the closest variable related to a particular risk is identified, we proceed to cluster the variable labels within that risk category. This process results in the formation of mini-clusters within each risk, providing a more nuanced understanding of the associations between SDOH variables and the identified risks. By employing this approach, we aim to enhance our ability to discern and address the various social, economic, and environmental factors contributing to health disparities and risks within specific geographical areas.

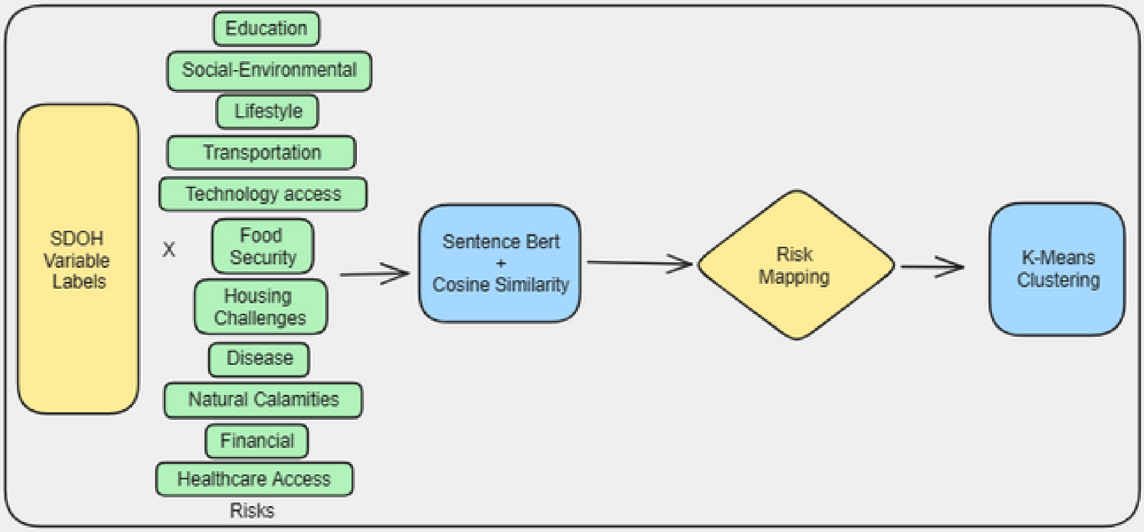
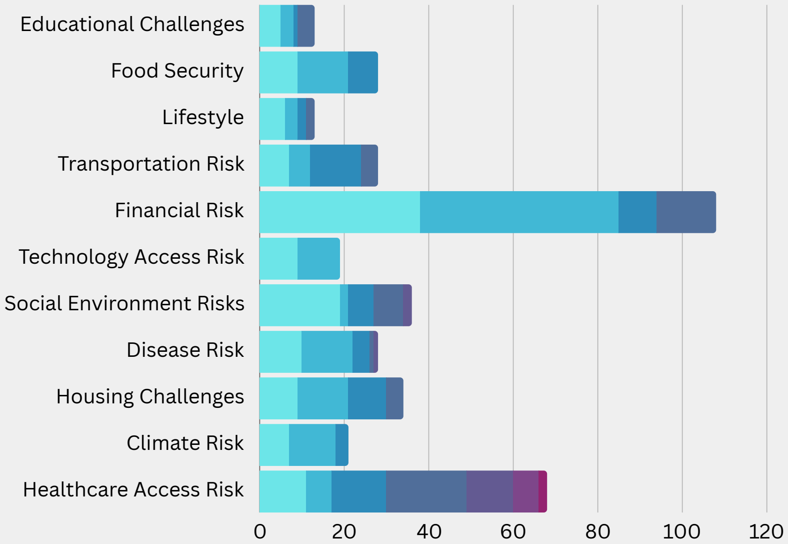


Fig. Schematic Representation of Risk (Identified) mapping to SDOH Variables

EDA: Exploratory Data Analysis on Feature Classification

This plot shows the division of the variable over the formed cluster for each of the determined risks. This explains the overall spread of the variables on clusters that our model has formed and these variables have been placed by the model to form these clusters. This indicates the correlation between variable potentially leading to a similar risk arising in the patient.

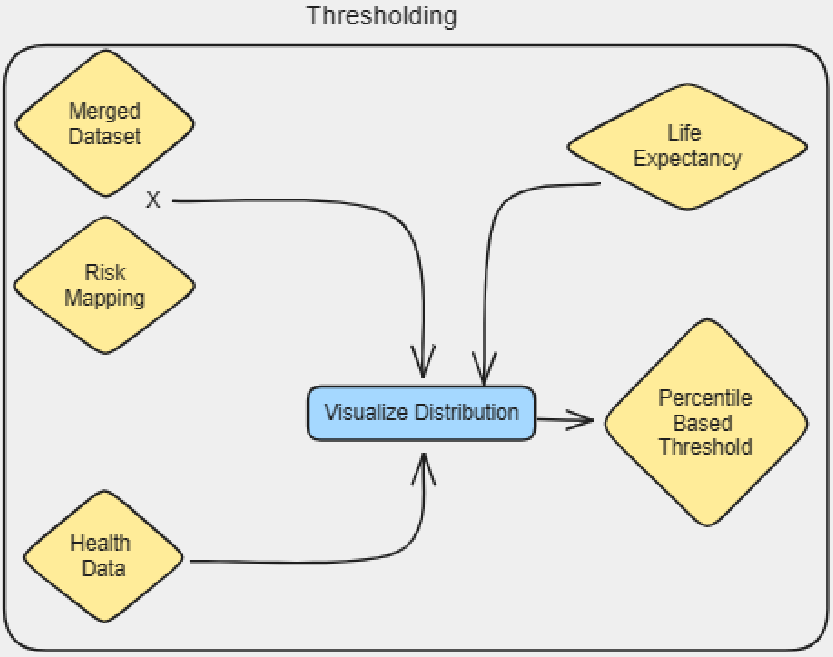


Thresholding

Now, each Social Determinants of Health (SDOH) variable undergoes scoring to determine its correlation with specific risks. A score of +1 indicates a positive correlation, suggesting that the risk increases with higher variable values. To establish thresholds, we analyze the frequency distribution of each variable and calculate percentile-based values using the risk-correlation score.

For variables positively correlated with a risk score of +1, we apply a ratio of 3:5:12 to categorize them into high, mid, and low-risk categories. Conversely, for variables negatively correlated with a risk score of -1, we utilize a ratio of 12:5:3 for the same categorization. This systematic approach ensures that SDOH variables are appropriately classified based on their correlation with specific health risks, facilitating targeted interventions and risk mitigation strategies.

Similarly we also perform thresholding in the health data and life expectancy data.



