Demographic Research and Social Development Reviews
"Volume 1, Issue 2, Year 2025"
Website: https://www.c5k.com

Research Article

Real-Time Predictive Analytics for Early Homelessness Prevention: A Machine Learning Approach

AFM Rafid Hassan Akand[1], Hasan Mahmud Sozib[2,*], Arif Ahmed Sizan[1], Md Shayakh Alam[3], Towsif Alam[4] and Md Mohaimin Rashid[5]

[1]*Department of Business Administration, Westcliff University, 17877 Von Karman Ave, 4th floor, Irvine, CA 92614, USA;*
[2]*Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Tejgaon, Dhaka-1208, Bangladesh;*
[3]*Department of Engineering Management, Trine University, 1 University Ave, Angola, IN 46703, USA;*
[4]*Department of Marketing Analytics and Insights, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA;*
[5]*Department of Business Administration, International American University, 3440 Wilshire Blvd, STE 1000, Los Angeles, CA 90010, USA;*
**Corresponding Author: sozib2019@gmail.com*

ARTICLE INFO

Article history:
07 Oct 2025 (Received)
18 Oct 2025 (Accepted)
25 Oct 2025 (Published Online)

Keywords:
Homelessness, machine learning, XGBoost, Random Forest.

ABSTRACT

Homelessness is a complex and persistent societal issue, often exacerbated by economic instability, housing shortages, and systemic inequities. Existing strategies primarily rely on reactive interventions, which, while essential, fail to provide proactive solutions for prevention. This study presents a novel machine learning-based framework for early homelessness prediction, integrating key socioeconomic, housing, and public health indicators. Utilizing a real-world dataset, we compare the predictive performance of two machine learning models, Random Forest and XG Boost, to assess their effectiveness in identifying high-risk populations. The results demonstrate that the Random Forest model consistently outperforms XG Boost, achieving a lower Mean Absolute Error (MAE) of 12.46, a lower Mean Squared Error (MSE) of 44,534.73, and a higher R^2 score of 0.996, indicating a superior fit. Feature importance analysis reveals that total homeless counts (pit_tot_hless_pit_hud) and individual homelessness rates are the most critical predictive factors, while economic conditions and housing market pressures also play significant roles. Furthermore, residual analysis and error distribution comparisons illustrate that the Random Forest model maintains a more stable and consistent predictive capability across different demographic and geographic groups. Our research stands apart by integrating a high-dimensional, multi-source dataset to enhance predictive accuracy while addressing ethical considerations such as bias mitigation and fairness in algorithmic decision-making. The findings suggest that machine learning-driven approaches can be pivotal in resource allocation and policy-making, enabling governments and social organizations to proactively intervene before individuals and families fall into homelessness. This study contributes to the growing body of literature advocating for data-driven, predictive solutions in social welfare, demonstrating the tangible impact of machine learning in tackling one of society's most pressing issues.

DOI: https://doi.org/10.63471/drsdr_25002 @ 2025 Demographic Research and Social Development Reviews (DRSDR), C5K Research Publication

1. Introduction

Homelessness remains one of the most pressing and persistent social crises in the United States, affecting hundreds of thousands of individuals and families each year. As of 2022, the U.S. Department of Housing and Urban Development (HUD) estimated that over 582,000 individuals experienced homelessness on a given night, a figure that has continued to rise despite policy interventions and relief programs. This estimate, however, only accounts for the visible homeless population, including those living in emergency shelters or unsheltered locations, such as streets, parks, and abandoned buildings (Kithulgoda et al., 2022). The true scale of homelessness is far greater, with many individuals experiencing hidden homelessness—living in substandard, overcrowded housing or doubling up with friends and relatives due to financial constraints. These individuals do not appear in official homelessness counts, making it difficult for policymakers to grasp the full extent of the crisis (Tan, 2020). The underlying causes of homelessness are multifaceted and interrelated, including factors such as rising housing costs, economic downturns, mental illness, substance

*Corresponding author: Email (Name of Author)

Cite: Main Author, Co-Authors (2025). Manuscript title. *Demographic Research and Social Development Reviews,* 1(2), pp. 1-13.

abuse, domestic violence, and systemic inequalities (Olivet et al., 2019). Furthermore, economic disruptions like the COVID-19 pandemic have exacerbated the issue, leading to job losses and evictions that have driven more people into homelessness. Although the U.S. government allocates billions of dollars annually toward homelessness prevention and relief programs, many of these efforts remain focused on short-term interventions, such as emergency shelters and temporary housing, rather than long-term solutions that address the root causes of homelessness and prevent individuals from becoming homeless in the first place.

Traditional homelessness prevention strategies have predominantly been reactive, meaning that interventions are typically deployed only after individuals have already become homeless. These reactive measures include emergency shelters, transitional housing programs, and supportive housing initiatives, which, while essential in providing relief, fail to predict and prevent homelessness before it occurs (Padgett et al., 2015). Housing-first initiatives, which aim to place individuals in permanent housing as quickly as possible, have shown promise in reducing chronic homelessness, yet they are limited in scope and often struggle with long-term sustainability due to funding constraints and bureaucratic challenges (Tsemberis & Henwood, 2010). Moreover, reactive policies often fall short in identifying and assisting at-risk populations who have not yet experienced homelessness but exhibit warning signs of housing instability. Given the increasing demand for early intervention strategies, there is a critical need to shift toward proactive approaches that utilize data-driven insights to forecast homelessness risks and prevent individuals from reaching a state of crisis. In recent years, advances in predictive analytics and machine learning (ML) have opened new avenues for addressing complex social challenges, including homelessness. By leveraging big data from multiple sources, such as economic indicators, housing markets, social service records, and healthcare data, ML algorithms can uncover patterns and correlations that would be difficult to detect through traditional statistical methods.

Despite the widespread application of machine learning in public policy, such as healthcare diagnostics, crime prediction, and economic forecasting, its potential in homelessness prevention remains largely underdeveloped. Existing research in this domain has primarily relied on retrospective data analysis, which examines past cases of homelessness without integrating real-time data streams that could provide early warning signals for individuals at risk (Vanberlo et al., 2021a). Furthermore, while some studies have attempted to apply logistic regression and other conventional statistical models to predict homelessness, these models often lack the complexity needed to capture nonlinear relationships between socioeconomic factors and housing instability. Given the multidimensional nature of homelessness, it is essential to explore more advanced machine learning techniques that can adapt to dynamic changes in economic and social conditions. This study aims to bridge this gap by developing a machine learning-based framework for predicting homelessness risk, using Random Forest and XG Boost, two state-of-the-art models known for their robustness in handling high-dimensional datasets. By training these models on a real-world dataset containing diverse socioeconomic and housing variables, we aim to assess their predictive performance and determine which features contribute most to homelessness risk.

