

Problem Statement:

In the modern world, cities and communities face a myriad of challenges that can vary significantly even within small geographic areas. Issues such as economic disparities, public health concerns, environmental hazards, and infrastructure needs often differ from one neighbourhood to another. Traditional approaches to identifying and addressing these problems have relied heavily on broad geographic analyses, such as at the city or state level, which often overlook the unique circumstances present in smaller, more localized regions like zip codes.

This lack of granularity in traditional methods results in a one-size-fits-all approach that fails to address specific local needs. For instance, a public health initiative that is effective in one zip code might be less so in another due to differing demographics, economic conditions, or environmental factors. Similarly, infrastructure improvements prioritized based on city-wide data might miss critical areas that require urgent attention.

To overcome these challenges, there is a need for a more precise and data-driven approach that can predict and address issues at a highly localized level. This is where an AI-driven system comes into play, leveraging detailed data and advanced modelling techniques to provide tailored insights specific to individual zip codes. By incorporating AI and machine learning, such a system can detect subtle patterns, anticipate potential problems, and offer proactive solutions that are tailored to the unique needs of each community.

Current Scenario:

In today's fast-paced and data-rich environment, decision-makers in urban planning, public health, and other sectors are increasingly turning to data-driven strategies to address localized challenges. The ability to predict and respond to issues at a granular level, such as specific zip codes, has become essential for effective governance and resource allocation. As cities grow and become more complex, the need for precision in understanding and addressing local issues has never been greater.

Traditionally, analyses and interventions were conducted at broader geographic levels, such as cities or states, using general statistics and historical data. These traditional methods, while useful for large-scale planning, often miss the unique challenges faced by smaller communities. For example, public health initiatives might be designed based on city-wide data, overlooking neighbourhoods with higher vulnerability due to specific local factors like pollution or economic disparity. Infrastructure projects are similarly prioritized using broader data, which may not reflect the urgent needs of particular zip codes that are more prone to flooding or traffic congestion.

These traditional approaches are largely reactive, addressing issues only after they have become apparent. Without the capability to analyze data at a more granular level, there is a risk of inefficiency and misallocation of resources, leading to suboptimal outcomes for communities. Furthermore, the lack of integration between various data sources—such as demographic, economic, and environmental data—limits the ability to make informed decisions that truly reflect the needs of each community.

However, with advancements in AI and machine learning, there is a significant opportunity to move from these traditional, reactive methods to a more proactive and precise approach. By leveraging detailed, localized data, an AI-driven system can anticipate potential issues and provide tailored recommendations specific to individual zip codes. This shift not only enhances the effectiveness of interventions but also allows for the efficient use of resources, ensuring that the right solutions are applied where they are needed most.

Dataset Description:

Objective 1: Analysis of Social Determinants of Health

Social Determinants of Health: Information about socioeconomic characteristics including education level, work status, income distribution, and access to healthcare services is included in the dataset that was obtained from the Social Determinants of Health database. This dataset sheds light on the underlying socioeconomic factors that affect a population's health results.

Health Data by Census Tract: Information on illness prevalence and localized health outcomes can be found in data gathered from health records by census tract.

Ancillary data on life expectancy: These data add context to the examination of population health trends, including life expectancy and the number and percentage of deaths in a given age group in various census tracts.

Data Type:

The datasets used for the analysis of Objective 1 include a combination of geographical, category, and numerical data that capture different aspects of population health outcomes and social determinants of health.

Objective 2: Mapping to Support

Home Health Provider Dataset: This dataset includes contact details, agency names, addresses, services rendered, accreditation status, and phone numbers for health providers. It provides information on the accessibility and standard of patient care for home health services in the region, which is crucial for developing a support network for individuals in need of home-based care.

