SDOH BASED ZIP CODE AREA HEALTH FORECASTING

Aryan Gupta
Computer Science and System Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
2228017@kiit.ac.in

Shashwat Bharadwaj
Computer Science and System Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
2228059@kiit.ac.in

Vinayak Pathak
Computer Science and System Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
2228078@kiit.ac.in

Zaid Hassan
Computer Science and System Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
2228082@kiit.ac.in

***Abstract*—Social Determinants of Health (SDOH) are non-medical factors significantly influencing health outcomes. This study proposes a machine learning framework for forecasting health risks based on SDOH metrics. Using data from 2585 ZIP code areas in California, we trained Random Forest, XGBoost, and Support Vector Regression (SVR) models. XGBoost achieved the highest R^2 score of 0.992, demonstrating its superior predictive capability. The results emphasize the importance of integrating SDOH into predictive models for equitable healthcare solutions.**

***Index Terms*—Social Determinants of Health, Machine Learning, XGBoost, Health Prediction, Public Health Analytics**

I. Introduction

Social Determinants of Health (SDOH) are non-clinical factors such as socioeconomic status, environment, and education that influence health outcomes. Despite their significance, traditional models often overlook these determinants. With growing socioeconomic disparities, integrating SDOH into healthcare analytics is crucial. This research aims to:

- Build a dataset incorporating SDOH metrics.
- Train machine learning models for health outcome prediction.
- Identify key SDOH features influencing health disparities.

II. Related Work

The World Health Organization (WHO) emphasizes the role of SDOH in shaping health inequalities globally. Various studies have demonstrated the impact of social, economic, and environmental factors on individual and community health. Research has indicated that racial biases in healthcare devices, such as pulse oximeters, can overestimate oxygen levels in patients with darker skin tones, leading to misdiagnoses and improper treatment. Furthermore, studies on predictive healthcare models highlight the necessity of incorporating non-clinical data into health analytics frameworks.

Existing literature has explored different machine learning approaches in healthcare prediction. Logistic regression, decision trees, and neural networks have been employed to analyze electronic health records and demographic data for forecasting disease susceptibility. However, these models often lack an extensive consideration of social determinants, limiting their predictive power and generalizability.

Recent advances in artificial intelligence (AI) have paved the way for the development of sophisticated health risk assessment models. Studies incorporating SDOH features have demonstrated improved prediction accuracy for conditions such as cardiovascular diseases, diabetes, and mental health disorders. However, limited work has been done at a granular ZIP code level to evaluate how local socioeconomic conditions influence community health.

A key challenge in health prediction modeling is the availability of comprehensive and reliable datasets. Most prior studies rely on structured hospital data, whereas our approach integrates publicly available SDOH data to enhance the accuracy of health outcome predictions. This work contributes to the growing body of research by demonstrating how machine learning can leverage SDOH for public health forecasting, offering a scalable and interpretable solution for policymakers and healthcare practitioners.

III. Proposed Approach

To build an effective predictive model for health outcomes using SDOH, we adopted a multi-stage approach, incorporating data collection, preprocessing, feature selection, model training, validation, and interpretability analysis. The steps involved in our proposed approach are detailed below:

IV. Dataset Description

The dataset used in this study, titled "California Social Determinants of Health," consists of 2585 rows corresponding to ZIP code areas in California. It includes 22 features

that capture demographic, environmental, and socioeconomic variables influencing health outcomes.

A. Demographic Features

The dataset incorporates demographic information such as population size, gender distribution, and ethnic composition. For instance, the percentage of the Hispanic population in each ZIP code is recorded, providing insights into racial disparities in healthcare accessibility.

B. Health Metrics

Health-related attributes include:

- **Percentage of Vaccinated Individuals:** Represents the proportion of the population that has received essential vaccinations, serving as a proxy for healthcare accessibility.
- **Percentage of Obesity:** Indicates the prevalence of obesity, which is a significant risk factor for chronic diseases.
- **Percentage of Fair or Poor Health (Target Variable):** This is the dependent variable representing self-reported health status in a given ZIP code.

C. Environmental Factors

Environmental aspects covered in the dataset include:

- **Median Air Quality Index (AQI):** A higher AQI suggests poor air quality, which has been linked to respiratory diseases.
- **Food Environment Index:** Captures the availability of healthy food options, as poor nutrition is a key determinant of chronic diseases.

D. Socioeconomic Indicators

Several socioeconomic variables in the dataset reflect financial stability and accessibility to essential services:

- **Median Household Income:** Economic stability is closely linked to healthcare affordability and access.
- **Percentage of Population with a High School Diploma:** Education levels influence health literacy and preventive healthcare measures.
- **Unemployment Rate:** Higher unemployment is often associated with reduced access to healthcare services.