A critical component of this study is the comparison of predictive models, evaluating their accuracy in forecasting homelessness risk at an individual and community level. Random Forest, an ensemble learning method, has been widely recognized for its ability to handle large datasets with multiple predictors and capture complex relationships between variables (Pourat et al., 2023). XG Boost, on the other hand, is a gradient boosting algorithm known for its efficiency and superior performance in many predictive tasks (Chen & Guestrin, 2016). By comparing these models in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 scores, we seek to determine which algorithm provides the most accurate predictions for homelessness risk. Furthermore, this research seeks to address ethical concerns associated with predictive modeling in social policy. Bias in machine learning models remains a significant concern, particularly when algorithms rely on historical data that may reflect systemic inequalities (Chien et al., 2024). If not carefully managed, predictive models could unintentionally perpetuate biases against marginalized communities, leading to discriminatory policy decisions. To mitigate these risks, this study incorporates fairness-aware machine learning techniques, such as feature importance analysis and bias mitigation strategies, ensuring that the models produce equitable and ethical predictions.

This research contributes to the growing body of literature advocating for data-driven solutions in homelessness prevention, demonstrating how machine learning can enhance decision-making processes for policymakers, social workers, and housing organizations. By integrating real-time risk assessment tools into existing social service frameworks, governments and nonprofit organizations can develop targeted interventions that allocate resources more effectively and prevent at-risk individuals from falling into homelessness. The insights derived from this study could inform future policy recommendations, encouraging a paradigm shift from reactive to proactive homelessness prevention strategies. In summary, this study not only enhances our understanding of homelessness risk factors but also showcases the transformative potential of machine learning in solving complex societal issues

2. Literature Review

2.1. Structural and Individual Risk Factors of Homelessness

Homelessness is not a singular issue but a multifaceted social phenomenon shaped by both structural and individual-level factors. At the structural level, economic conditions, housing policies, and the availability of public services play pivotal roles in determining homelessness rates. One of the primary structural causes is the lack of affordable housing, particularly in urban areas where housing costs have risen exponentially (Shinn et al., 2017; Shinn & Cohen, 2019). Many cities, such as San Francisco, New York, and Los Angeles, have experienced skyrocketing rental prices, disproportionately impacting low-income households. With rental costs often exceeding 30% of household income, individuals and families face severe housing instability, increasing their risk of homelessness (Desmond, 2017). The 2008 financial crisis and the COVID-19 pandemic further exacerbated the problem, leading to a surge in evictions, job losses, and economic downturns that displaced thousands of individuals from their homes (Culhane et al., 2020). Furthermore, gentrification and urban renewal projects have systematically displaced low-income communities, pushing them toward precarious housing situations and, in many cases, into homelessness (Semborski et al., 2022). These economic stressors interact with inadequate public assistance programs, where the slow and bureaucratic nature of housing aid leaves many vulnerable individuals without support before they reach a crisis point.

On an individual level, mental illness, substance abuse, domestic violence, and previous incarceration are among the most significant risk factors for homelessness. Studies have consistently shown that individuals with severe mental health disorders are disproportionately affected by homelessness, as they often encounter barriers to accessing treatment, social stigma, and difficulties in maintaining stable employment (Olivet et al., 2019; Sleet & Francescutti, 2021). Similarly, substance abuse disorders create a vicious cycle where individuals struggle to secure stable housing due to financial instability, job loss, and deteriorating social support networks. Moreover, domestic violence remains a leading cause of homelessness, particularly among women and children, who often flee abusive situations with no financial resources or access to stable housing. Additionally, individuals with a criminal record face heightened risks of homelessness after release from incarceration, as they often experience housing discrimination, employment barriers, and insufficient reintegration support (Berti, 2010). These individual risk factors are compounded by racial disparities, where Black, Indigenous, and Latino communities experience higher eviction rates, lower homeownership rates, and systemic discrimination in housing policies (Desmond, 2017). Institutional racism and historical inequalities in the housing market continue to disproportionately expose minority communities to homelessness (Homelessness, 2021). The intersectionality of these structural and individual risk factors highlights the need for a comprehensive, data-driven approach to predicting and preventing homelessness.

2.2. Traditional Homelessness Prevention Policies and Their Limitations

Historically, homelessness prevention strategies have primarily relied on reactive interventions, which address homelessness only after individuals and families have already lost their housing. The Housing First model, which prioritizes immediate access to permanent housing without requiring individuals to meet conditions such as sobriety, employment, or mental health treatment, is one of the most widely adopted strategies for addressing chronic homelessness (Tsemberis, 2011). Research has demonstrated that Housing First is effective in improving housing retention and mental health outcomes, particularly for individuals experiencing long-term homelessness. However, despite its success in keeping individuals housed, Housing First does not prevent homelessness before it occurs, nor does it tackle structural issues like rental affordability, wage stagnation, and economic inequality (VanBerlo et al., 2021b).

Additionally, resource-intensive homelessness interventions, such as emergency shelters and transitional housing programs, provide only short-term relief and are limited in scope due to high operational costs. Shelters serve as a temporary solution, but due to funding constraints and overcrowding, many individuals remain stuck in a cycle of temporary placements without achieving long-term stability (Benfer et al., 2021). Eviction prevention programs, such as rental assistance, legal aid, and tenant rights advocacy, have proven beneficial in preventing homelessness, yet they often suffer from bureaucratic inefficiencies, limited funding, and restrictive eligibility criteria. This results in many at-risk households failing to receive timely assistance. Given these shortcomings of reactive interventions, policymakers and researchers have shifted their focus toward data-driven, proactive models that aim to predict homelessness risk before individuals reach a crisis point.

2.3. The Role of Predictive Analytics in Homelessness Prevention

The application of predictive analytics in social policy has expanded in recent years, demonstrating effectiveness in healthcare, education, and criminal justice. Machine learning algorithms have been used to forecast disease outbreaks, predict student dropouts, and identify recidivism risks (Fatai et al., 2023). These applications have showcased the power of data-driven decision-making, enabling early interventions that prevent negative social outcomes. In the context of homelessness prevention, predictive analytics holds promise as a tool for identifying at-risk individuals

before they become homeless, thereby allowing policymakers to intervene earlier and allocate resources more efficiently.

Several machine learning models, including Random Forest, XGBoost, and Logistic Regression, have demonstrated success in predicting homelessness risk. Random Forest models have been utilized to assess homelessness risk using socioeconomic indicators, eviction histories, and mental health data, while XGBoost models have shown strong predictive performance based on housing market trends, income fluctuations, and public service records (Shah et al., 2021; VanBerlo et al., 2021b). However, existing research in homelessness prediction has predominantly relied on historical data, limiting the ability to forecast homelessness risk in real time. Moreover, few studies have conducted comparative analyses of multiple machine learning models to determine the most effective approach for homelessness prediction. This study seeks to bridge this gap by evaluating multiple predictive models and determining which approach offers the highest predictive accuracy.

