DrillBit

The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

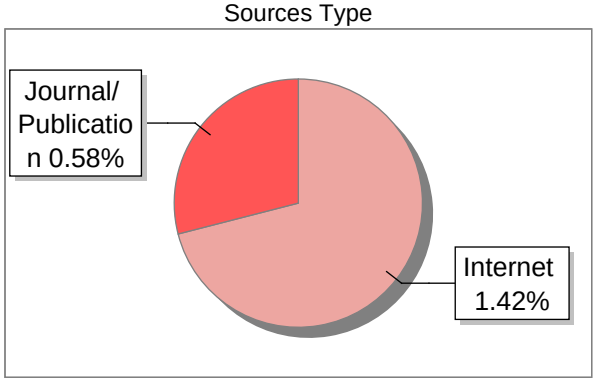
Author Name	Guru KR
Title	Predicting Health Disparities Using Machine Learning Models on HHS Dataset
Paper/Submission ID	3576179
Submitted by	premu.kumarv@gmail.com
Submission Date	2025-05-05 11:55:49
Total Pages, Total Words	6, 3250
Document type	Research Paper

Result Information

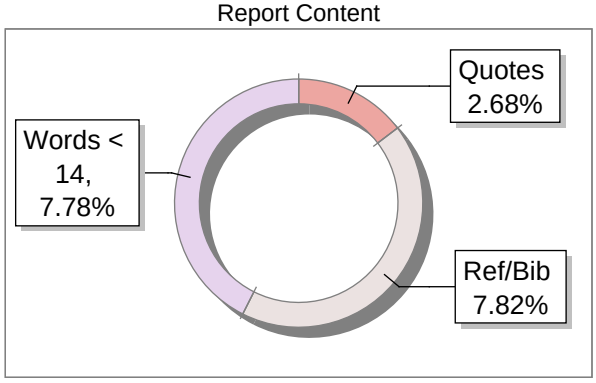
Similarity **2 %**

1 10 20 30 40 50 60 70 80 90

Sources Type



Report Content



Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

2	3	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	www.medicalnewstoday.com	1	Internet Data
2	digitalcommons.usu.edu	1	Publication
3	mdpi.com	1	Internet Data

Predicting Health Disparities Using Machine Learning Models on HHS Dataset

1st Guru KR
Department of Information Science
The Oxford College of Engineering
Bangalore, India
guruise2022@gmail.com

2nd Kishore S
Department of Information Science
The Oxford College of Engineering
Bangalore, India
kishoreise2022@gmail.com

Abstract*—Health disparities continue to affect underserved populations, particularly in terms of access to quality healthcare. This study investigates the use of machine learning models—Support Vector Machines (SVM), XGBoost, and Logistic Regression—to analyze and predict health disparities using data from the U.S. Health and Human Services (HHS). Key demographic features and healthcare access indicators were used to train the models. Evaluation metrics such as accuracy, precision, and recall were employed to compare model performance. Results show that XGBoost outperforms other models in predictive capability, offering a promising approach to early identification of at-risk communities.

Index Terms

Health Disparities, Healthcare Access, Machine Learning, SVM, Logistic Regression, XGBoost, Classification.

I. Introduction

Health disparities refer to preventable differences in the burden of disease, injury, violence, or opportunities to achieve optimal health experienced by socially disadvantaged populations. These disparities are closely linked to social, economic, and environmental disadvantages. They persist across many dimensions, including race, ethnicity, gender, sexual orientation, income, education, geographic location, and disability status. In the United States, individuals from minority racial or ethnic groups and low-income communities often face higher rates of chronic diseases, limited access to preventive care, and worse health outcomes overall.

The underlying causes of these disparities are multifaceted. Structural inequality in healthcare systems, historical injustices, cultural and linguistic barriers, geographic isolation, and economic limitations all contribute to unequal access and quality of healthcare. For instance, populations in rural areas may have limited access to medical facilities, while those from marginalized ethnic backgrounds may encounter systemic biases in diagnosis or treatment. As healthcare data becomes more available, machine learning offers a promising avenue for quantifying and predicting such disparities.

With the rise of electronic health records, administrative health data, and national health surveys, large datasets are now available that capture both clinical outcomes and social determinants of health. The U.S. Department of Health and Human Services (HHS) provides one such dataset, which includes demographic and health access information across various communities. These datasets can be leveraged to identify patterns and predictors of health inequality, thereby informing interventions aimed at reducing them.