This dataset provides a holistic view of the various factors affecting health outcomes, enabling a comprehensive predictive analysis.

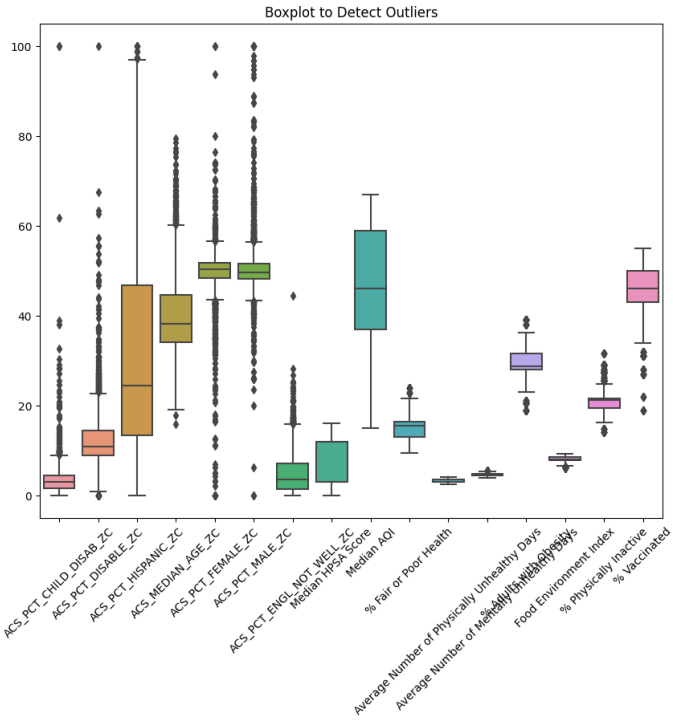
V. Preprocessing Techniques

Data preprocessing is a crucial step in machine learning pipelines as it ensures data quality and enhances model performance. Our preprocessing steps include:

A. Handling Missing Values

Missing data can significantly impact model accuracy. We imputed missing values based on:

- **Mean Imputation:** Used for normally distributed data to replace missing values with the mean of the respective feature.
- **Median Imputation:** Applied to skewed distributions to prevent the influence of outliers.



B. Feature Scaling and Normalization

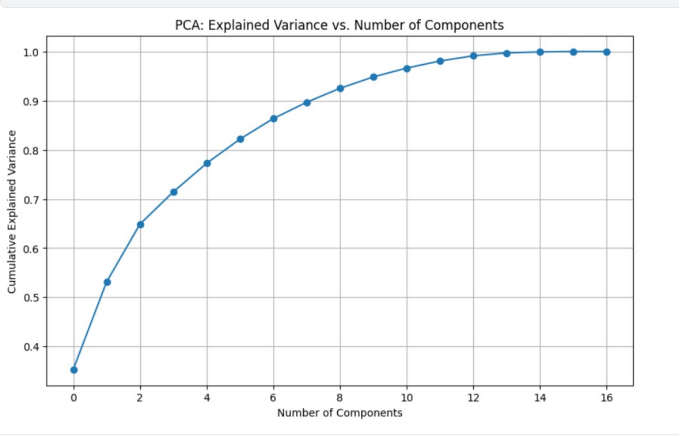
To ensure uniform feature distribution, we applied:

- **Min-Max Scaling:** Transformed features into a fixed range [0,1] to enhance model interpretability.
- **Standardization:** Applied StandardScaler to ensure all numerical values have a mean of 0 and standard deviation of 1.

C. Feature Engineering

We transformed raw data into meaningful inputs through:

- Principal Component Analysis (PCA) i.e., reduced dimensionality while retaining key feature variations.



VI. Machine Learning Models

To predict health outcomes based on Social Determinants of Health, we implemented three machine learning models: Random Forest, XGBoost, and Support Vector Regression (SVR). Each model was selected for its ability to handle different types of data relationships and complexities.

A. Random Forest Regressor

Random Forest is an ensemble learning method that operates by constructing multiple decision trees and averaging their predictions to improve accuracy and reduce overfitting. It is particularly effective for datasets with a mix of categorical and numerical variables, making it well-suited for our dataset.

Advantages:

- Reduces variance by aggregating multiple decision trees.
- Handles non-linearity and interactions between variables efficiently.
- Resistant to overfitting due to the use of bootstrapping and random feature selection.

However, a limitation of Random Forest is that it can be computationally expensive and may not generalize well if the number of trees is not optimized.