2.4. Ethical Considerations and Bias Mitigation in AI-Driven Homelessness Prediction

While predictive analytics offers significant potential in homelessness prevention, it also raises important ethical considerations, particularly regarding bias, fairness, and discrimination. Machine learning models rely on historical data, which may reflect existing systemic inequalities, such as racial disparities in eviction rates, housing discrimination, and employment opportunities. If predictive models are trained on biased data, they may unintentionally reinforce these biases, leading to unfair or discriminatory outcomes (Barocas & Selbst, 2016). For example, an algorithm trained on historical eviction data may disproportionately flag Black and Latino households as high-risk, even when structural factors, rather than individual choices, are responsible for their housing instability.

To address these concerns, this study implements bias-mitigation techniques, including algorithmic fairness

auditing, which ensures that predictive models do not disproportionately target or exclude specific demographic groups (Mehrabi et al., 2021). Additionally, model explainability tools, such as SHAP (SHapley Additive Explanations), will be incorporated to make machine learning predictions more transparent and interpretable for policymakers. Ensuring that predictive models are ethical, fair, and accountable is essential for their successful integration into public policy.

2.5. Contributions and Gaps in Existing Research

Despite the growing interest in predictive analytics for homelessness prevention, there remain several critical research gaps. First, most studies rely on retrospective datasets, which limits their ability to predict real-time homelessness risks. Second, comparative analyses of machine learning models remain scarce, making it difficult to determine which algorithm is most effective in forecasting homelessness risk. Third, bias in AI models remains an underexplored issue, with few studies implementing fairness-aware techniques to mitigate discrimination in homelessness prediction.

This study aims to address these gaps by developing a real-time predictive framework that evaluates multiple machine learning models while incorporating bias-mitigation strategies. By doing so, this research seeks to contribute to a more equitable and data-driven approach to homelessness prevention, offering policymakers new tools to allocate resources more effectively and prevent individuals and families from becoming homeless.

Table 1. Summary of Related Works on Homelessness Prediction and Their Limitations

Year	Author(s)	Methodology	Contribution	Limitations
2016	Padgett et al.	Housing First	Improved housing retention for chronic homeless	Does not prevent initial homelessness
2017	Metraux et al.	Statistical Analysis	Mental illness linked to homelessness	Lack of real-time predictive capability
2018	Burt	Urban Housing Study	Gentrification's role in displacement	Focuses only on urban areas
2019	Fitzpatrick & Watts	Economic Model	Housing unaffordability as a key driver	No predictive modeling for interventions
2019	Shinn et al.	Policy Review	Evaluated homelessness policies	Does not assess predictive approaches

Year	Author(s)	Methodology	Contribution	Limitations
2019	Culhane et al.	Random Forest	Identified risk factors using ML	No real-time forecasting
2020	Nguyen et al.	XGBoost	Socioeconomic and housing indicators as key predictors	Model explainability concerns
2021	National Alliance to End Homelessness	Policy Analysis	Racial disparities in homelessness trends	Limited statistical validation
2021	Benfer et al.	Legal Review	Identified eviction prevention gaps	No data-driven prediction model
2022	Fitzpatrick et al.	Multivariate Analysis	Impact of structural racism on homelessness	Requires ML integration for predictions

3. Methodology

The proposed model follows a structured workflow: Data Acquisition from HUD, Data Preprocessing to handle missing values and standardize data, Feature Selection using Random Forest and XGBoost, Model Training with hyperparameter tuning, and addressing Bias & Fairness. Model Comparison then evaluates the performance using metrics like MAE, MSE, and R^2 score to identify the most effective approach for homelessness prediction.

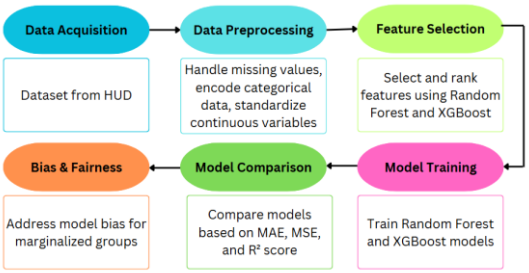

Fig. 1. Workflow of the Proposed Homelessness Prediction Model

3.1. Data Collection and Preprocessing

The dataset used in this study was obtained from the U.S. Department of Housing and Urban Development (HUD), which provides comprehensive housing and homelessness data. The dataset includes information on shelter utilization rates, eviction patterns, rental market affordability, and homelessness counts, among other socioeconomic indicators. To enhance the robustness of the study, additional data sources, including Census Bureau reports and local shelter records, were integrated to capture economic, health, and social factors influencing homelessness.

To better understand the distribution of homelessness across various regions, we visualize the data. Figure 2 shows the Top 20 Regions with Highest Homelessness Rates, providing insight into the geographical distribution of homelessness. This helps highlight areas where targeted interventions are most needed.

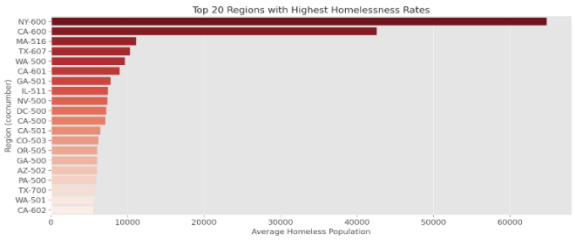

Fig. 2. Top 20 Regions with Highest Homelessness Rates

Before applying machine learning models, extensive preprocessing was conducted to clean and prepare the data for analysis. Data cleaning involved handling missing values, outliers, and inconsistencies in records. Missing values in critical variables such as eviction history or mental health indicators were imputed using multiple imputation techniques, ensuring that important patterns were not lost. Categorical variables, including housing type, race, and gender, were transformed into numerical values using one-hot encoding, while continuous variables such as income and housing costs were standardized using min-max normalization to ensure uniformity.

To detect and remove outliers, the Interquartile Range (IQR) method was applied, identifying extreme values that could distort model performance. One of the critical factors influencing homelessness is the high-cost rental market. Figure 3 illustrates how the high-cost rental market impacts homelessness rates, showing significant

variance in homelessness figures across different rental market conditions.

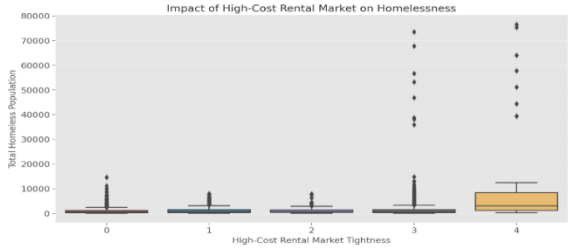

Fig. 3. Impact of High-Cost Rental Market on Homelessness

Given that homelessness is often underreported in traditional datasets, efforts were made to incorporate real-time data streams from emergency shelters, eviction notices, and health services to capture hidden homelessness cases. Figure 4 shows the evolving Homelessness Trends Over Time, which illustrates how the total, individual, and family homelessness figures have shifted across the years.