Hospital General Information Dataset: This dataset offers a holistic view of hospital facilities, facilitating analyses ranging from geographical distributions of healthcare resources to evaluations of performance across different quality metrics such as mortality rate, readmission rate, patient experience and timeliness and effectiveness of care

Database of Long-Term Care Hospital Providers: This dataset contains information on healthcare providers, including their CMS Certification Numbers (CCN), names, addresses, contact details, and performance measures assessed by CMS. Each entry includes scores, start and end dates for evaluation periods, facilitating analyses of provider performance and quality across different regions and timeframes.

Medical Equipment providers Dataset: This dataset comprises provider information, featuring unique IDs, participation details, business and practice names, addresses, contact information, specialties, provider types, supply lists, geographic coordinates, and contract statuses for Competitive Bidding Areas (CBA).

Census Tract Crosswalk Dataset: The crosswalk dataset creates a connection between geographic locations at various levels of granularity by making it easier to map census tracts to ZIP codes. With the use of this dataset, investigations at the ZIP code level can incorporate health and socioeconomic data, allowing for more thorough spatial analysis and resource allocation plans.

Data Type:

The datasets used for the Objective 2 analysis are mostly spatial and categorical, containing details about healthcare facilities, resources, and providers that are essential for the mapping of support services.

Feature Engineering:

Null Value Treatment: In the initial step, columns containing more than 40% missing values are identified and subsequently eliminated. This criterion is 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. Then after that, exclusion criteria based on data completeness were 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 were 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 excised to preserve the integrity and reliability of future analyses.

Imputation Methods:

Method 1: K-Nearest Neighbors (KNN) Imputation

- The KNN imputation technique is applied on the null-values-removed dataset. This method assumes that the similarity between instances can be used as a basis for imputing missing values, leveraging the underlying patterns within the dataset.

Method 2: Correlation-based Imputation with DATAWIG

- Identification of Highly Correlated Variables: For each column with missing values, variables that are highly correlated ($|\text{correlation coefficient}| \geq 0.8$) with the target column identified.
- Selective Imputation Using Datawig: This method leverages the Datawig library and utilizes a deep learning-based approach for imputation. It selectively fills in missing

values within a column by utilizing highly correlated variables as features to predict and impute the missing values. This approach proves to be highly effective, especially in cases where the missing data pattern is intricate and tied to observable variables.

- Further KNN Imputation: After the Datawig imputation process, KNN imputation is applied to address any remaining missing values, ensuring a thorough treatment of missing data throughout the dataset. This step helps to further enhance the completeness and accuracy of the dataset, particularly in cases where the Datawig imputation may not have completely filled in all missing values.

Data Normalisation:

The Data Normalisation Techniques used were :-

- A. Min Max Scalar
- B. Z Score Normalisation
- C. Robust Scaling
- D. Log Transformation
- E. Quantile Transform

Below are some important features due to which this method should be used for this kind of dataset :-

- **Robustness:** Quantile transform is robust to outliers as it ranks and transforms data based on ranks, rather than raw values, which can benefit clustering with outlier-prone data.
- **Distribution Preservation:** Quantile transform ensures data follows a specific distribution (here Normal is used), advantageous for clustering algorithms assuming a particular data distribution.
- **Non-linear Transformations:** Quantile transform allows non-linear transformations, capturing complex relationships beneficial for clustering requiring non-linear transformations.
- **Normalization of Non-Normal Data:** Quantile transform normalizes non-normally distributed data, making it suitable for clustering algorithms assuming normality.
- **Distribution Equalization:** Quantile transform equalizes distributions of features, beneficial for clustering algorithms sensitive to differences in feature scales or distributions.
- **Skewness Insensitivity:** Quantile transform handles skewed data distributions better, potentially improving clustering performance on skewed datasets.

As the Data given was a microarray dataset as it is based on the census so the best performed Normalization Technique was Quantile Transform as the metrics scores such as Standard Deviation, Skewness and Kurtosis are greater than the threshold value.

Various data normalization techniques were employed to process the microarray dataset, which is inherently complex due to its derivation from census data. The techniques utilized include Min-Max Scaling, Z-Score Normalization, Robust Scaling, Log Transformation, and notably, the Quantile Transform.