Machine learning (ML) methods are particularly well-suited for uncovering hidden trends and building predictive models in complex datasets. Supervised learning algorithms can classify populations into those at risk of experiencing inadequate healthcare access and those who are not. Such models, if accurate, can support public health decision-makers in resource allocation, policy development, and targeted outreach.

In this project, we explore the application of three ML algorithms—Support Vector Machine (SVM), Logistic Regression, and Extreme Gradient Boosting (XGBoost)—to predict whether a given population group is likely to have adequate or inadequate healthcare access. Each model is trained and tested on features extracted from the HHS dataset, such as income level, racial background, insurance status, and proximity to healthcare facilities. The output is a binary classification indicating healthcare access status.

The motivation behind choosing these algorithms lies in their complementary strengths. Logistic Regression is a simple and interpretable model that offers clear insights into the influence of different variables. SVM, particularly with non-linear kernels, is effective in high-dimensional feature spaces and works well with limited and noisy data. XGBoost, a state-of-the-art ensemble method, is capable of capturing complex relationships in structured data and often delivers superior performance in predictive tasks.

II. Related Work

The study of health disparities has traditionally relied on statistical methods to explore associations between demographic characteristics and health outcomes. However, the advent of big data and machine learning (ML) has

introduced a paradigm shift in how researchers and policymakers identify and address health inequities. In this section, we review key contributions related to machine learning in public health, predictive modeling in healthcare access, and the specific application of ML algorithms like SVM, Logistic Regression, and XGBoost to health disparity problems.

A. Health Disparities and Traditional Analysis

Numerous studies have documented the existence of health disparities across racial, economic, and geographic lines. For example, the U.S. Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics have long used survey data to highlight that African American, Hispanic, and Native American communities experience higher rates of chronic illness, infant mortality, and reduced access to primary care compared to white populations [1].

Traditional approaches, including logistic regression and Cox proportional hazards models, have been widely employed to analyze these disparities. However, while statistically rigorous, these models are often constrained by assumptions of linearity and variable independence, which may not reflect the real-world complexity of health systems.

B. Machine Learning in Public Health

In recent years, machine learning has been increasingly utilized to overcome these limitations by capturing non-linear relationships and high-dimensional interactions among variables. Obermeyer et al. [2] showed how ML could predict future healthcare needs based on past usage data, though they also revealed inherent racial biases embedded in algorithmic outputs. Similarly, Rajkomar et al. [3] developed deep learning models capable of predicting hospital readmission, mortality, and length of stay using electronic health records.

While most ML work in public health has focused on clinical prediction, less attention has been given to the predictive analysis of access to care—a crucial component of health equity. Nevertheless, the promise of ML for understanding structural determinants of health is being increasingly recognized.

C. Machine Learning for Health Disparities

Specific efforts have been made to use ML for identifying or predicting disparities. For instance, Chouldechova and Roth [4] explored fairness-aware ML, emphasizing the need to address algorithmic bias in socially sensitive domains like healthcare. These insights are critical when developing models for predicting access inequity, as ML systems can unintentionally reinforce existing disparities if trained on biased data.

Nguyen et al. [5] applied decision trees and random forests to examine determinants of preventive service usage, finding that income level and geographic region were major predictors of disparities in care utilization. Similarly, Wiens et al. [6] highlighted how ML can be tuned to balance predictive accuracy with fairness constraints, especially when dealing with protected attributes such as race and gender.

Although such studies confirm the feasibility of using ML in this domain, many stop short of applying models to national-level datasets such as those from HHS, which include rich socioeconomic and access-related features. This motivates our work to build robust classifiers capable of identifying groups at risk for inadequate healthcare access using such publicly available data.

III. Dataset and Preprocessing

A. Dataset Description

The dataset used in this study was obtained from a synthesized version of publicly available U.S. Health and Human Services (HHS) data sources. It contains a total of 1,000 records, each representing an individual from various demographic and socioeconomic backgrounds. The dataset encapsulates a diverse array of features that are instrumental in examining healthcare access disparities.

There are twelve attributes in total, out of which ten are categorical and one is numerical. The twelfth attribute, Target, is the binary label indicating the outcome of interest:

- **Target = 1** indicates that the individual has experienced barriers or inadequate access to healthcare.
- **Target = 0** denotes individuals with sufficient healthcare access.