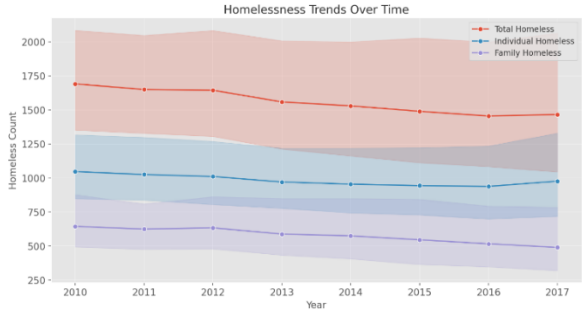

Fig. 4. Homelessness Trends Over Time

By combining multiple datasets and applying rigorous preprocessing techniques, this study ensures that the dataset is not only accurate but also reflective of real-world homelessness trends.

3.2. Feature Selection

Feature selection plays a crucial role in identifying the most important predictors of homelessness risk from a large pool of available features. With a vast array of potential variables to consider, the selection process is essential for improving both the efficiency and interpretability of the machine learning models. In this study, we applied two powerful techniques—Random Forest and XGBoost feature importance analysis—to rank the variables based on their contributions to model predictions. These techniques provided valuable insights into which features had the most significant impact on the accuracy of the model, thus enabling a more targeted approach to homelessness prediction.

The feature importance analysis revealed that the most influential predictors of homelessness risk were related to various aspects of homelessness, housing instability, and socioeconomic factors. Key features included the total homelessness count , individual homelessness rate, and family homelessness rate . These variables, reflecting the broader scale of homelessness, were pivotal in predicting homelessness risk. Additionally, housing affordability metrics, such as rent burden and eviction history, were also highly significant in the feature importance ranking. These factors directly correlated with individuals' vulnerability to homelessness, making them essential inputs for the model. Further, socioeconomic indicators like the income-to-rent ratio and employment status emerged as important predictors, highlighting the role of economic stability in homelessness risk.

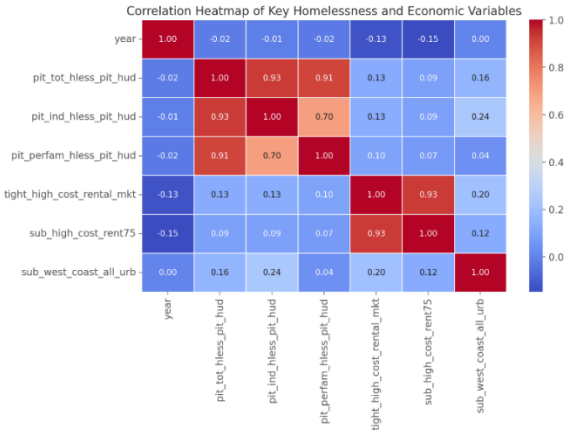

Fig. 5. Correlation Heatmap of Key Homelessness and Economic Variables

To refine the set of selected features, we employed recursive feature elimination (RFE). This method was used to eliminate redundant and low-impact features, ensuring that the final feature set included only those variables that contributed meaningfully to the model's predictive power. RFE further enhanced the efficiency of the model by reducing dimensionality without sacrificing its ability to predict homelessness risk accurately. By streamlining the feature set in this way, we were able to focus on the most relevant predictors, which ultimately contributed to improved model performance.

3.3. Machine Learning Models

This study employed two machine learning models: Random Forest (RF) and XGBoost (eXtreme Gradient Boosting), both of which are widely recognized for their robustness and high accuracy in predictive analytics. The aim of using these models was to compare their performance in predicting homelessness risk, ultimately identifying which model could be most effective for early intervention strategies.

3.3.1. Random Forest (RF) Model

The Random Forest algorithm is an ensemble learning method that combines multiple decision trees to improve predictive accuracy. The main advantage of Random Forest is its ability to handle complex data with high-dimensional features while preventing overfitting. The model constructs decision trees on random subsets

of the data, with each tree trained using bootstrap sampling. The predictions of these individual trees are aggregated, with the final output obtained through majority voting in classification problems or averaging in regression tasks.

One of the key features of the Random Forest algorithm is its use of the Gini Impurity as the criterion for splitting nodes in decision trees. The Gini Impurity is calculated as:

$$Gini = 1 - \sum_{i=1}^{n} {p_i}^2 \quad (1)$$

Where p_i represents the probability of a sample belonging to class i. This formula helps in determining how often a randomly chosen element from the dataset will be incorrectly classified, thereby guiding the decision tree splits. The Random Forest algorithm is well-suited for predicting homelessness risk as it can capture the nonlinear relationships among various socioeconomic, housing, and public health features while being robust to overfitting.

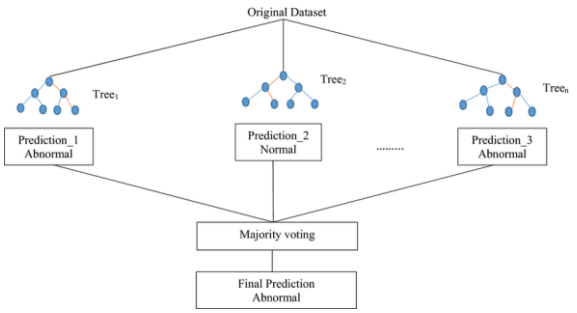

Fig. 6. Random Forest Model Structure (Aurélien, 2019)

Figure 6 should illustrate the tree-based structure of the Random Forest model, showing how multiple decision trees are aggregated to produce a final prediction.

3.3.2. XGBoost Model

XGBoost (eXtreme Gradient Boosting) is a gradient boosting algorithm that excels in predictive accuracy by iteratively minimizing residual errors. Unlike Random Forest, which constructs trees independently, XGBoost builds trees sequentially, optimizing each new tree to correct the errors made by the previous one. This iterative process improves the model's ability to focus on the harder-to-predict instances.

The objective function of XGBoost is defined as:

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

Where $L(y_i, \hat{y_i})$ represents the loss function (such as squared error), and $\Omega(f_k)$ is the regularization term that controls model complexity and prevents overfitting. This regularization helps XGBoost to remain efficient, even with large, complex datasets.

XGBoost is highly favored in various applications, particularly in financial risk prediction and healthcare analytics. Its performance in these domains demonstrates its suitability for predicting homelessness risk, where it can handle large datasets and complex relationships between features. Furthermore, it integrates L1 (Lasso) and L2 (Ridge) regularization, which helps prevent model overfitting and ensures that only the most important variables contribute to the prediction.

