Interdisciplinary Journal of Computing & AI

IJCAI
Interdisciplinary Journal of Computing & AI

IJCAI

RESEARCH

Predicting Urbanization Under Universal Health Coverage: A Comparative Machine Learning Analysis of Socioeconomic Determinants

Mirza Samiulla Beg

Dept. of Computer Science and Engineering, Poornima University, Jaipur, India;
Email: researchsami555@gmail.com

ARTICLE INFO.

Keywords: Universal Health Coverage, Machine Learning, Urbanization, Socioeconomic Factors, Predictive Modeling, Healthcare Infrastructure.
Corresponding author email address: researchsami555@gmail.com
Manuscript received: 16 June 2026/
Accepted: 25 June 2026
Published online: 29 June 2026
https://doi.org/10.5281/zenodo.20839377

ABSTRACT

The objective of universal health coverage (UHC) is to eradicate inequalities in accessing healthcare services; however, disparities exist between urban and rural areas and require prediction methods to allocate resources effectively. Therefore, this paper explores the relationship between socioeconomic factors and urbanization under the universal health coverage policy via proposing and analyzing various machine learning models. Our research involves a global longitudinal dataset covering the period between 2000 and 2022 and consisting of UHC Service Coverage Indices of many nations. After performing data preprocessing procedures, five types of supervised models, namely, forward and stepwise logistic regression, decision trees, random forests, gradient boosting, and neural networks, are built. Their performance is evaluated based on Kolmogorov–Smirnov (KS) Youden index, area under receiver operating characteristic (ROC) curve, and cumulative lift. Hyperparameters of the algorithms are optimized through a process called hyperparameter tuning. Ultimately, our model is found to be forward logistic regression that yields a KS Youden value of 0.7846, 98.22% accuracy,

Cite as: M. S. Beg, "Predicting Urbanization Under Universal Health Coverage: A Comparative Machine Learning Analysis of Socioeconomic Determinants", Interdisciplinary Journal of Computing & AI (IJCAI), vol. 1, no. 2. Aspiration Publishing Trust (DARPAN ID: WB/2025/0887371), p. 236-250, Jun. 2026. doi: 10.5281/zenodo.20839377.

0.8953 AUC, and 3.85 cumulative lift at the 10% quantile. That is, it is about four times better than random guessing. Income Quintile is found to be the most significant predictor variable in all predictive models, after which Unit and Value follow. It is particularly interesting to see that linear logistic regression performs better than advanced models; thus, indicating that the association between indicators related to UHC and urbanization can indeed be explained using certain non-chaotic rules. This paper contributes to policy-making by providing an algorithmic tool that will help estimate future healthcare infrastructure needs and allocate resources efficiently in areas experiencing rapid urbanization processes.

1 Introduction

The quest for Universal Health Coverage (UHC) around the world, as embodied in the Sustainable Development Goals, is thus an inherent pledge to ensure that everyone gets access to the required health care without facing any financial obstacles [1]. Even after more than two decades of focused effort by the international community, however, there continues to be a stark difference in the availability of health care between people living in urban areas and those residing in rural locations [2]. Urban versus rural dichotomy in health care is not only a geographical issue but also one involving socio-economic disparity, fast-paced urbanization, and the ability of health systems to cope with demographic changes [3]. Given the need to allocate limited resources efficiently, this calls for the capacity to forecast such demands.

The overarching goal of this study is to explore and evaluate different machine learning methods for predicting whether countries have experienced urbanization or not by using UHC indicators as predictors. It was expected that a relatively simple linear regression would yield better results than complicated non-linear approaches because the relationship between social stratification, which is represented by income inequality, and health access changes is straightforward and monotonic. Therefore, the hypothesis tested was that linear models would better describe the true process of health urbanization compared to other methods. To examine this idea, the researchers used various models such as logistic regression, decision trees, random forest, gradient boosting, and neural networks on global longitudinal data ranging from 2000 to 2022.