The decision to apply Quantile Transform was strategic, grounded in the dataset's lack of conformity to any standard distribution, its substantial skewness, and the presence of outliers. The metrics for assessing data distribution, including Standard Deviation, Skewness, and Kurtosis, surpassed the predefined thresholds, underscoring the necessity of this normalization method. The Quantile Transform proved most appropriate for this dataset, optimizing it for further analysis by reducing the influence of outliers and transforming the data into a more uniform distribution.

Final selected features and dataset

Random Forests:

Random Forest was used to select important features after removing correlated features and features with more than 40% null values. Random Forests are a robust ensemble learning technique widely used for classification and regression. Their construction involves the aggregation of multiple decision trees trained on bootstrapped samples of the data and random subsets of features. This inherent randomization reduces overfitting, making Random Forests well-suited for high-dimensional datasets. Additionally, they provide intrinsic measures of feature importance, aiding in model interpretation.

SelectFromModel:

SelectFromModel is a versatile feature selection meta-transformer in the scikit-learn library. It operates in tandem with a fitted estimator. Feature importance scores, derived from the estimator (e.g., a Random Forest's 'feature_importance_' attribute), are analyzed. Features exceeding a predefined importance threshold are retained. This technique provides model-specific feature selection, customizable thresholds, and the potential to improve model performance and efficiency through dimensionality reduction.

Synergistic Use in Feature Selection

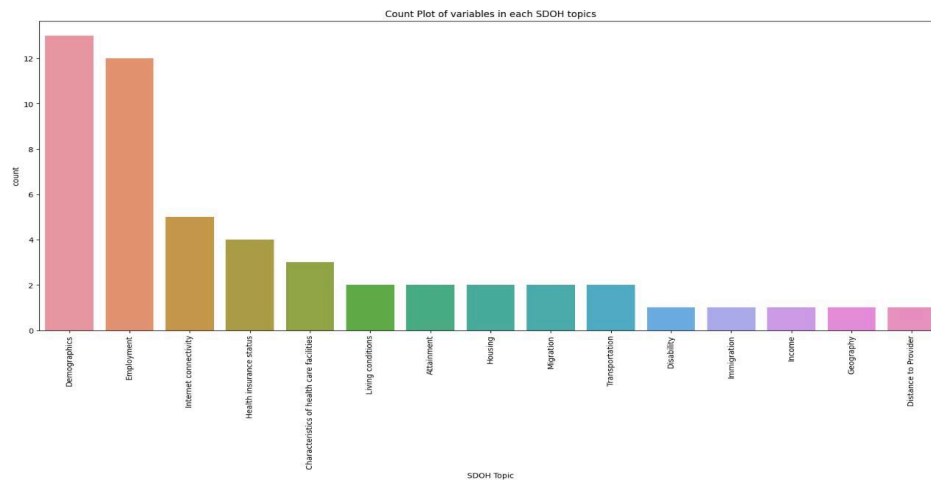
The following outlines the combined use of Random Forests and SelectFromModel for feature selection:

1. A Random Forest model is fitted to the dataset.
2. Feature importance scores are extracted from the fitted model.
3. SelectFromModel is employed, utilizing the importance scores and a selected threshold to refine the feature set.

Advantages of This Approach

- Enhanced Predictive Accuracy: By isolating the most influential features, predictive performance may be improved.
- Computational Efficiency Gains: Reduction in feature dimensionality can accelerate model training and prediction.
- Improved Model Interpretability: The selected feature subset provides valuable insights into the key drivers within the problem domain.

Synergic approach was finally deployed to reduce the features to 60 SDOH Variables.



In our chosen feature set, we included a diverse array of 14 broad areas, including demography, employment, connectivity, and more. Our aim was to incorporate variables from nearly every relevant category. Our analysis revealed 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.