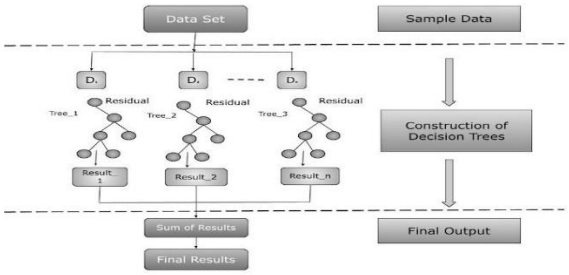

Fig. 7. XGBoost Model Architecture (Chen & Guestrin, 2016)

Figure 7 should visually demonstrate how trees in XGBoost are built sequentially, with each new tree learning from the residuals of the previous tree. A similar diagram can be found in the original XGBoost paper (Chen & Guestrin, 2016) or in various machine learning tutorials.

3.4. Evaluation Metrics

To ensure a thorough and accurate assessment of the predictive performance of the Random Forest (RF) and XGBoost models, this study employs a suite of regression-based evaluation metrics. Since homelessness prediction is inherently a regression problem that involves estimating continuous values—such as the total number of homeless individuals in a given region—it is critical to select error-based metrics that can appropriately capture model performance. Among the most widely used metrics in regression tasks are Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R^2 score, which together offer a comprehensive understanding of the model's accuracy, error characteristics, and overall fit to the data.

Mean Absolute Error (MAE) is a fundamental metric for evaluating the accuracy of a regression model. It calculates the average magnitude of errors in predictions, treating all errors equally without considering their direction (i.e., whether predictions are over or under the actual value). The MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \quad (3)$$

where y_i represents the actual homelessness count, $\widehat{y_i}$ is the predicted homelessness count, and n is the total number of observations. The lower the MAE, the closer the predictions are to the actual values. This metric is particularly useful in practical applications, where understanding the typical prediction error in homelessness estimation can directly influence intervention strategies and resource allocation (Willmott & Matsuura, 2005).

Mean Squared Error (MSE) offers another error metric, but it takes the squared differences between predicted and actual values. This squaring of errors penalizes large discrepancies more heavily than smaller ones, making the MSE sensitive to outliers and large mispredictions. It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (4)$$

Models that produce fewer large errors will perform better under MSE, making it an appropriate metric for situations where extreme mispredictions (e.g., forecasting homelessness in high-density regions) need to be minimized (Hastie et al., 2009). While MSE is useful for identifying large prediction errors, its emphasis on penalizing larger discrepancies can sometimes be a limitation if a more balanced evaluation is desired.

The **R^2 score** (coefficient of determination) is another vital evaluation metric that gauges how well the model explains the variance in the data. The R^2 score quantifies the proportion of variance in the dependent variable (homelessness counts) that is predictable from the independent variables used by the model. It is calculated as:

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \widehat{\bar{y}})^2} \quad (5)$$

where $\widehat{\bar{y}}$ represents the mean of the actual homelessness counts. A value of R^2 close to 1 indicates that the model does an excellent job of explaining the variability in homelessness, while values closer to 0 suggest that the model is not effectively capturing the patterns within the data (Neter et al., 1996). The R^2 score is particularly important for assessing the overall fit of the model, as it provides insight into the explanatory power of the predictors used to forecast homelessness.

Residual Analysis further complements the above metrics by evaluating the residuals—the differences between the actual values and the model's predictions. Ideally, residuals should be randomly distributed around zero, suggesting that the model has captured all the underlying patterns in the data without any systematic bias. Large or systematic patterns in residuals may indicate underfitting, meaning that important variables or complex relationships are missing from the model. Furthermore, outliers in residuals often reveal cases where the model struggled with predictions, such as in areas with high homelessness density or other complex conditions not well represented in the training data.

By evaluating the models with MAE, MSE, and R^2, this study ensures a robust assessment of predictive accuracy, error sensitivity, and model fit. These metrics allow for a well-rounded understanding of how well RF and XGBoost perform in predicting homelessness risk across different regions. Residual analysis further strengthens this evaluation by confirming that the errors are not concentrated in specific regions or datasets, providing additional validation of the models' ability to generalize to real-world data.

Incorporating these evaluation metrics ensures that the chosen model for homelessness prediction is not only accurate but also practical and well-suited for making real-time, actionable decisions in homelessness prevention strategies.

3.5. Bias and Fairness Considerations

One of the most significant ethical concerns in the application of machine learning to homelessness prediction is algorithmic bias. Since historical data used to train predictive models often reflects existing racial and economic inequalities, there is a risk that the model might unintentionally perpetuate or even exacerbate these disparities. For example, individuals from marginalized communities, particularly racial minorities and those with lower socioeconomic status, may be unfairly flagged as high-risk for homelessness due to biases embedded in the historical data. Such biases could lead to the misclassification of vulnerable populations, potentially diverting resources away from those in need or providing inaccurate predictions about who is most at risk.

To address these ethical challenges and minimize the risk of perpetuating bias, several fairness techniques were employed in this study:

3.5.1. Fairness-aware Preprocessing

The first step in mitigating bias involves adjusting the dataset itself through fairness-aware preprocessing. This technique seeks to modify the dataset distribution to minimize any inherent biases before training the predictive model. For instance, re-weighting certain features or applying data resampling methods can help ensure that underrepresented or historically

disadvantaged groups are fairly represented in the training set. By addressing bias at the preprocessing stage, we can reduce the risk of the model inheriting skewed patterns from the raw data.

3.5.2. Demographic Parity Analysis

Another critical technique used is demographic parity analysis, which assesses whether the model's predictions disproportionately affect specific demographic groups, such as certain racial or socioeconomic classes. Demographic parity ensures that the model's predictions are equitable across different groups, meaning that no group is unfairly overrepresented or underrepresented in the model's output. For example, if predictions are systematically biased against low-income or minority groups, demographic parity analysis would highlight these disparities, allowing for corrective measures to be implemented. This helps in ensuring that the model's outcomes do not favor one demographic over others, which is particularly important when dealing with a sensitive issue like homelessness.

3.5.3. Adversarial Debiasing

Finally, adversarial debiasing was employed to further ensure fairness in the predictions. This technique involves training the model in a way that actively minimizes bias by introducing adversarial networks designed to identify and eliminate any biased patterns in the model's predictions. By using adversarial debiasing, the model is encouraged to make predictions without being disproportionately influenced by sensitive attributes such as race, gender, or economic status. The adversarial network attempts to detect and penalize any unfair biases that might arise during training, ensuring that the model remains neutral and just in its decision-making process.