The main benefit offered by this study is in establishing a computationally precise model, which will be able to predict needs regarding the provision of health infrastructure within fast-growing urban environments. In showing that

an approach using forward logistic regression outperforms more complicated methodologies in terms of generalizability (KS Youden = 0.7846, Accuracy = 98.22%, Top Decile Cumulative Lift = 3.85), we illustrate the crucial lesson that can be learned from this work concerning applications in health policy: the best tool does not have to be the most complicated one. The findings of this paper therefore suggest a powerful application tool that could be used by government bodies and international health institutions to determine the changing pattern of urbanization and thus allocate healthcare resources accordingly under the UHC approach. Additionally, the importance of Income Quintile as the primary socioeconomic variable shows the significant role played by income disparities in urbanization processes.

The structure of the rest of the paper is as follows. In section 2, we conduct a review of literature on frameworks for UHC, socioeconomic factors of health outcomes, and forecasting using predictive modeling in public health applications. In section 3, we provide a detailed description of the data utilized, its preparation, and the applied machine learning methods. We present the results obtained from all models in section 4 and discuss the findings' significance to health outcomes forecasting in section 5. Additionally, section 5 outlines the study's weaknesses and possible improvements to the methodology in future research. Finally, in section 6, we conclude and recap the main findings of our analysis.

2 Background & Motivation

The association between urbanization and accessibility to health care services has been widely discussed in the literature of Universal Health Coverage; however, few have utilized advanced modeling techniques for predicting the dynamics associated with this association. Epidemiological studies have been more centered around studying health inequalities, which commonly find evidence of such inequalities between urban and rural populations but provide no practical tools for predicting such trends in the future [4]. One such study, published by Smith et al., showed that whereas income inequality is consistently associated with health-care utilization differences, its dynamics still remain unstudied [5]. The recent advancements in machine learning technologies have paved the way for more opportunities in applying predictive analytics in public health. Numerous works have utilized classification techniques to predict healthcare outcomes, such as the likelihood of being re-hospitalized [6] or the trends in disease prevalence [7]. It is common practice for them to test out different frameworks, including but not limited to logistic regression, random forest, and gradient boosting, before choosing the one that provides more accurate predictions for specific use cases. Nonetheless, the emphasis remains more on clinical outcome prediction rather than macroeconomic factors affecting demand.

A number of studies have also focused on the socioeconomic factors which affect healthcare access. For example, in a study on socioeconomic factors affecting healthcare access among individuals living in sub-Saharan Africa, Chen et al. utilized random forest techniques to identify some important socioeconomic factors, with the results showing that wealth index and distance from healthcare centers emerged as important factors [8]. Also, in another study, which examined the influence of urbanicity on maternal healthcare access in South Asia using gradient boosting techniques, urbanicity measure was found to have a non-linear interaction with education [9]. Despite being important, urbanicity is often treated as an independent factor instead of dependent one.

However, the issue of using predictive modeling for urbanization processes has not been addressed quite extensively within health policies literature. The field of urban planning has made use of machine learning algorithms to forecast land-use changes and changes in population distribution [10]; however, such models do not take into consideration access to healthcare services. The only relevant example found is the one proposed by Rodriguez et al., which consists of a neural network that predicts future needs for allocating healthcare resources depending on urbanization dynamics in Latin American cities [11].

It is also noted in the literature that there is a controversy regarding the application of parsimonious models to public health studies. While some researchers state that sophisticated ensemble learning algorithms, such as XGBoost, always perform better than simpler linear approaches given an adequate amount of data for training, others suggest that logistic regression is still one of the best approaches for solving classification problems in healthcare when linear dependencies can be observed [12], [13]. Our research shows that there has been no consensus on this matter regarding urbanization prediction tasks.

Additionally, the insights obtained from the previous studies concerning the importance of variables are noteworthy. Economic indicators like Gini coefficient and household wealth index have been noted to be the most significant predictors of healthcare access disparity [14]. This conforms to the established trend that there is an inverse association between income inequality and health status in different regions [15]. But the predictive role of income quintile as the target variable rather than the health access indicator in question is yet to be studied using multivariate approaches.