Clustering

→ Domain Dataset formation

Clustering was conducted to divide our dataset into supersets of Problems faced in zip codes. There were two primary dataset constructs over which our clustering algorithms are applied. For the first part, a dataset containing 60 parameters and around 84000 rows is taken, and clustering is performed. The parameters are grouped into seven distinct domains or problem supersets for the second part. The variables were divided among these domains manually, and then after this, using a threshold for correlation among the variables of each domain, 52 columns were finally found.-

1. Socioeconomic Factors - 13 Columns
2. Internet and Communication Access - 3 Columns
3. Demographic Information - 13 Columns
4. Healthcare Access and Utilization - 7 Columns
5. Health Behavior and Outcomes - 5 Columns
6. Environmental Factors - 2 Columns
7. Housing and Transportation - 9 Columns

After the above categorization, the following clustering algorithms were tested on the datasets -

- A. K-means Algorithm
- B. Gaussian Mixture Model (GMM)
- C. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- D. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

The clustering was done on the dataset containing 60 columns to find the clustering algorithm that fits best into these dataset, and their scores are:-

Algorithm	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score	No. of Clusters
K-means	0.051997	4.011797	4775.546769	3
GMM	0.059129	4.62119	1517.420701	3
HDBSCAN	0.1184326520	4.76484614	619.992267	3
DBSCAN	0.069166	0.895483	100.585796	3

K-means was chosen as the best algorithm because although HDBSCAN, DBSCAN, and GMM had higher scores, they produced clusters with very non-uniform sizes, with most rows in one cluster only.

→ Clustering on problem supersets

So, Now as K-means is the best algorithm that fits this dataset so clustering was done on the 7 distinct domains dataset by using K-means clustering algorithm and the scores obtained were :-

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

As the scores obtained by K-means on these dataset were greater than the threshold so these clusters were taken for further processes.

→ Problem justification and cluster linking

Problem Identification: Our analysis revealed socioeconomic challenges like healthcare, education, employment, and housing disparities.

Solution Proposal: Tailored solutions were crafted for each cluster to address these challenges and improve community well-being.

Application of test census tracts

Probability Calculation: By applying our established model, we calculated the probability of each test census tract aligning with the topics of the identified clusters.

Cluster Assignment: Based on the calculated probabilities, we assigned the test census tracts to the top three clusters deemed most fitting by the analysis. It would facilitate targeted intervention strategies.

Assigning problems and solutions: We allocated the identified issues and corresponding solutions to each test census tract, ensuring a comprehensive approach to addressing the socioeconomic challenges.

→ Cluster problem Validation

Cluster problems were compared against socioeconomic indicators and SYNTHEA data for relevance. Generative AI techniques refined data points and validated cluster problems. LLM models enhanced interpretability and coherence, validating results. The use of the Synthea dataset affirmed the accuracy and reliability of our methods in identifying clusters and associated problems related to SDOH, providing a foundation for further analysis and intervention planning.

1. Experimentation

Far-mean approach

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

After compiling the final dataset, comprising 60 identified important parameters, the data was categorized into eight distinct domains or 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

To streamline the analysis, the selected parameters were condensed into one of the seven supersets mentioned above. This condensation was achieved through a blend of statistical analysis and manual selection.

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. Based on these characteristics, such as whether a variable's value for a ZIP code exceeds its average or not, the values were given with prompting to assign the fine-tuned LLAMA-2-7B chat. This framework generates tailored problems and solutions aligned with the research objectives.

By adhering to this methodology, the goal is to effectively evaluate and address various socio-economic and environmental challenges across diverse regions.

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 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 model for output.

Probabilistic Score Calculation

The model's methodology integrates unsupervised clustering algorithms such as K-Means. After completing the feature selection, the data was partitioned into training and testing sets using a split ratio of 0.15. The columns were then categorized into distinct groups, each representing various Social Determinants of Health (SDOH) variables, including Housing and Transportation, Socioeconomic factors, Internet Access, Healthcare Access and Utilization, Health Behavior and Outcomes, and Demographic Information.