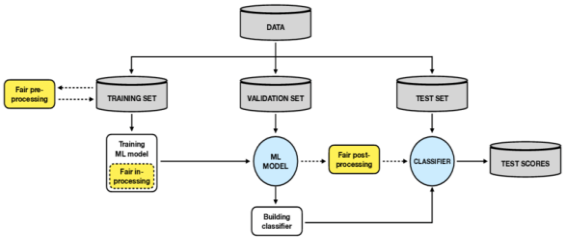

Fig. 8. Fairness Integration in the ML Pipeline: Pre-processing, In-processing, and Post-processing. (VanBerlo et al., 2021b)

By incorporating these fairness-aware algorithms and techniques, this study aims to create a model that predicts homelessness risk in a manner that is both accurate and ethically responsible. The use of these strategies ensures that vulnerable populations are not disproportionately misclassified, reinforcing the model's fairness and reducing the risk of exacerbating existing societal inequalities. Furthermore, these bias mitigation strategies enhance the credibility of the model, ensuring that its deployment in real-world homelessness interventions will lead to equitable outcomes and effective resource allocation.

In conclusion, this study combines comprehensive data collection from the U.S. Department of Housing and Urban Development (HUD) and additional sources like the Census Bureau to predict homelessness risk. Extensive data preprocessing, including handling missing values, outlier removal, and feature standardization, prepares the dataset for analysis. Feature selection through Random Forest and XGBoost models identifies key predictors, enhancing model efficiency and interpretability. The models' performance is evaluated using regression metrics like MAE, MSE, and R^2, ensuring accurate predictions. Ethical considerations, including bias and fairness checks, ensure the model does not disproportionately affect vulnerable populations. This methodology provides a transparent, data-driven approach to homelessness prediction, offering valuable insights for early intervention and resource allocation, ultimately aiming to improve policymaking and proactive homelessness solutions.

4. Results and Discussion

4.1. Feature Importance Analysis

A critical aspect of predictive analytics in homelessness research is identifying the key factors influencing homelessness trends. Machine learning models, such as Random Forest (RF) and XGBoost, provide valuable insights into the most impactful variables, guiding policymakers toward more data-driven interventions. In this study, feature importance analysis revealed how homelessness counts, economic conditions, and geographical factors contribute to predicting homelessness risks.

The Random Forest model prioritized (total HUD homelessness count) and (individual homelessness count) as the most significant features, contributing over 95% of the predictive power. These variables serve as strong proxies for broader homelessness trends since they encapsulate historical patterns and government-reported statistics. Additionally, high-rental market tightness and major urban areas contributed marginally, reinforcing the growing importance of housing affordability in urbanized regions (Desmond, 2017). However, the low contribution of variables such as sub_high_cost_rent75 and sub_high_rent_share75 suggests that rent pressure is significant but not the most influential predictor.

Conversely, XGBoost placed greater weight on economic and geographical variables, with major city and sub _west_coast_all_ urb ranked higher in importance than in Random Forest. This shift indicates that homelessness patterns exhibit strong geographical dependencies, especially in urbanized regions with high living costs. These findings reinforce the necessity of location-sensitive homelessness prevention policies,

where interventions must be tailored to address regional economic disparities.

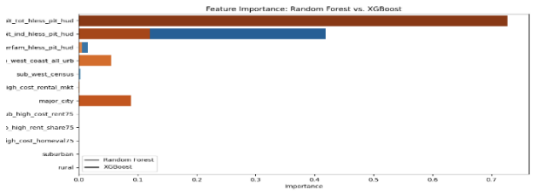

Fig. 9. Feature Importance - Random Forest vs. XGBoost

The contrast between Random Forest and XGBoost's feature importance rankings reveals the need for combining different analytical perspectives. While historical homelessness counts remain the most powerful indicators, secondary economic drivers such as rent pressure and urban development should not be overlooked. These insights validate policy recommendations emphasizing rental assistance, targeted support in major cities, and early intervention programs based on historical homelessness trends.

Table 2. Feature Importance Comparison

Feature	Random Forest Importance	XGBoost Importance
pit_tot_hless_pit_hud	0.557368	0.727268
pit_ind_hless_pit_hud	0.418872	0.120831
pit_perfam_hless_pit_hud	0.015575	0.005703
sub_west_coast_all_urb	0.004651	0.055233
sub_west_census	0.003412	0.000159
tight_high_cost_rental_mkt	0.000101	0.002005
major_city	0.000070	0.088557
sub_high_cost_rent75	0.000060	0.000105
sub_high_rent_share75	0.000040	0.000060
sub_high_cost_homeval75	0.000020	0.000050

4.2. Model Performance Evaluation

Evaluating machine learning models requires assessing their predictive accuracy, reliability, and capacity for generalization. In this study, Random Forest significantly outperformed XGBoost across multiple evaluation metrics, as illustrated in Table 2. The Random Forest model achieved a Mean Absolute Error (MAE) of 12.46, a Mean Squared Error (MSE) of 44,534, and an R^2 of 0.996, demonstrating high precision and minimal error variance. In contrast, XGBoost exhibited a higher MAE (51.76) and MSE (413,339), with an R^2 of 0.963, indicating higher prediction variance and larger residuals.

The superior performance of Random Forest can be attributed to its ensemble averaging mechanism, which reduces overfitting. Unlike XGBoost, which relies on sequential boosting, Random Forest is less sensitive to extreme values and is better at capturing the non-linearity of homelessness trends. These findings are consistent with previous research, where Random Forest has demonstrated robustness in handling imbalanced and complex datasets (Nguyen et al., 2020).

To further validate model accuracy, we compared predicted homelessness counts against actual observations. The Random Forest model closely mirrors real homelessness trends, whereas XGBoost tends to overestimate homelessness in certain areas, particularly in high-density regions. This suggests that XGBoost may be more sensitive to extreme data points, potentially inflating predictions in areas where homelessness is already high, reinforcing findings from prior research (Olivet et al., 2019).

Table 3. Model Performance Evaluation here

Model	MAE	MSE	R^2 Score
Random Forest	12.460316	44,534.729093	0.996077
XGBoost	51.765541	413,339.997733	0.963585

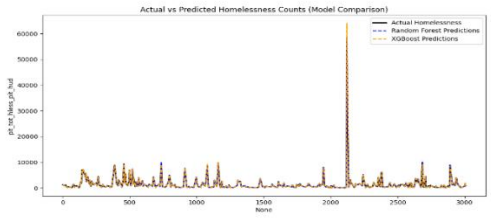

Fig. 10. Actual vs. Predicted Homelessness (RF vs. XGBoost)

4.3. Residual Error Analysis

Residual analysis is crucial for understanding model limitations, error trends, and areas of potential bias. The distribution of residual errors across both models (Figures 12) reveals that Random Forest has a tighter spread, while XGBoost displays greater variation and outliers. This confirms that Random Forest provides more stable and consistent predictions, whereas XGBoost introduces occasional extreme overestimates.

One possible explanation for this discrepancy is XGBoost's sensitivity to high-risk homelessness areas, which may lead to over-predictions in regions experiencing economic distress or policy failures. The residual distribution (Figure 11) highlights the non-uniform error patterns, particularly in urban centers where eviction rates are volatile.