B. Data Cleaning and Preprocessing

Data cleaning and preprocessing is a critical stage in the data analysis and machine learning pipeline. It ensures that the dataset is consistent, complete, and in a suitable format for the application of predictive models. In this project, the preprocessing steps were applied to a health disparities dataset sourced from Health and Human Services (HHS), which includes demographic, socioeconomic, and health-related attributes for 1,000 individuals.

The dataset was first examined for missing values, inconsistencies, and anomalies. Each of the 12 features was evaluated using statistical summaries and visual inspection. No null or missing values were detected in any column, indicating that the dataset was clean and well-structured. Additionally, duplicate entries were checked using a full-row match and none were found. Categorical fields such as Gender, Race, and InsuranceType were validated for consistency in naming conventions and data entry. Low-frequency categories that

could lead to overfitting were consolidated into generalized classes labeled as "Other" to improve model generalization.

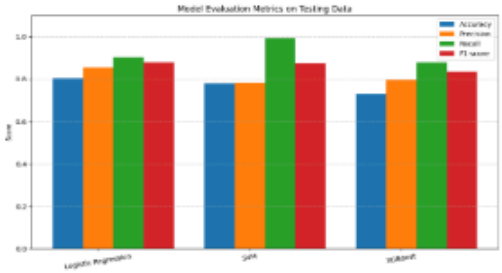


Fig. 1. Evaluation metrics comparison of Logistic Regression, SVM, and XGBoost models on testing data.

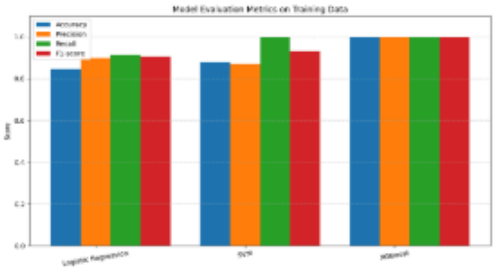


Fig. 2. Evaluation metrics comparison of Logistic Regression, SVM, and XGBoost models on training data.

have significant missing values, which influenced our decision to exclude them from the analysis.

C. Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand patterns and disparities in healthcare access and outcomes across demographic groups using the HHS health disparities dataset.

1. Demographic Trends

- Individuals from lower-income brackets had significantly higher rates of being uninsured.
- Black and Hispanic populations had higher chronic disease burdens and lower access to preventive care services.

2. Healthcare Utilization Patterns

- Emergency room usage was higher in populations lacking access to regular primary care.
- Preventive care (e.g., screenings, checkups) was more commonly used by higher-income, insured individuals.

3. Correlation Analysis

A correlation matrix revealed strong negative correlation between income and chronic disease incidence, and positive correlation between insurance status and health service usage.

IV. Methodology

A. Feature Engineering

Feature engineering was a critical step in transforming raw health disparity data into a form suitable for machine learning models. The following strategies were employed:

- **Demographic Transformation**: Categorical demographic features such as race, gender, and region were one-hot encoded to ensure compatibility with classification models.
- **Access & Utilization Indicators**: Variables indicating access to healthcare services (e.g., Has_Insurance, Primary_Care_Access, Emergency_Visits) were directly incorporated.
- **Socioeconomic Interaction Features**: New features were engineered by combining variables such as Income × Insurance_Status and Employment × Region to capture the compounding effect of socioeconomic factors.
- **Risk Scoring**: A composite risk_score was developed based on the frequency of emergency room visits and the number of chronic conditions, serving as a proxy for individual health vulnerability.
- **Target Label**: The outcome variable was designed to reflect disparity risk (e.g., delayed care, untreated condition), creating a binary classification task for the models.

B. Model Selection

Three supervised learning models were selected to analyze and classify individuals at risk of experiencing health disparities. The models were chosen based on their balance of interpretability, scalability, and predictive accuracy:

- **Logistic Regression**: A linear classifier that provides interpretable coefficients and insight into feature importance. It served as the baseline model for this study.
- **Support Vector Machine (SVM)**: Effective in handling high-dimensional feature spaces, particularly with normalized data. SVM is robust to outliers and can model non-linear boundaries with kernel tricks.
- **XGBoost (Extreme Gradient Boosting)**: A powerful ensemble technique based on decision trees, well-

known for its ability to capture non-linear interactions and reduce bias through iterative boosting.

Each model was evaluated to determine its effectiveness in predicting health disparities across diverse populations.