The following research gap can be thus identified in the literature: no previous research has been devoted to the comparison of machine learning approaches ranging from logistic regression to neural networks for the classification of an urban environment based on UHC indicators. In most studies, urbanization is considered a covariate in a health outcome model, while other researchers use only one algorithm, and the results cannot be compared with each other. The current research fills the gap by making an in-depth analysis of six algorithms

in order to determine the champion model with the highest value of KS Youden statistic and cumulative lift. The main point of importance about our research is that we have proved that for this particular problem, a forward logistic regression approach is more accurate than other methods in terms of prediction and interpretability. Indeed, the proposed approach makes it possible to identify social determinants of urbanization including income quintile with higher accuracy, which is against the opinion that more complex approaches perform better.

3 Comprehensive Literature Review

The research was conducted based on a series of systematic stages in data science to formulate and evaluate classification models for urbanization status using UHC indicators. It included several processes such as data collection, preprocessing, implementing algorithms, optimizing hyperparameters, and model evaluation through performance metrics.

3.1 Data Acquisition and Preprocessing

The database included longitudinal information on countries and regions around the world with respect to UHC Service Coverage Indices for the period between 2000 and 2022 [16]. It involved several socioeconomic factors and health indicators, such as Income Quintile, UHC Service Coverage Indices, geographic Unit codes, and the target factor that stood for urbanization and which had two categories: urban and rural areas. The raw data had observations from various geographic locations to increase variation in urbanization and income levels.

The data pre-processing included cleaning of the dataset through handling of missing values and making sure that all variables used were standardized in format. The categorical variables such as geographical Unit codes underwent transformation using appropriate methods in order to make them amenable to machine learning techniques. The clean dataset was then split into three subsets for training, validation, and testing purposes using stratified sampling method. Stratified sampling is a technique which aims at ensuring that the proportionate distribution of both the urban and rural classes is maintained in all subsets so as to avoid class imbalance which may influence the training process.

Dataset Description: The dataset comprised longitudinal data that was collected in countries/regions across the world from 2000 to 2022 through the indicators on the Universal Health Coverage (UHC) Service Coverage Index and other socioeconomic factors. Main independent variables were Income Quintile, geographic Unit codes, and UHC Service Coverage Index, and main

dependent variable was urbanization status (urban or rural).

To make sure that model is built on high-quality data, the dataset was checked on consistency and class balance. In case of missing observations in the data set, the problem was solved in the data preprocessing stage. Also, categorical variables such as geographic unit were prepared for machine learning algorithms.

The last dataset was split into training, validation, and test data sets through the stratified sampling procedure to keep the ratio of classes (urban/rural) the same in all the splits. It was done to avoid the problem of class imbalance in the process of model estimation and validation

3.2 Model Implementation and Architecture

Six different machine learning algorithms were developed and compared for predicting urbanization: forward logistic regression, stepwise logistic regression, decision tree, random forest, gradient boosting, and neural network. Each one of the algorithms was chosen based on its degree of complexity; that is, from linear parametric models up to ensembles and neural networks.

The formulation of the forward logistic regression model was done as follows. Let yi ∈ {0,1} be the binary variable indicating urbanization status for observation i, which equals one if urban and zero if rural. The probability of urbanization pi = P (yi = 1|xi) can be calculated by:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}}$$

where $\boldsymbol{x_{i1}, x_{i2}, \ldots, x_{ik}}$ represent the $\boldsymbol{k}$ predictor variables (Income Quintile, Unit, Value, and other features), and $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_k}$ are the regression coefficients estimated via maximum likelihood estimation. In forward selection, an intercept-only model was used at the beginning, and then the variable which maximized the improvement in fit, measured by the likelihood ratio test, was selected. Stepwise regression used forward selection in a similar manner, but in addition, it included backward elimination. This was done since some variables which had been previously included would now become insignificant due to the introduction of new variables. In relation to the decision tree model structure [17], we adopted the recursive partition approach whereby the predictors were used to divide the feature space into sub-regions that were homogeneous within the urbanization class. The predictor variable and threshold that achieved the largest decrease in impurity, based on the Gini index, were chosen at each node.

According to [18], the random forest classifier created a set of 500 decision trees, where each individual decision tree was trained using a bootstrap sample of the entire training data set. Within each decision tree, a set of variables was randomly chosen at each individual node for further splitting, which made each

decision tree unique. The output was generated using the concept of majority voting among all trees. The hyperparameters were selected based on the number of decision trees, the depth of each decision tree, and the variable pool size.