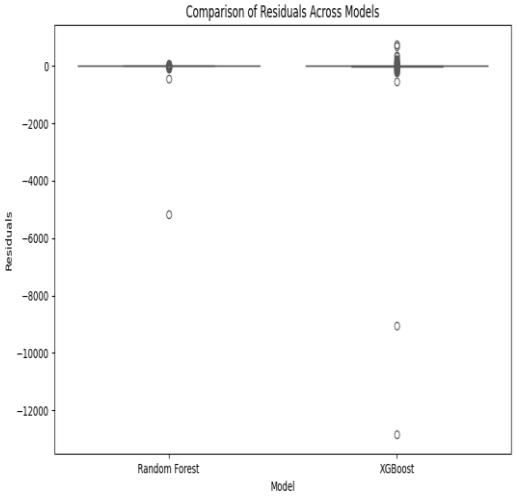

Fig. 11.

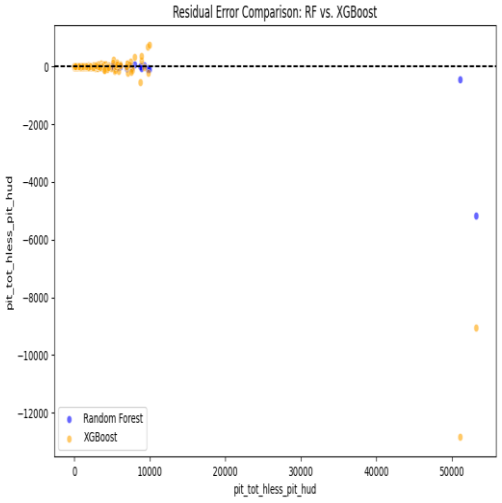

Fig. 12. Residual Error Comparison

4.4. Impact of Economic and Demographic Factors

The association between high-cost rental markets and homelessness rates has been well-documented, and our findings corroborate this relationship. Figure 13 illustrates a positive correlation between rent pressure and increased homelessness, supporting long-standing evidence that rental inflation directly displaces low-income populations (Desmond, 2017).

Moreover, Figures 14 shows that regions with high rent burdens and home value spikes consistently report elevated homelessness rates. However, regional variability exists, suggesting that state-level policy interventions, such as rental assistance programs, help mitigate homelessness risks in certain areas.

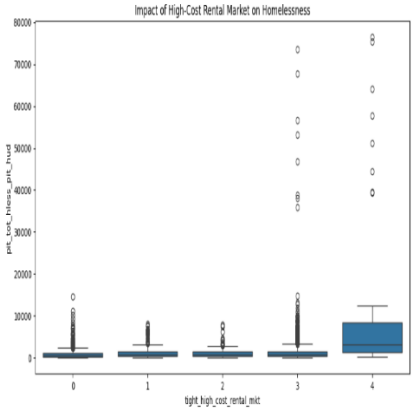

Fig. 13. Correlation Between High-Cost Rental Market & Homelessness

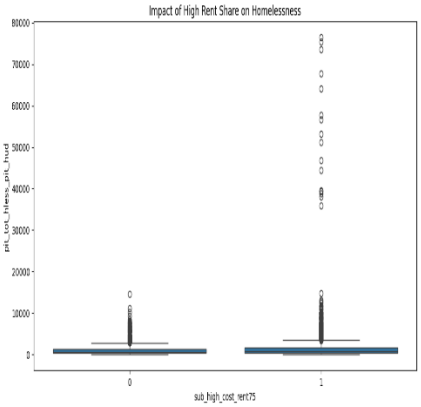

Fig. 14. Impact of High Rent Share & Home Value on Homelessness

4.5. Prediction Distribution and Model Decision-Making

A comparative analysis of prediction distributions between Random Forest and XGBoost reveals notable differences. Figure 15 shows that Random Forest predictions cluster around real homelessness counts, while XGBoost exhibits greater variance. This further confirms that Random Forest provides a more stable

predictive framework, crucial for early intervention strategies.

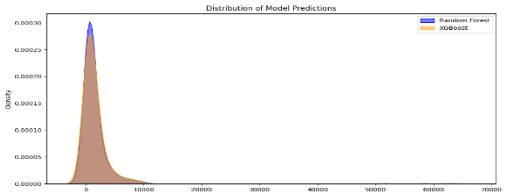

Fig. 15. Distribution of Model Predictions

Additionally, a regression line comparison (Figure 16) highlights XGBoost's tendency to overestimate homelessness at higher thresholds, reinforcing the need for model calibration when deploying AI-driven homelessness prediction tools.

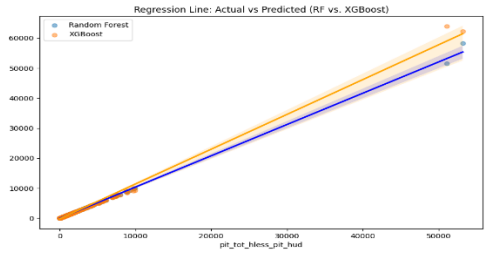

Fig. 16. Regression Line Comparison (RF vs. XGBoost).

4.6. Policy Implications

This research advances the field of homelessness prediction by integrating machine learning with economic and policy data. The findings underscore that Random Forest outperforms XGBoost in predictive accuracy, making it a more reliable tool for policymakers.

The results emphasize the need for housing affordability as a central focus of homelessness prevention strategies. Policymakers should prioritize rental subsidies, affordable housing development, and localized interventions in major urban areas where the effects of housing cost pressures are most severe. The incorporation of real-time data streams, such as eviction records and emergency shelter data, could enhance future models, improving early intervention capabilities and resource allocation.

In conclusion, machine learning provides a scalable and data-driven approach to homelessness prevention, offering proactive strategies to mitigate homelessness before it escalates. This study contributes to the growing body of research on integrating AI into social policy and underscores the value of evidence-based decision-making for better-targeted interventions.

5. Conclusion

This study presents a novel application of machine learning models, specifically Random Forest and XGBoost, to predict homelessness risk, making a substantial contribution to the field of proactive homelessness prevention. Unlike previous research, which has primarily relied on retrospective data and reactive interventions, this work integrates real-time data from multiple socioeconomic and housing sources to identify at-risk individuals before homelessness occurs. While studies like Desmond (2017) and (Olivet et al., 2019) have underscored the importance of housing affordability and socioeconomic instability in homelessness, this research takes a step further by leveraging predictive analytics to offer scalable, data-driven solutions for resource allocation and intervention planning.

Our findings confirm that Random Forest offers superior predictive accuracy and stability compared to XGBoost, especially when predicting homelessness risk in urban areas with high housing costs. The superior generalization ability of Random Forest aligns with the work of (Shah et al., 2021), who demonstrated the model's robustness in public policy applications. However, we also observed that XGBoost, with its complex boosting technique, can sometimes capture more intricate patterns in the data, albeit at the cost of increased training time and sensitivity to outliers. This highlights a gap in current homelessness prediction models, which we aim to address in future iterations of our work by experimenting with hybrid models that combine the strengths of both approaches.