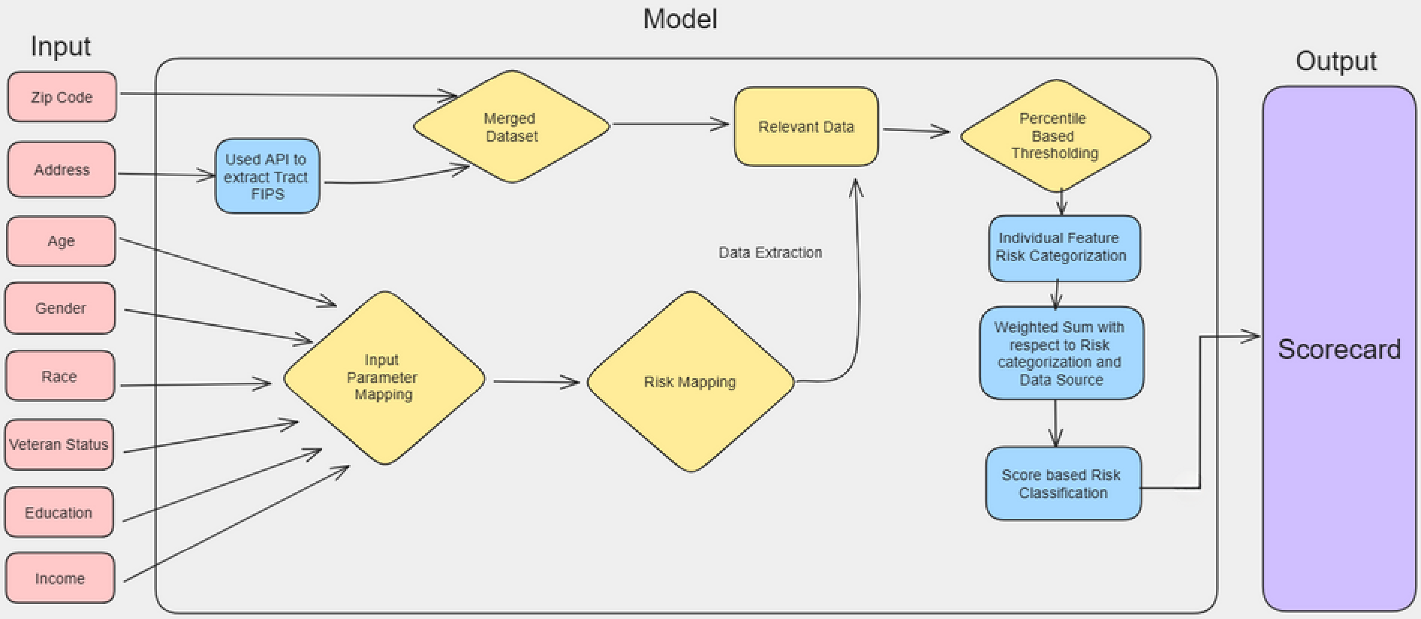
Model Pipeline

Utilizing User Inputs:
Integrate user-provided parameters, including ZIP code and address, to extract tract FIPS codes. In instances where address information is unavailable, aggregate all tract FIPS entries for the provided ZIP code to establish a unique identifier pair for dataset access.

Input Parameter Mapping:
Leverage additional user inputs such as age, race, and gender to extract pertinent Social Determinants of Health (SDOH) variables and associated risks for the specified region. This process utilizes the input-parameter mapping and risk mapping generated previously.

Risk Categorization:
Employ thresholding techniques to categorize identified risks into high, mid, and low categories, utilizing percentile-based quartile division for precise classification.

Weighted Sum Calculation:
Conduct a weighted sum computation, incorporating both risk categorization and SDOH variables, to generate a comprehensive health score reflective of the individual's risk assessment. This holistic approach ensures a thorough evaluation of the individual's health status based on relevant socio-economic and demographic factors.


Model
Input
Zip Code
Address
Age
Gender
Race
Veteran Status
Education
Income
Used API to extract Tract FIPS
Merged Dataset
Relevant Data
Percentile Based Thresholding
Individual Feature Risk Categorization
Weighted Sum with respect to Risk categorization and Data Source
Score based Risk Classification
Data Extraction
Input Parameter Mapping
Risk Mapping
Output
Scorecard

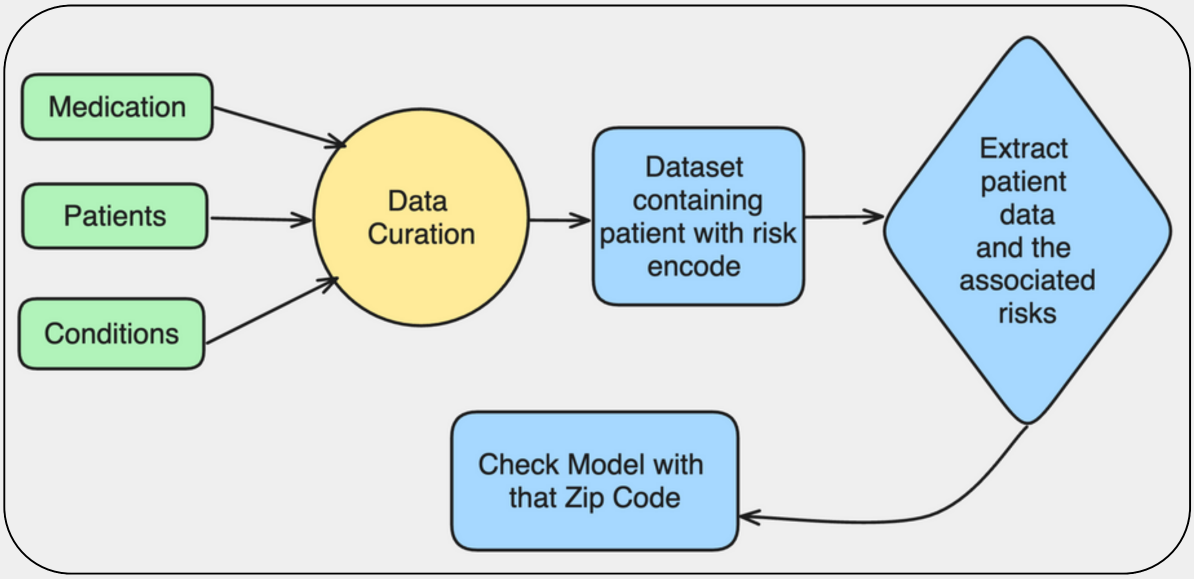
Model Pipeline Testing

To assess how well our model performs, we're using a dataset called Synthea, which provides information on patients' conditions, medications, and general information. The condition file gives us details about when patients and doctors interact, which we call encounters. Meanwhile, the patient file includes essential information about each patient, such as age and address. The medications file describes the medications prescribed to patients.