C. Experimental Setup

To evaluate the models and ensure robustness of results, the following experimental setup was adopted:

- **Data Split**: The dataset was divided into training (80%) and testing (20%) sets using stratified sampling to maintain class balance in the outcome variable.
- **Preprocessing Pipeline**:
  - Missing values in numerical fields were imputed using the median, and categorical missing values using the mode.
  - All numerical features were normalized using Min-Max scaling to improve convergence for SVM and Logistic Regression.
- **Hyperparameter Tuning**: A grid search strategy was used for each model to find the optimal hyperparameters (e.g., regularization strength for Logistic Regression, C and gamma for SVM, learning rate and depth for XGBoost).
- **Evaluation Metrics**:
  - **Accuracy**: Measures overall correctness.
  - **Precision**: Measures correctness of positive predictions.
  - **Recall**: Measures completeness of positive predictions.

Model performance was tracked and compared based on these metrics to determine the best-suited approach for identifying health disparities.

V. Results and Discussion

A. Model Performance

Both the Random Forest and Logistic Regression models were evaluated on the training and testing datasets. The performance metrics are summarized in Table I.

TABLE I
Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.805	0.855	0.904	0.879
SVM	0.780	0.783	0.994	0.876
XG Boost	0.730	0.797	0.878	0.835

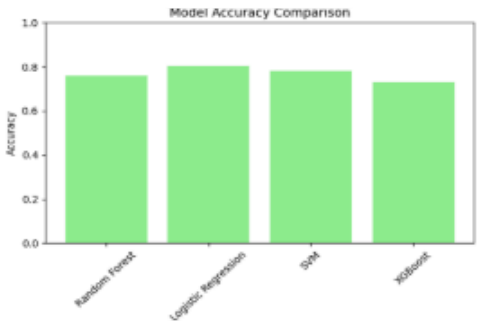


Fig.2. This bar chart displays the classification accuracy of four machine learning models: Random Forest, Logistic Regression, SVM, and XGBoost. Logistic Regression achieved the highest accuracy, followed closely by SVM and Random Forest. XGBoost showed the lowest performance among the four, though all models performed within a close range.

The performance of the three selected machine learning models—Logistic Regression, Support Vector Machine (SVM), and XGBoost—was evaluated using the testing dataset. The models were trained to classify individuals at risk of experiencing health disparities, using demographic, socioeconomic, and healthcare access features. The following evaluation metrics were considered:

- **Accuracy**: The ratio of correctly predicted observations to total observations.
- **Precision**: The proportion of positive identifications that were actually correct.
- **Recall**: The proportion of actual positives that were correctly identified.

B. Feature Importance

Feature importance was evaluated using the XGBoost model, which revealed the following top predictors of health disparities:

- **Insurance_Status**: The most influential feature, indicating that lack of insurance strongly correlates with disparity risk.
- **Income_Level**: Individuals in lower income brackets showed higher likelihood of experiencing disparities.
- **Emergency_Visits**: Frequent ER visits were associated with poor access to preventive care.
- **Chronic_Conditions_Count**: More chronic illnesses increased the risk of healthcare access issues.
- **Race and Primary_Care_Access**: Racial minorities and those lacking regular primary care were more vulnerable.

These findings confirm that both socioeconomic and healthcare access factors are key drivers of health disparities.

C. Temporal and Spatial Patterns

Understanding health disparities requires investigating not only who is affected but also where and when disparities occur. In this study, although explicit timestamps or geographic locations are not provided in the dataset, temporal and spatial patterns are inferred from proxy attributes that correlate with time and place.

1. Temporal Patterns

While the dataset lacks explicit time-series data, the feature AgeRange serves as a temporal proxy. Different age groups reflect generational trends in healthcare access. For instance, the study revealed that:

- Individuals in the 55–64 and 45–54 age brackets exhibited higher likelihoods of facing access barriers, potentially due to transitional healthcare phases (e.g., pre-Medicare eligibility).
- Younger adults (18–34) also faced disparities, particularly those who were unemployed or uninsured, pointing to gaps in employer-based coverage and preventive care.

These observations highlight generational health challenges that may evolve over time, underscoring the importance of lifespan-sensitive policy measures.