Subsequently, each category underwent individual clustering using the K-Means algorithm. Employing EDA techniques, we identified unique problems associated with each cluster, identifying 23 subclusters in total, each representing a distinct issue. Following this, we computed the probability score of the test data pertaining to each subcluster and selected the top 3 probabilities, representing the top 3 problems.

Once the problems were discerned, a dataset containing potential solutions was combined by manual verification and with the help of generative AI. Finally, we trained an LLM using this dataset for future prediction and analysis.

Single LLM train on 10k dataset

In this approach, firstly, the dataset of about 80,000 rows is divided into datasets of approximately 70,000 and 10,000 rows, respectively. Then, two-level clustering was applied

to the larger dataset, resulting in the formation of 21 clusters, each of which represented about 3 problems. For the smaller dataset, the top 3 confidence scores for each data point are obtained for the previously obtained clusters and problems and solutions are assigned to that data point accordingly. Thus, a new dataset of around 10000 data points is formed along with their set of 3 problems and solutions, and the Llama model is finetuned on this dataset.

Multi-model approach for different sets of parameters

The objective of the approach is 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 to us.

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.

The project involves 330 parameters categorized into age, gender, race, veteran status, education level, and income. The objective is to train various models on these parameter sets 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 to us.

RAG-based approach

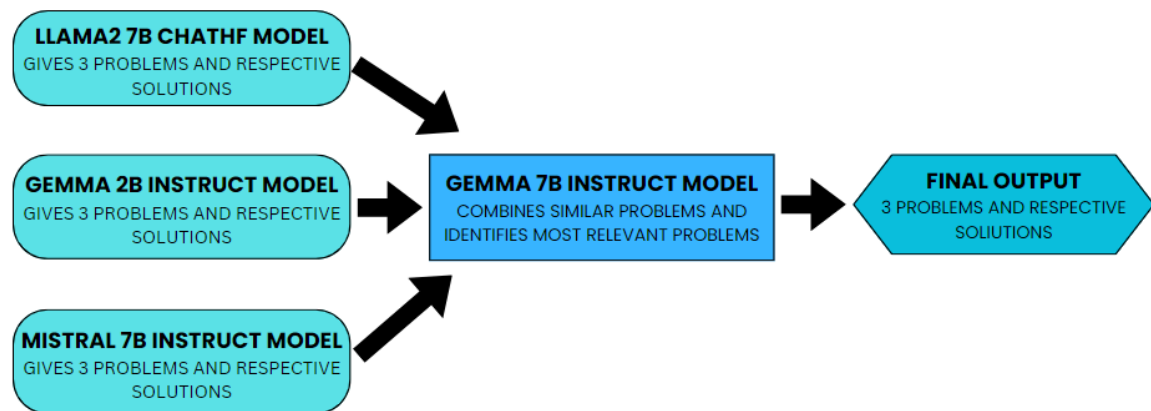
A dataset containing detailed information on important parameters, such as the effects of age and gender on these parameters, was utilized. When a ZIP code input is received, parameter values are retrieved from the ZIP code dataset, and relevant information about parameters is retrieved from the created database. The entire context is then passed to LLAMA2 for the problems and solutions of that respective ZIP code.

Final approach for social care scorecard

Ensemble methodology, end-to-end pipeline

An ensemble methodology is 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.



Validation: Synthea Testing + Our Testing (Some Manual)

An ensemble was created utilizing three models, each fine tuned on a dataset of 10,000 instances. This dataset was constructed through random sampling, selecting 15% of the total rows. Problems were assigned to each row, guided by its top 3 cluster probabilities.

The models in question are Llama2 7b, Gemma2 b, and Mistral 7b. Outputs from these models were obtained, aiming to assess the precision and similarity of their outputs. To facilitate this evaluation, the outputs were transformed into word2vec embeddings, followed by the computation of cosine similarities among them.