To evaluate our model's performance, we link the medication and condition information to the 11 risks we identified earlier. This allows us to create a consolidated dataset that combines patient data, zip codes, diseases, and associated risks. With this dataset, we can analyze specific zip codes to determine which risks are more prevalent among patients in those areas. This testing process helps us gauge how effectively our model predicts risks based on demographic and healthcare data.

In simpler terms, we're using a set of patient data to see how well our model works. By connecting this data to the risks we've identified, we can figure out which health concerns are more common in different areas. This helps us understand how accurate our model is in predicting health risks for specific locations.

Example:
We extracted zip codes that have high number of patients with a particular type of risk, for example Transportation Risk, Drug/ Alcohol abuse risk, Lung Disease Risk, etc. When input into the model, these zip codes were flagged for the given risks in most of the cases showcasing that our model works well on real world patients too.


Medication
Patients
Conditions
Data Curation
Dataset containing patient with risk encode
Extract patient data and the associated risks
Check Model with that Zip Code

Flowchart of the Model Pipeline Testing

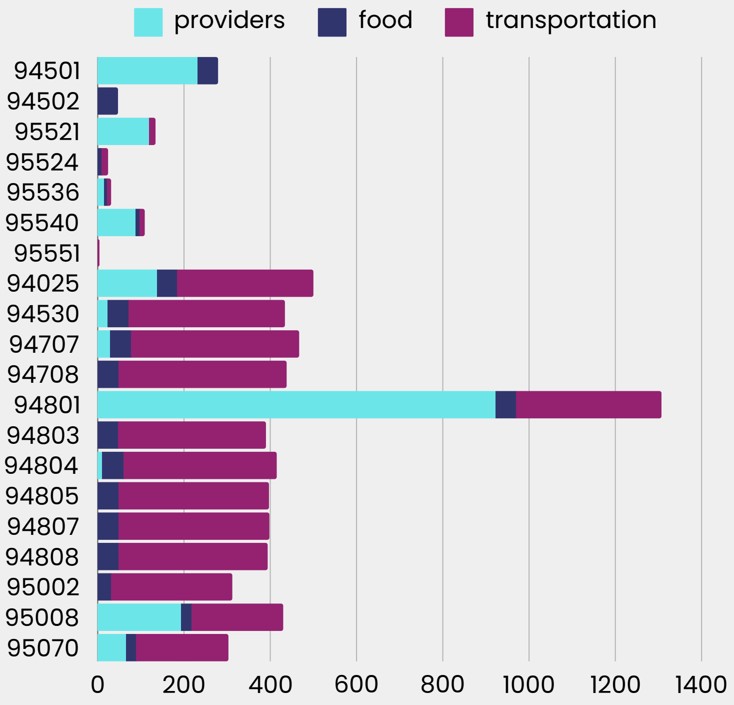
Making a Real-World Difference: Connecting People with Resources

Moving forward, our focus is on building a comprehensive database of community support resources and information for a specified region. Leveraging diverse data sources provided in Appendix 2, we aim to create a robust support database that includes transportation options, physicians, home health agencies, and more. Utilizing our predictive algorithm, we will tailor support suggestions based on an individual's scorecard, ensuring that they receive recommendations only if they are at a medium or high risk for a particular need. Additionally, as an optional stretch goal, we plan to visualize our scorecard and matched findings through two interfaces: one for patients and families and another for physicians and organizations. This will enable us to showcase various scenarios and demonstrate the impact of different user inputs on their social scorecard generated by our predictive model.

Data Scraping and Curation

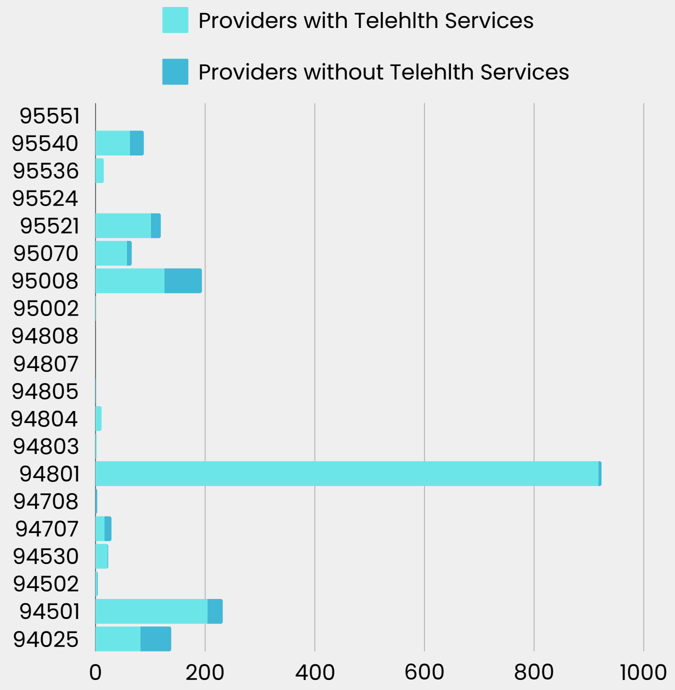
Data scraping was executed using a provided code to collect information on various services including hospitals, home health providers, food services, and transportation. Additionally, efforts were made to extract data on elder law attorneys and long-term care facilities, although no data was found for our designated area of North California, encompassing 20 zip codes. Subsequently, the scraped data was compiled into a single file and merged based on zip code identifiers.

In instances where certain services were unavailable in a given zip code, a contingency plan was developed to identify the nearest available zip code. This was achieved through the utilization of **geopandas** to map the nearest zip code within a maximum distance of 50 miles, ensuring comprehensive coverage of service data despite limitations in availability for specific locations.



EDA: Exploratory Data Analysis

The plot illustrates the distribution of different facilities in each zip code. It is evident that zip codes **95551** and **95524** have no available providers. Therefore, we will gather information about providers from the nearest zip code to them.



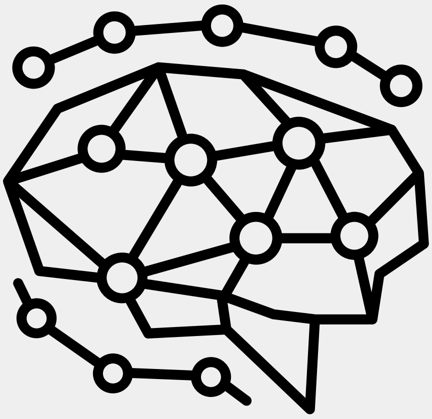
EDA: Exploratory Data Analysis

Similarly, this plot illustrates the comparison between the number of **providers offering telehealth services** and those **not offering telehealth services** for each zip code.