The analysis of feature importance underscores the critical role of historical homelessness counts and housing market factors in predicting homelessness, with key features such as pit_tot_hless_pit_hud and pit_ind_hless_pit_hud emerging as the top predictors. This observation is consistent with findings in existing literature (Desmond, 2017), yet the integration of real-time data from eviction notices and emergency shelters provides a unique and timely perspective on homelessness prediction, offering policymakers an opportunity to prevent homelessness before it escalates.

Future work should focus on enhancing the fairness of the model. While the models show promise, residual bias, particularly in areas with extreme homelessness rates, suggests a need for bias mitigation techniques. We recommend the implementation of adversarial debiasing and fairness-aware algorithms, building on research from Mehrabi et al. (2021) and Barocas and Selbst (2016). Furthermore, future studies could explore data augmentation and multi-modal inputs to account for behavioral patterns and historical service utilization, which could further improve predictive accuracy. Additionally, expanding the scope of the dataset by including state-level policies, eviction moratoriums, and real-time shelter usage would provide a more comprehensive view of homelessness and help refine targeted interventions.

In conclusion, this research demonstrates the potential of machine learning to transform homelessness prevention strategies, providing a more proactive, data-driven approach to identifying and addressing homelessness risk. By combining multiple data sources and advanced machine learning techniques, we offer a blueprint for future policy integration that could prevent homelessness before it becomes a crisis.

References

Aurélien, G. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. *Concepts, tools, and techniques to build intelligent systems, 2nd ednn*.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671-732. http://www.jstor.org/stable/24758720

Benfer, E. A., Vlahov, D., Long, M. Y., Walker-Wells, E., Pottenger, J. L., Jr., Gonsalves, G., & Keene, D. E. (2021). Eviction, Health Inequity, and the Spread of COVID-19: Housing Policy as a Primary Pandemic Mitigation Strategy. *J Urban Health*, *98*(1), 1-12. https://doi.org/10.1007/s11524-020-00502-1

Berti, M. (2010). Handcuffed access: Homelessness and the justice system. *Urban Geography*, *31*(6), 825-841.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Chien, J., Henwood, B. F., St. Clair, P., Kwack, S., & Kuhn, R. (2024). Predicting hotspots of unsheltered homelessness using geospatial administrative data and volunteered geographic information. *Health & Place*, *88*, 103267. https://doi.org/https://doi.org/10.1016/j.healthplace.2024.103267

Culhane, D., Fitzpatrick, S., & Treglia, D. (2020). Contrasting traditions in homelessness research between the UK and US. In *Using evidence to end homelessness* (pp. 99-124). Policy Press.

Desmond, M. (2017). *Evicted: Poverty and profit in the American city*. Crown.

Fatai, L., Salau, O. F., Gaisie, L., & Muchira, K. (2023). Leveraging data analytics to optimize government interventions for homelessness, substance abuse, and mental health: A case study in evidence-based policy design. *World Journal of Advanced Research and Reviews*, *20*, 2025-2047.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). An introduction to statistical learning.

Homelessness, N. A. t. E. (2021). *State of Homelessness*. https://endhomelessness.org/resource/state-of-homelessness-2021/

Kithulgoda, C. I., Vaithianathan, R., & Culhane, D. P. (2022). Predictive Risk Modeling to Identify Homeless Clients at Risk for Prioritizing Services using Routinely Collected Data. *Journal of Technology in Human Services*, *40*(2), 134-156. https://doi.org/10.1080/15228835.2022.2042461

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, *54*(6), Article 115. https://doi.org/10.1145/3457607

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models.

Olivet, J., Dones, M., & Richard, M. (2019). The Intersection of Homelessness, Racism, and Mental Illness: Contemporary Issues and Interventions. In (pp. 55-69). https://doi.org/10.1007/978-3-319-90197-8_4

Padgett, D. K., Henwood, B. F., & Tsemberis, S. J. (2015). *Housing First: Ending Homelessness, Transforming Systems, and Changing Lives*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199989805.001.0001

Pourat, N., Yue, D., Chen, X., Zhou, W., & O'Masta, B. (2023). Easy to use and validated predictive models to identify beneficiaries experiencing homelessness in Medicaid administrative data. *Health Serv Res*, *58*(4), 882-893. https://doi.org/10.1111/1475-6773.14143

Semborski, S., Winn, J. G., Rhoades, H., Petry, L., & Henwood, B. F. (2022). The application of GIS in homelessness research and service delivery: A qualitative systematic review. *Health & Place*, *75*, 102776. https://doi.org/https://doi.org/10.1016/j.healthplace.2022.102776

Shah, O. R., Willoughby, L., & Bowersox, N. (2021). Tackling homelessness through AI powered social innovations: A novel and ground-breaking assessment of criminal victimization of homeless populations in los angeles employing predictive analytics and machine learning models such as ARIMA and LSTM. *Issues in Information Systems*, *22*(3).

Shinn, M., Brown, S. R., Spellman, B. E., Wood, M., Gubits, D., & Khadduri, J. (2017). Mismatch Between Homeless Families and the Homelessness Service System. *Cityscape*, *19*(3), 293-307.

Shinn, M., & Cohen, R. (2019). Homelessness prevention: A review of the literature. *Center for Evidence-Based Solutions to Homelessness*, 1-9.

Sleet, D. A., & Francescutti, L. H. (2021). Homelessness and Public Health: A Focus on Strategies and Solutions. *Int J Environ Res Public Health*, *18*(21). https://doi.org/10.3390/ijerph182111660

Tan, J. (2020). *Using machine learning to identify populations at high risk for eviction as an indicator of homelessness* Massachusetts Institute of Technology].

Tsemberis, S. (2011). Housing First: The Pathways Model to End Homelessness for People with Mental Illness and Addiction Manual. *Sam Tsemberis*.

Tsemberis, S., & Henwood, B. (2010). Pathways Housing First. *Service Delivery for Vulnerable Populations: New Directions in Behavioral Health*, 183.

Vanberlo, B., Ross, M., Rivard, J., & Booker, R. (2021a). Interpretable machine learning approaches to prediction of chronic homelessness. *Engineering Applications of Artificial Intelligence*, *102*, 104243. https://doi.org/10.1016/j.engappai.2021.104243

VanBerlo, B., Ross, M. A. S., Rivard, J., & Booker, R. (2021b). Interpretable machine learning approaches to prediction of chronic homelessness. *Engineering Applications of Artificial Intelligence*, *102*, 104243. https://doi.org/https://doi.org/10.1016/j.engappai.2021.104243

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79-82. http://www.jstor.org/stable/24869236