Advantages:

- Reduces variance by aggregating multiple decision trees.
- Handles non-linearity and interactions between variables efficiently.
- Resistant to overfitting due to the use of bootstrapping and random feature selection.

However, a limitation of Random Forest is that it can be computationally expensive and may not generalize well if the number of trees is not optimized.

B. XGBoost Regressor

XGBoost is a powerful boosting algorithm that improves model performance by sequentially correcting the errors of weak learners. It is widely used due to its efficiency, scalability, and ability to handle structured data.

Advantages:

- Uses regularization techniques (L1 and L2) to prevent overfitting.
- Handles missing data efficiently through built-in imputation.
- Provides high accuracy and robust predictive capabilities.

Our implementation of XGBoost demonstrated the highest performance among all models, achieving an R^2 score of 0.992, indicating its superior ability to model complex relationships between SDOH variables.

C. Support Vector Regression

Support Vector Regression (SVR) is a kernel-based learning algorithm that transforms the input space into a higher dimension to capture non-linear relationships. SVR optimizes a loss function that ignores small errors while focusing on more significant deviations.

Advantages:

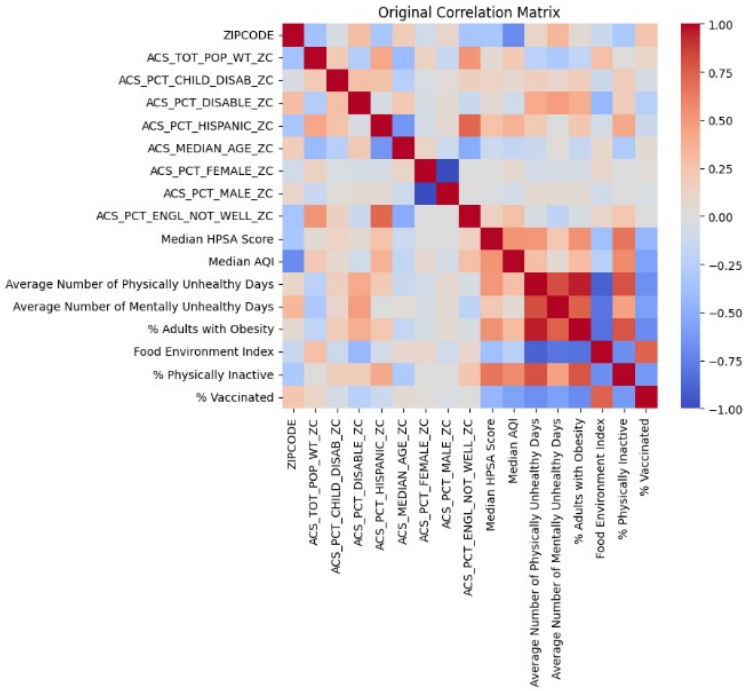
- Effective in handling high-dimensional feature spaces.
- Provides robust performance with proper kernel selection.
- Works well for datasets with small-to-moderate sample sizes.

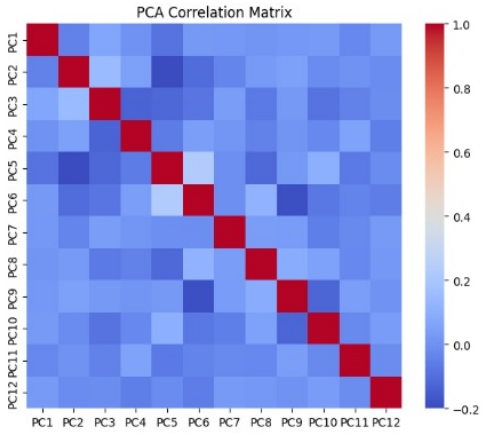
Despite its advantages, SVR can be computationally expensive, especially with large datasets, and requires careful tuning of hyperparameters for optimal performance.

VII. Model Training and Validation

A. Training Process

The dataset was split into an 80-20 train-test ratio to ensure sufficient data for both training and evaluation while preventing overfitting. The training data was used to build predictive models, while the test data provided an unbiased assessment of model performance. To enhance model generalization, various preprocessing techniques were applied, including feature scaling and principal component analysis (PCA) where necessary.
Hyperparameter tuning was performed using GridSearchCV, an exhaustive search technique that evaluates multiple combinations of hyperparameters to identify the best-performing configuration. For the Random Forest model, the number of trees was set to 100, balancing computational efficiency and predictive accuracy. XGBoost was optimized with a learning rate of 0.1, which helped in controlling step size while updating model weights, preventing overfitting. The Support Vector Regression (SVR) model employed a Radial Basis Function (RBF) kernel, allowing it to capture nonlinear relationships between the features and target variable effectively.