Model

Our predictive model is designed to make things easy. It automatically groups similar data together, identifies potential risks, and sets boundaries to categorize them. Our predictive model employs automated clustering methods, risk identification, and thresholding procedures, allowing for straightforward application to fresh datasets. This means it can work smoothly with new sets of data without needing anyone to step in and adjust things. The model sorts through scraped data (e.g., transportation, provider, etc.) using the inputted zipcode and recommends the closest services. If the zipcode is not found, the model will suggest services from the nearest zipcode within a 50-mile radius of the current one.

Here's how it works: Let's say you input a zipcode. The model then sifts through the gathered data, which includes information about transportation, healthcare providers, and more. Based on the zipcode you entered, it suggests the services closest to that location. But if it can't find a specific zipcode, don't worry. The model is smart enough to look for the nearest one within a 50-mile range and recommend services from there instead.



The India Connection - Data for a Healthier Future

Our predictive model utilizes automated clustering techniques, risk detection, and thresholding processes, making it easily applicable to new datasets. Its automation guarantees scalability, efficiency, and flexibility, eliminating the necessity for manual intervention or assumptions.

Challenges

In the context of Indian healthcare, predicting health risks using social determinants of health (SDOH) variables presents a unique difficulty because of the disparities in healthcare and the nation's diverse socioeconomic landscape. Even if the model might be based on US statistics, the following factors need to be carefully considered before applying it to India:

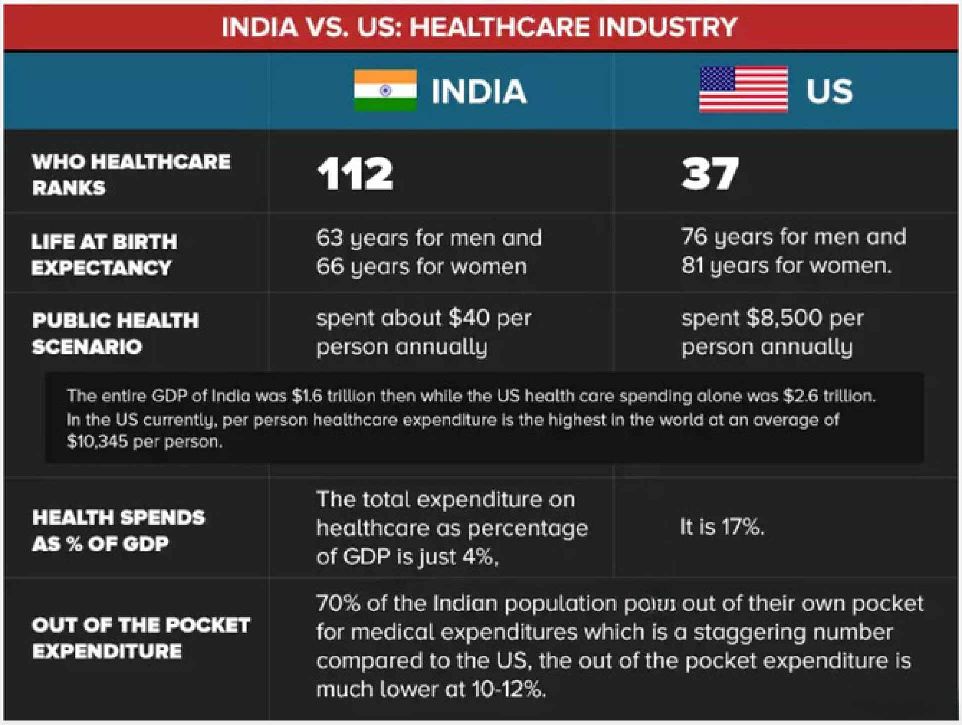
- Healthcare predictions in India face challenges because some places have better healthcare than others, and some people have more access to healthcare than others.
- There are different types of healthcare, and some are better organized than others. In rural areas, there are fewer doctors and hospitals, making it harder to predict healthcare needs.
- We don't have enough good information about healthcare in India, which makes it tough to predict what might happen in the future.
- Things like how much money people have and how far they have to travel to get to a hospital affect their health. We need to find ways to deal with these issues to make better predictions about healthcare.
- The way people live and the things they believe in also affect their health. We need to think about these things when we make predictions about healthcare.
- We need to make predictions about healthcare that work for different places and different kinds of people in India.
- People who work in healthcare, government leaders, scientists, and people in communities all need to work together to make good predictions about healthcare.
- Using technology like video calls with doctors and health apps on phones can help make healthcare better, especially in places where it's hard to get to a doctor.
- We need to make hospitals and clinics better, teach more people to work in healthcare, and find better ways to collect information about healthcare to make things better for everyone in India.

It's really important to address these complex problems so that we can make prediction models that understand the details of healthcare in India. We need to use thorough plans to make sure these prediction models can properly understand health risks and suggest the right actions for India's varied healthcare system. Factors like how much money people have and the limits of healthcare buildings directly affect how healthy people are and how easy it is for them to get medical help.

The India Connection

US vs INDIA

This image highlights the contrasting features of the American and Indian healthcare systems. Despite India's rapid development, it significantly trails behind the USA, evident in factors like healthcare access, infrastructure, and expenditure. The statistics underscore the substantial gap between the two nations' healthcare landscapes.



INDIA VS. US: HEALTHCARE INDUSTRY

	INDIA	US
WHO HEALTHCARE RANKS	112	37
LIFE AT BIRTH EXPECTANCY	63 years for men and 66 years for women	76 years for men and 81 years for women.
PUBLIC HEALTH SCENARIO	spent about $40 per person annually	spent $8,500 per person annually
The entire GDP of India was $1.6 trillion then while the US health care spending alone was $2.6 trillion. In the US currently, per person healthcare expenditure is the highest in the world at an average of $10,345 per person.		
HEALTH SPENDS AS % OF GDP	The total expenditure on healthcare as percentage of GDP is just 4%,	It is 17%.
OUT OF THE POCKET EXPENDITURE	70% of the Indian population pours out of their own pocket for medical expenditures which is a staggering number compared to the US, the out of the pocket expenditure is much lower at 10-12%.	

Approach

After conducting thorough research, we found that expanding our model to cater to India's diverse healthcare landscape is crucial. While some adjustments may be necessary to align with the country's specific nuances, our model provides a robust foundation. Its adaptable nature allows for the modification of variables and methodologies to seamlessly align with India's unique dataset, ensuring subsequent models effectively address the nation's multifaceted healthcare challenges. Moreover, the automation of most processes within our model streamlines the adaptation process, enhancing its efficiency and scalability for use in the Indian context.