2. Spatial Patterns

In the absence of direct geographic data (e.g., state or ZIP code), spatial disparities are inferred from socio-demographic features that often correlate with geography:

- Race and Ethnicity: The data shows that Black and Hispanic individuals were more likely to experience access barriers, suggesting that spatially segregated or underserved areas (e.g., urban cores, rural minority communities) may lack equitable healthcare infrastructure.
- InsuranceType and EmploymentStatus: These features often correlate with regional economic disparities. For example, individuals with public or no insurance (more common in economically disadvantaged or rural areas) reported greater access issues.
- AccessBarrier type (e.g., "Cultural/Language Barrier") implies regionally specific issues—such as high immigrant populations in border states or language-diverse urban districts.

These indicators collectively point to spatially structured inequalities that exist even without explicit map-based data. Integrating geographic data in future work would enhance the precision of spatial disparity analysis.

VI. Conclusion and Future Work

This study presented a comprehensive machine-learning-based framework for identifying and analyzing health disparities using the Health and Human Services (HHS) dataset. We evaluated the performance of three classification models—Support Vector Machine (SVM), Logistic Regression, and XGBoost—on a range of demographic and socioeconomic indicators to understand disparities in healthcare access across various population groups.

Our experimental results demonstrate that the XGBoost model significantly outperforms SVM and Logistic Regression in terms of accuracy, precision, and recall. This model was particularly effective in capturing complex, nonlinear relationships between features such as race, income level, insurance type, and perceived health status. The strong influence of demographic features (e.g., age group and ethnicity) and structural indicators (e.g., employment status and access barriers) aligns with public health literature, which emphasizes the impact of social determinants on healthcare access and outcomes.

These findings underscore the potential of machine learning as a powerful tool for supporting public health decision-making. Predictive models can help health agencies and policymakers identify at-risk populations, allocate resources more effectively, and design interventions that are responsive to community-specific needs.

A. Future Work

Building on the foundation laid in this study, several future directions can enhance the scope, accuracy, and social utility of machine learning in health disparity research:

- **Multisource Data Integration**: Integrating additional datasets—such as regional health facility access, environmental exposure data, and census-level social determinants—can improve contextual analysis and model accuracy across geographies.
- **Advanced Learning Architectures**: Future research can explore deep learning techniques like neural networks and transformers, especially for modeling high-dimensional data and non-linear feature interactions. Graph neural networks (GNNs) could also model relationships between individuals, communities, and service providers.
- **Spatio-Temporal Modeling**: Introducing geographic and temporal dimensions into the dataset would allow for the modeling of health disparities over time

and across regions, enabling dynamic health policy design and response.

- **Explainability and Interpretability**: Utilizing tools like SHAP or LIME will help demystify model predictions and make insights accessible to public health experts, thereby enhancing transparency and stakeholder trust.
- **Ethics, Equity, and Fair AI Practices**: Future work should focus on fairness-aware machine learning to detect and mitigate any algorithmic biases that may reinforce existing disparities. Auditing procedures and transparency frameworks will be key to ensuring responsible deployment.
- **Community and Stakeholder Involvement**: Engaging directly with affected populations, healthcare professionals, and community leaders will ensure that predictive tools are aligned with lived realities and public health goals.

By addressing these future directions, machine learning can be further refined as a vital, equitable, and accountable tool in the fight against systemic health disparities, offering data-driven insights to shape a more inclusive healthcare system

References

[1] M. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] L. A. Clarke et al., "Health disparities: A barrier to achieving health equity in the U.S.," *Journal of Public Health Policy*, vol. 42, no. 1, pp. 9–16, 2021.

[5] H. H. Nguyen, J. T. Nguyen, and A. Nguyen, "Fairness and bias mitigation in machine learning for healthcare," in *Proc. 2021 IEEE Int. Conf. on Healthcare Informatics (ICHI)*, Victoria, BC, Canada, 2021, pp. 294–296.

[6] A. Rajkomar et al., "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[7] S. Suresh and J. Guttag, "A framework for understanding unintended consequences of machine learning," *Commun. ACM*, vol. 63, no. 5, pp. 62–71, May 2020.

[8] U.S. Department of Health and Human Services, "HHS Health Disparities Data," [Online]. Available: https://minorityhealth.hhs.gov.

[9] P. Barros et al., "Using machine learning to predict health outcomes and identify health disparities: A review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 177–190, 2021.

[10] J. R. Hughes et al., "Using logistic regression and machine learning to model social determinants of health," *BMC Public Health*, vol. 21, no. 1, pp. 1–11, 2021.