The similarity scores obtained were as follows: 0.82 between Gemma2 b and Mistral 7b, 0.76 between Gemma2 b and Llama2 7b, and 0.69 between Llama2 7b and Mistral 7b. These results highlight the relevance of the ensemble approach, demonstrating its effectiveness in generating outputs that are both non-repetitive and cohesive.

Connecting people with resources

The ensembled model's three problems are processed by a BERT model fine-tuned on a synthetic dataset with 1000 rows. The dataset has two columns: one with problem sets and the other with corresponding categories like transportation, age, income, and disability. These categories are mapped to CSV files.

A database of CSV files is created for various data types like geriatric care, elder law attorneys, and home care for each zip code through web scraping. The CSV files are analyzed to find the nearest hospitals and home care facilities based on the user's address, with their contacts and addresses displayed in an action-based scorecard

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 are mapped to different categories of csv. A database of csv files are created for multiple data types(like geriatric care, elder law attorneys, home care etc...) for each zipcodes by using web scraping. These csvs are analyzed and the nearest favourable hospitals,home care,..(from the user's address) are found from these csvs and their contacts,addresses are displayed in the action based scorecard.

Proof of the project:

Scorecard :-

By giving the inputs as follows :-

Zip Code :- 10003

Address :- 834 Treutel Plaza

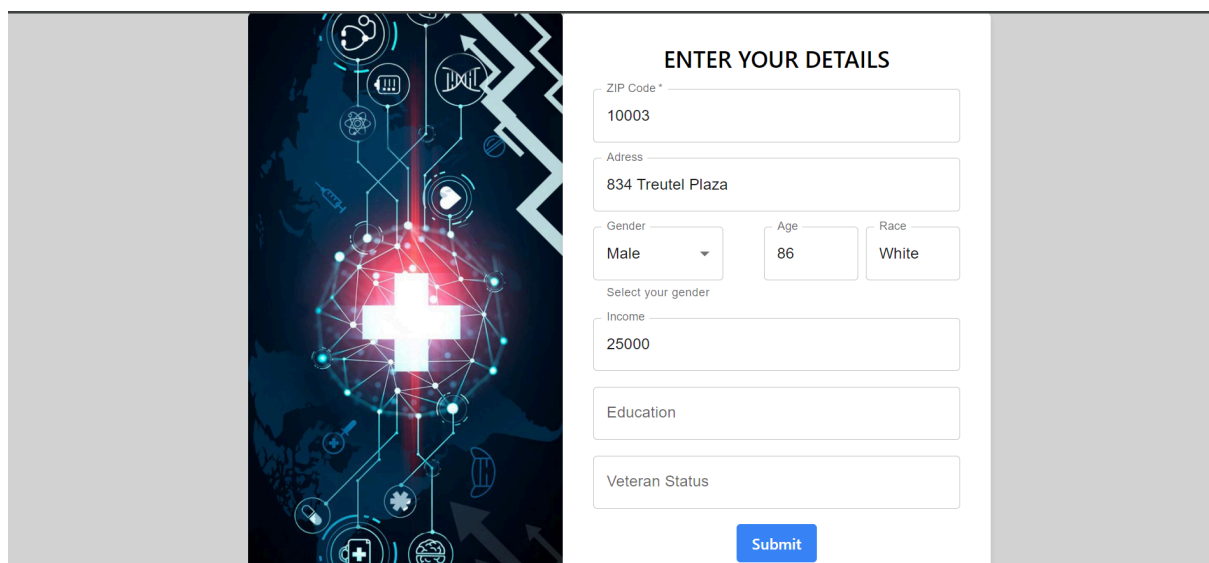
Gender :- Male

Age :- 86

Race :- white

Income :- 25000

The inputs form looks like this:-



ENTER YOUR DETAILS

ZIP Code *
10003

Address
834 Treutel Plaza

Gender
Male

Age
86

Race
White

Select your gender

Income
25000

Education

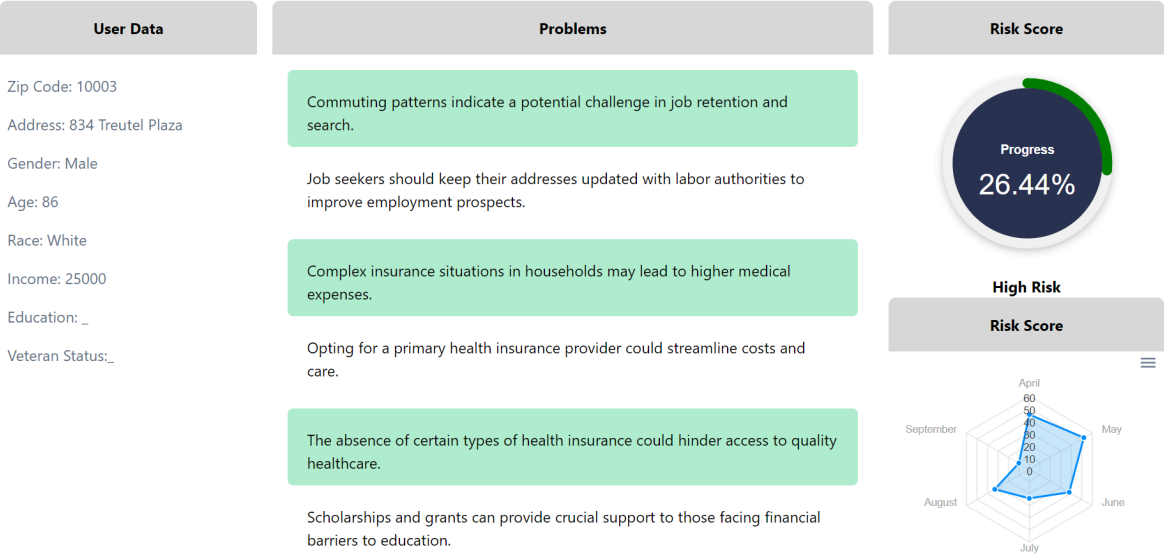
Veteran Status

Submit

After this the scorecard for task 1 will show the output as :-

- 1) **Problem:** Commuting patterns indicate a potential challenge in job retention and search.