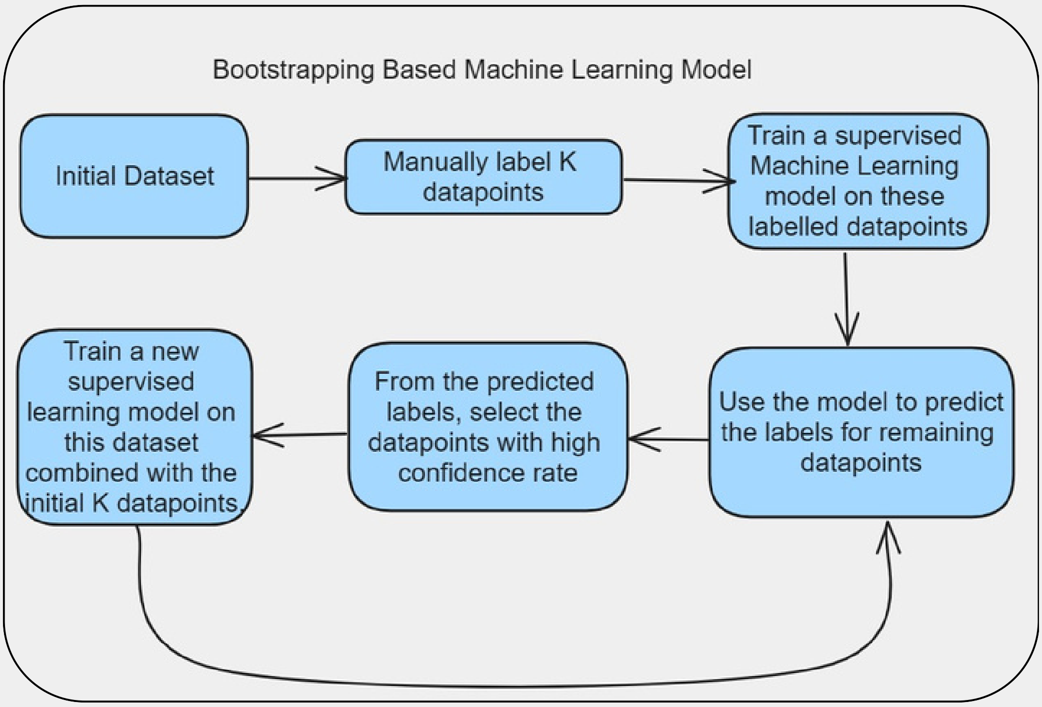
The India Primary Health Care Data available on Kaggle serves as a rich resource that can significantly contribute to our understanding of the medical conditions prevalent in India. Analyzing this dataset can provide valuable insights into the healthcare needs and challenges faced by different regions and demographic groups across the country. By leveraging this data, we can gain a deeper understanding of India's healthcare landscape, enabling us to tailor our model more effectively to address the diverse needs of the population.

Integrating the Kaggle dataset into our model holds immense potential for enhancing healthcare applications within India. By combining our model's flexibility and automation with the insights gained from the Kaggle dataset, we can develop more accurate and targeted predictive models. This integration will enable us to better assess health risks, guide interventions, and improve healthcare outcomes for the Indian population. Overall, leveraging both resources will contribute to the advancement of healthcare initiatives in India and pave the way for more effective healthcare delivery systems.

Future Aspect

In future Aspect, we will be planning to fine tune our modelling by bootstrapping in which we begin by manually labeling a small set of data points, typically around 100, to initialize the model. After training the model on this initial labeled dataset, we deploy it to predict labels for the entire dataset.
Using a confidence score threshold, we filter out predictions with lower confidence levels. The data points with confident predictions are then added to the labeled dataset.
This expanded labeled dataset is used to retrain the model, incorporating the new data points and refining its predictive capabilities. Once the model is retrained, we repeat the process by deploying it again to predict labels for all data points. This iterative cycle of prediction, labeling, and retraining continues, gradually increasing the size of the labeled dataset and improving the model's performance over time.

By iteratively refining the model with additional labeled data points, we aim to enhance its accuracy and robustness, making it more effective at predicting labels for new, unseen data.
During the fine-tuning process, various hyperparameters, such as learning rate, regularization strength, and model architecture, can be adjusted and optimized based on the performance metrics obtained from bootstrapping. Additionally, feature selection techniques and data preprocessing methods can be refined to enhance model generalization and interpretability.


Bootstrapping Based Machine Learning Model
Initial Dataset
Manually label K datapoints
Train a supervised Machine Learning model on these labelled datapoints
Use the model to predict the labels for remaining datapoints
From the predicted labels, select the datapoints with high confidence rate
Train a new supervised learning model on this dataset combined with the initial K datapoints

Overview of Bootstrapping

Conclusion

In conclusion, the Social Determinants of Health (SDOH) play a crucial role in shaping individual and community-level health outcomes, emphasizing the need for comprehensive strategies to address health disparities and achieve equity. Our initiative, the 2024 Community Companion Challenge, aims to utilize data science to understand SDOH and develop innovative solutions to improve healthcare delivery.

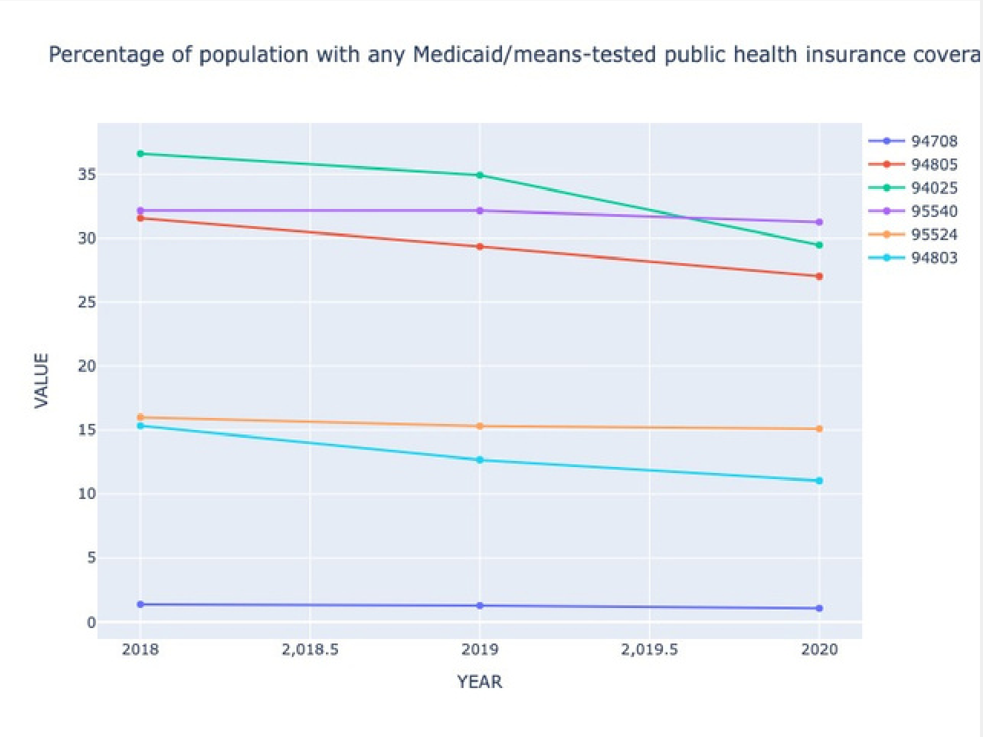
The heart of our challenge lies in the creation of the Social Care Scorecard, a data-driven tool designed to assess individual social support needs and health risks. By leveraging predictive modeling techniques, we can identify and flag specific needs such as transportation and healthy food options, enabling healthcare professionals to take targeted actions. Additionally, we aim to connect individuals with relevant support services through a comprehensive database of community resources.