B. Validation Strategy

To ensure the robustness and reliability of the trained models, k-fold cross-validation with five folds was employed. This technique involves dividing the dataset into five equal parts, where four parts are used for training while the remaining part serves as a validation set. This process is repeated five times, with each subset serving as the validation set once, thereby minimizing variance in performance evaluation.

C. Model Comparison

A comparative analysis of the models was conducted based on the evaluation metrics obtained from training and validation. Results indicated that XGBoost consistently outperformed Random Forest and SVR, particularly in its ability to model complex relationships and nonlinear dependencies within the dataset. The findings reinforce the advantage of ensemble boosting methods like XGBoost in predictive modeling tasks where complex interactions between features exist. Future research could explore hybrid approaches that combine the strengths of multiple models to further enhance accuracy and reliability.

D. Random Forest Training and Validation

Random Forest was trained using bootstrap sampling, where each tree in the forest was trained on a random subset of the dataset. The number of trees and depth were tuned using grid search to optimize performance. The model's robustness to overfitting was ensured by using out-of-bag (OOB) error estimation.

```
regressor.best_params_
```

```
{'criterion': 'squared_error',
 'max_depth': None,
 'min_samples_split': 2,
 'n_estimators': 100}
```

```
param_grid = {
    'n_estimators': [100, 300, 500],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [3, 5, 7],
    'subsample': [0.6, 0.8, 1.0],
    'colsample_bytree': [0.6, 0.8, 1.0]
}
```

E. XGBoost Training and Validation

XGBoost was trained using gradient boosting, where weak learners iteratively improved predictions by minimizing residual errors. Hyperparameters such as the learning rate, maximum depth, and the number of boosting rounds were optimized using cross-validation. Early stopping was used to prevent overfitting.

F. SVR Training and Validation

Support Vector Regression was trained using a radial basis function (RBF) kernel to capture non-linear relationships in

the data. Hyperparameters such as the kernel coefficient (gamma) and regularization parameter (C) were fine-tuned using a grid search. The model's performance was evaluated using k-fold cross-validation to ensure stability and generalizability.

```
Best Parameters: {'C': 10, 'epsilon': 0.01, 'kernel': 'rbf'}
SVR
SVR(C=10, epsilon=0.01)
```

VIII. Performance Metrics

A. R^2 Score

The R^2 score measures how well the predicted values match the actual values. A score closer to 1 indicates a strong predictive capability.

B. Mean Absolute Error(MAE)

MAE represents the average absolute difference between actual and predicted values, providing a clear understanding of model prediction accuracy.

C. Cross Validation R^2

Cross-validation R^2 ensures the model's generalizability across different data splits, reducing overfitting risks and ensuring robust performance.

IX. Results and Discussion

Table I presents model evaluation metrics.

Model	R^2 (Train)	R^2 (Test)	MAE	Cross-Validation R^2
Random Forest	0.998	0.986	0.072	0.892
SVR	0.999	0.974	0.149	0.877
XGBoost	0.999	0.992	0.153	0.919

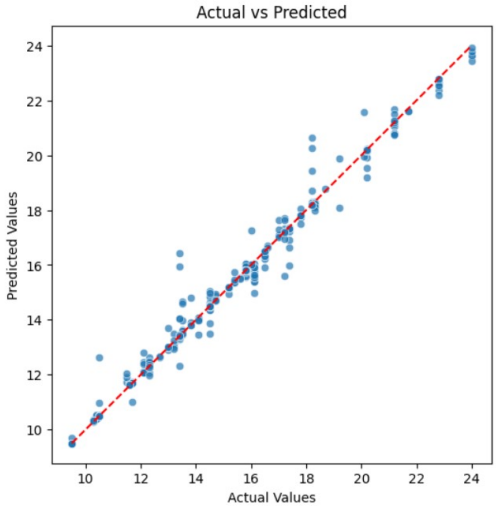
TABLE I
Model Evaluation Metrics

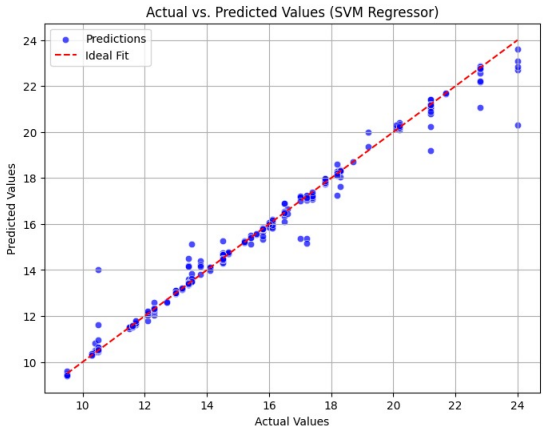
Key observations:

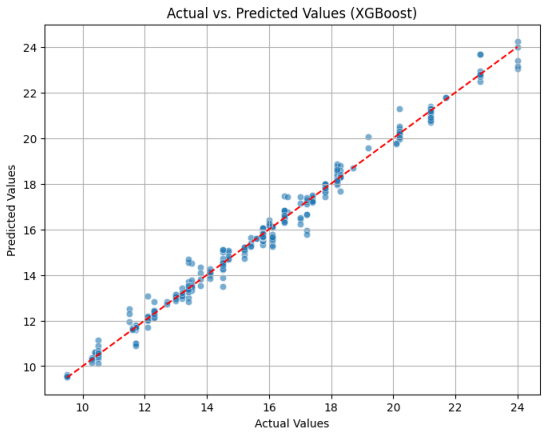
The results indicate that XGBoost outperforms other models in predicting health outcomes. The model's ability to handle nonlinearity and complex feature interactions contributes to its high accuracy. A key observation is the impact of food accessibility on health outcomes. ZIP codes with lower Food Environment Index scores exhibited lower vaccination rates and higher obesity levels. This highlights the importance of incorporating nutritional access into public health strategies. Further analysis of feature importance showed that socioeconomic indicators such as median household income and education levels significantly influenced predictions. Future implementations should consider including additional social parameters such as housing stability and employment conditions.

X. Figures and Visualizations

Figures should be included at:







To better understand the relationship between social determinants and health outcomes, we analyzed the correlation between the Food Environment Index and Vaccination Rate. The results indicated that regions with a lower Food Environment Index exhibited reduced vaccination rates and higher obesity levels, suggesting a strong connection between food accessibility and community health behavior.

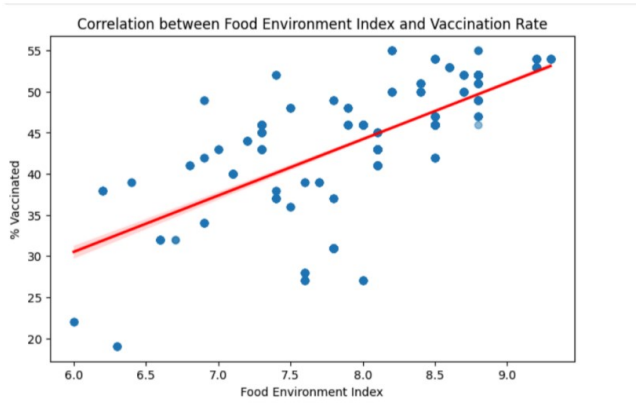


Fig. 1. Scatter plot depicting the relationship between Food Environment Index and Vaccination Rate, showing a downward trend in vaccination rates for areas with lower food accessibility.

XI. Conclusion and Future Work

Integrating SDOH significantly improves predictive accuracy in health analytics. This study highlights:

This study highlights the impact of Social Determinants of Health on predictive modeling and healthcare analytics. The high accuracy of the XGBoost model suggests that machine learning can effectively model complex relationships between socioeconomic factors and health outcomes.

The findings emphasize the importance of using comprehensive datasets that include SDOH for better healthcare decision-making. Policymakers and public health officials can leverage such models for resource allocation, intervention planning, and health policy optimization.

Despite promising results, certain limitations remain. The dataset is confined to California ZIP codes, which may reduce generalizability. Further studies should explore multi-regional datasets to ensure broader applicability. Additionally, real-time data integration and deep learning models can be explored to enhance prediction accuracy.

Future work should focus on expanding the dataset to national and global levels, incorporating real-time environmental data such as pollution and climate conditions, and developing a user-friendly dashboard for decision-makers. Furthermore, ethical considerations, including bias mitigation and data privacy, should be examined when implementing AI-driven healthcare solutions.

References

1) WHO Social Determinants of Health Framework https://iris.who.int/handle/10665/206363.
2) PMC Study on Bias in Pulse Oximetry https://pmc.ncbi.nlm.nih.gov/articles/PMC8765800/.
3) Annual Reviews Paper on Public Health Analytics https://www.annualreviews.org/content/journals/10.1146/annurev-publhealth-031210-101218.
4) Machine Learning–Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients With Heart Failure https://jamanetwork.com/journals/jamacardiology/fullarticle/279372
5) How the SDOH machine learning model improves patients' health and your bottom line https://www.indium.tech/blog/how-the-sdoh-machine-learning-model-improves-patients-health/