Solution: Job seekers should keep their addresses updated with labor authorities to improve employment prospects.
- 2) **Problem:** Complex insurance situations in households may lead to higher medical expenses.
Solution: Opting for a primary health insurance provider streamlines costs and care.

- 3) *Problem:* The absence of certain types of health insurance could hinder access to quality healthcare.
- Solution:* Scholarships and grants can provide crucial support to those facing financial barriers to education.



Conclusion

1. Data Analysis and Feature Engineering: With a robust dataset at hand, data analysis and feature engineering techniques were employed. This phase included data preprocessing, normalization, and outlier detection.
2. Clustering and Insights Generation: Through the use of clustering algorithms, the project categorized the data into significant groups, revealing insights into the SDOH and their relationship with potential health issues.
3. Integration of LLMs: Following the clustering, LLMs were integrated to process the 10k train dataset, identifying potential problems and solutions. Using that, An ensemble methodology was adopted to boost predictive accuracy, integrating multiple LLMs and making the problem-solving process end-to-end.
4. Development of the Social Care Scorecard: Leveraging the insights obtained, a comprehensive social care scorecard was developed.
5. Connecting People with Resources: This practical application of data insights translates into real-world health improvements for communities, demonstrating the tangible benefits of the project's research.
6. Tailoring Solutions to the Indian Context: Recognizing the distinct challenges posed by India's healthcare landscape, the project tailored its methodologies and solutions to meet these needs.
7. Future Directions and Innovation: The report concludes with an outlook on future research directions and the potential for ongoing innovation in healthcare analytics.

Future Prospects

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.