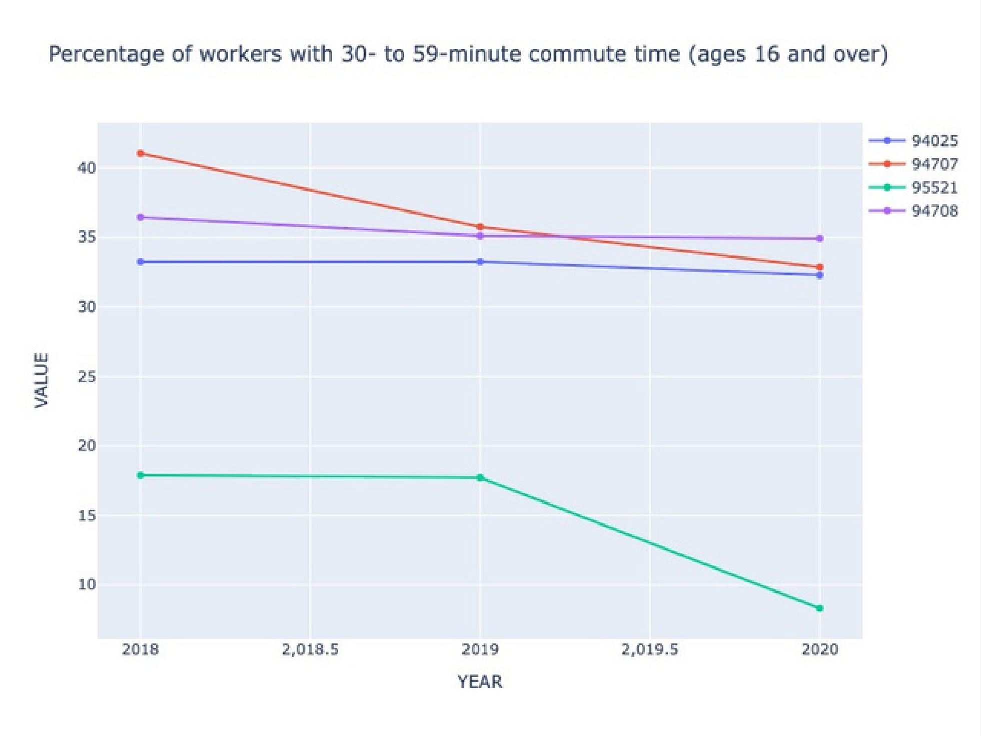
The India Connection aspect of our challenge highlights the importance of tailoring predictive models to address the unique healthcare challenges faced by India. By integrating innovative data sources, such as the India Primary Health Care Data available on Kaggle, we can enhance our understanding of India's healthcare landscape and develop more accurate predictive models. This integration holds immense potential for improving healthcare outcomes and delivering more effective interventions tailored to the diverse needs of the Indian population.

In summary, our initiative seeks to harness the power of data science to address health disparities, improve healthcare access, and ultimately make a real-world difference in people's lives. By bridging the gap between data analysis and healthcare delivery, we aim to create positive impacts on healthcare systems globally, promoting health equity and well-being for all.

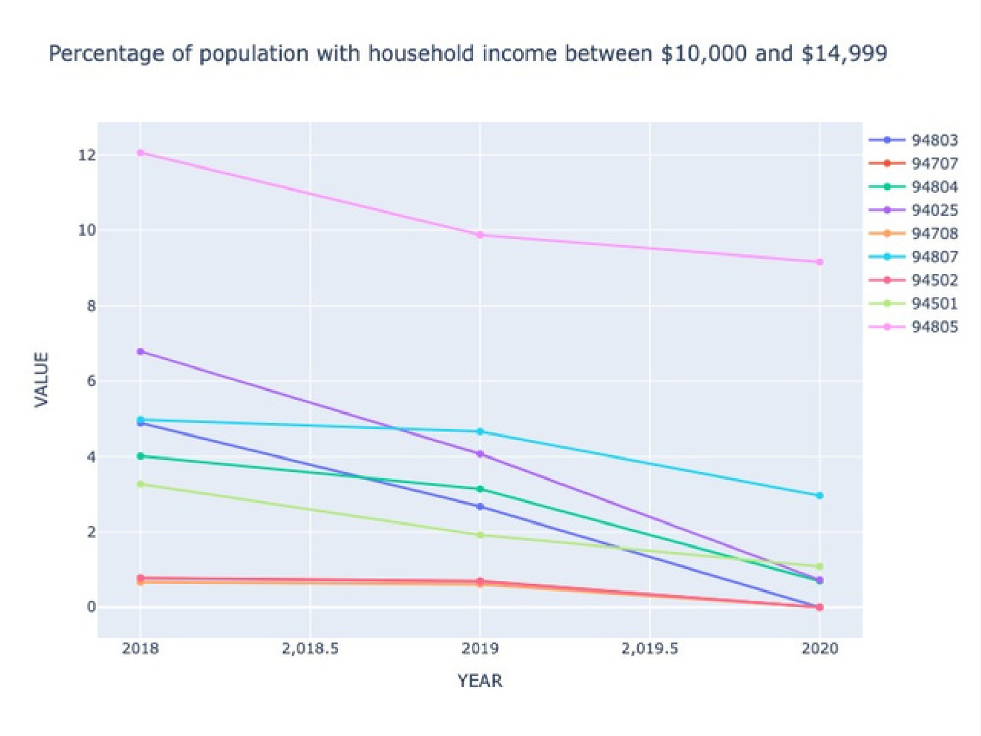
ANNEXURE

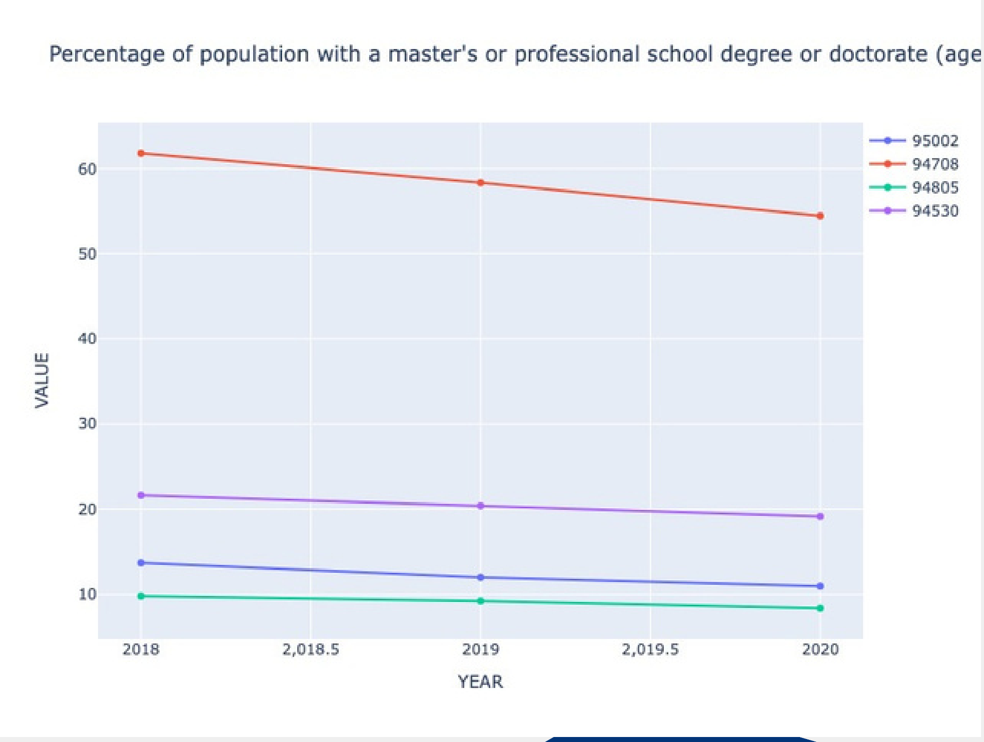
Social trends and forecasting



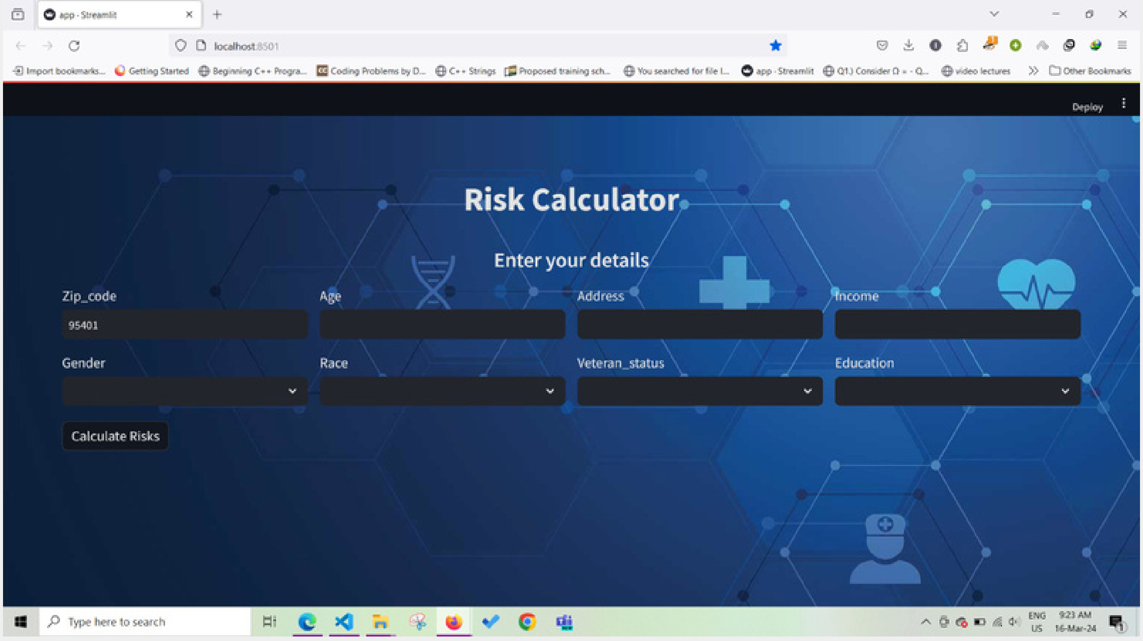


Social trends and forecasting





UI and UX



Risk Calculator

Enter your details

Zip_code: 95401

Age

Address

Income

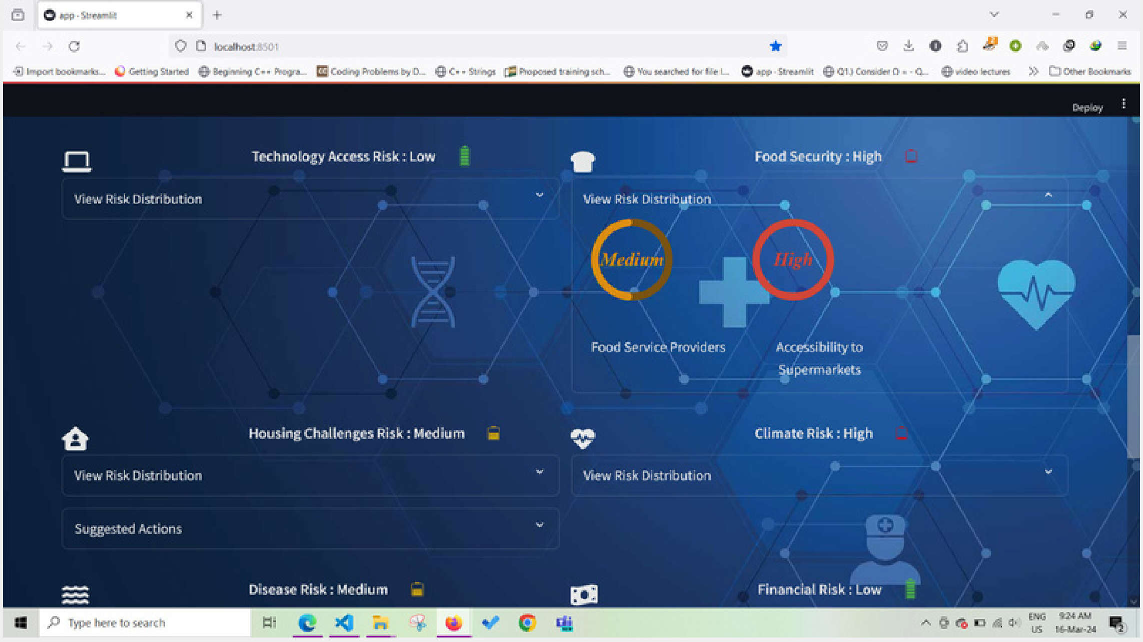
Gender

Race

Veteran_status

Education

Calculate Risks



Technology Access Risk : Low

View Risk Distribution

Food Security : High

View Risk Distribution

Medium — Food Service Providers

High — Accessibility to Supermarkets

Housing Challenges Risk : Medium

View Risk Distribution

Suggested Actions

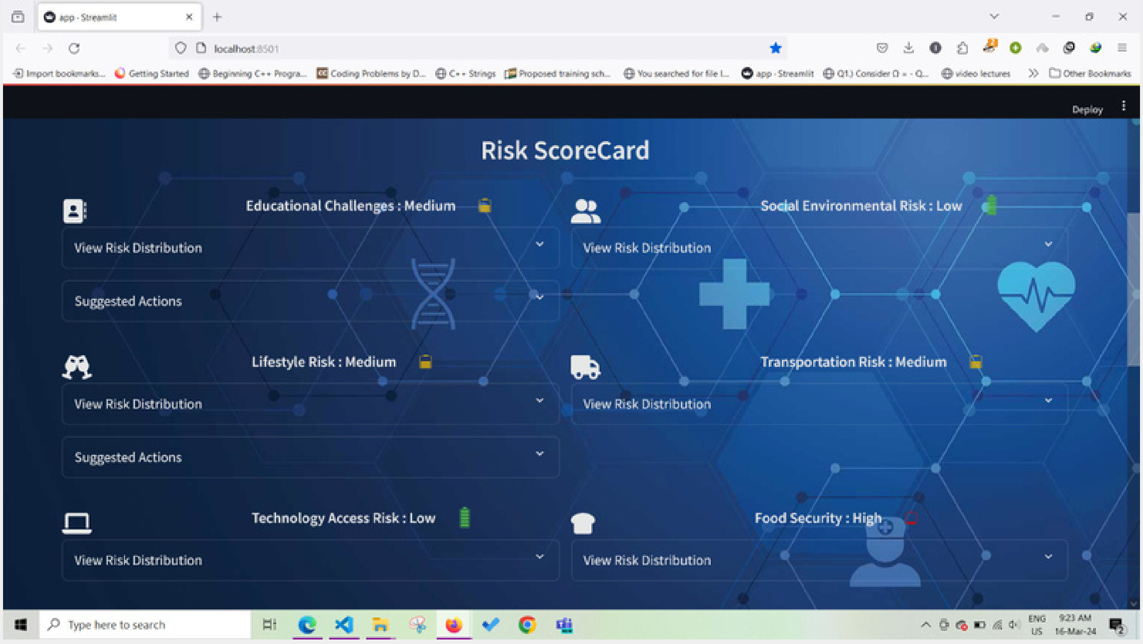
Climate Risk : High

View Risk Distribution

Disease Risk : Medium

Financial Risk : Low

UI and UX



Risk ScoreCard

Educational Challenges : Medium

View Risk Distribution

Suggested Actions

Social Environmental Risk : Low

View Risk Distribution

Lifestyle Risk : Medium

View Risk Distribution

Suggested Actions

Transportation Risk : Medium

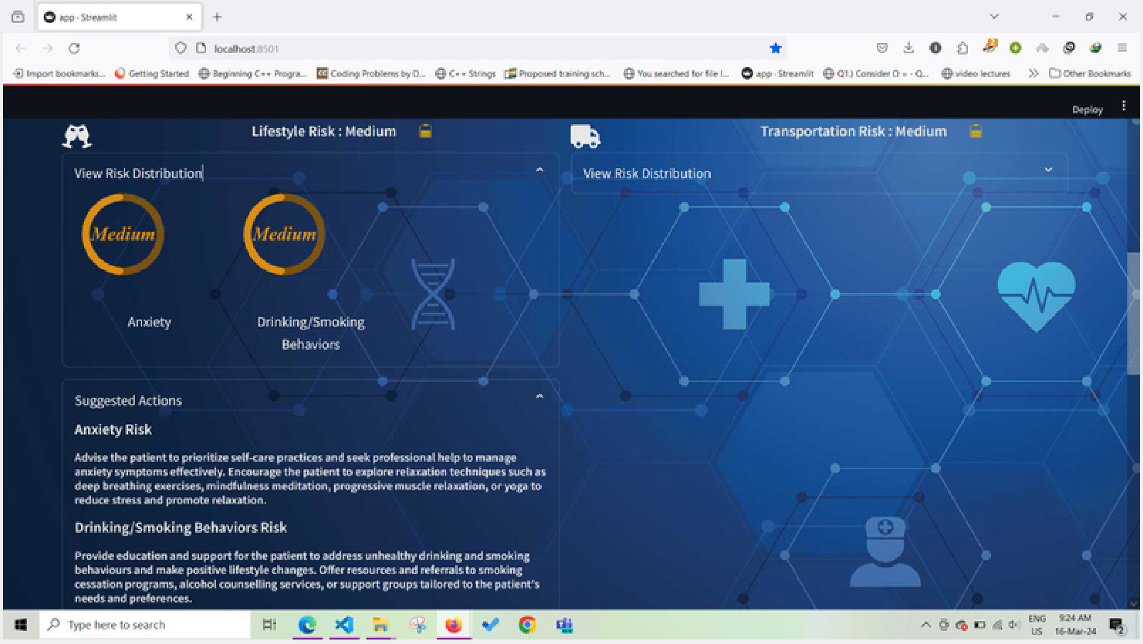
View Risk Distribution

Technology Access Risk : Low

View Risk Distribution

Food Security : High

View Risk Distribution



Lifestyle Risk : Medium

View Risk Distribution

Anxiety

Drinking/Smoking Behaviors

Transportation Risk : Medium

View Risk Distribution

Suggested Actions

Anxiety Risk

Advise the patient to prioritize self-care practices and seek professional help to manage anxiety symptoms effectively. Encourage the patient to explore relaxation techniques such as deep breathing exercises, mindfulness meditation, progressive muscle relaxation, or yoga to reduce stress and promote relaxation.

Drinking/Smoking Behaviors Risk

Provide education and support for the patient to address unhealthy drinking and smoking behaviours and make positive lifestyle changes. Offer resources and referrals to smoking cessation programs, alcohol counselling services, or support groups tailored to the patient's needs and preferences.