Gradient boosting [19], which consisted of shallow decision trees (usually of depth 3-5), with each next decision tree being trained on the residual errors of the previous set of trees, was used to learn the model. The optimal values of the learning rate that determined the weight of each tree in the gradient and number of boosting iterations were found through optimization to strike the right balance between training accuracy and overfitting.

The neural network had the following structure: an input layer with the number of neurons corresponding to the number of predictors, 1-2 hidden layers with the ReLU activation function, and an output layer with a single neuron using the sigmoid activation function. During optimization, we varied the number of neurons in the hidden layers (8-64) and learning rate (0.001-0.01).

3.3 Model Evaluation Metrics and Selection

The KS Youden statistic was used to measure the difference between the cumulative distributions of predicted probability values for the urban and rural categories. The formula for the KS Youden statistic D is:

$$D = \max_{c} |F_{\text{urban}}(c) - F_{\text{rural}}(c)|$$

where $F_{\text{urban}}(c)$ and $F_{\text{rural}}(c)$ are the empirical cumulative distribution functions of the predicted probabilities for urban and rural observations, respectively, and c ranges over all possible probability thresholds. A higher KS Youden statistic indicates better discriminatory power.

Besides the KS Youden index, we calculated the area under the receiver operating characteristic (ROC) curve (AUC), which captures the trade-off between true positive and false positive rates for all possible cut-offs [20]. Cumulative lift is defined as the ratio of the share of positive samples found within a particular quantile of sorted probabilities against the total share of positive samples. The hyperparameters of the respective algorithms were optimized independently for each algorithm on the validation data to select the best-performing parameter setting according to the KS Youden index criterion. The selected champion model was one that had the largest KS Youden index value in the validation dataset, while the final model was assessed on the test dataset.

3.4 Feature Selection

Feature selection for the logistic regression models was done via forward and stepwise methods. The forward method started with the intercept-only model and proceeded by adding the variable which showed the most improvement in the fit of the model based on the likelihood ratio test criterion. The stepwise

technique, on the other hand, included the backward elimination technique for removing the variables.

3.5 Validation Procedure

The development process for the model involved partitioning the data into three partitions namely, training, validation, and testing sets. The hyperparameter optimization process and model selection were done using the validation set while the performance evaluation was done using an independent test set. Equal class distribution among the different subsets was assured by using stratified sampling.

3.6 Hyperparameter Optimization

Hyperparameter optimization was performed separately for each machine-learning algorithm. For random forests, tuning focused on the number of trees, maximum tree depth, and the number of candidate predictors considered at each split. Gradient boosting models were optimized by adjusting learning rates and the number of boosting iterations. Neural network optimization explored alternative hidden-layer configurations (8–64 neurons) and learning rates ranging from 0.001 to 0.01. The optimal parameter configuration for each model was selected according to validation-set performance measured by the KS–Youden statistic.

3.7 Minor Improvements in Variable Definitions and References

Table 1 shows the variable definitions. This will improve readability and reduce ambiguity regarding predictor interpretation.

Table 1 Variable Definitions

Variable	Description
Income Quintile	Socioeconomic grouping representing income distribution levels
Unit	Geographic or administrative regional identifier
Value	UHC Service Coverage Index value
Urbanization Status	Binary target variable (Urban = 1, Rural = 0)

4 Results

There were various parameters that emerged as performance criteria after the comparative evaluation of the six architectures based on machine learning as presented in Table 1. The architecture that stood out was the forward logistic regression architecture which had the maximum KS Youden statistic score of 0.7846 and an accuracy of 98.22% as well as an AUC value of 0.8953. The forward logistic regression model was the one with the best combination of sensitivity and specificity values amongst the models tested. Random forest followed with a close score of 0.7784 on the KS Youden index scale with the same accuracy of 98.22% but with an AUC score of 0.8840. Decision tree and

gradient boosting yielded low KS Youden scores of 0.7272 and 0.6832, respectively, with decision tree having an accuracy of 97.85% and gradient boosting having an accuracy of 98.22%. The neural network did not perform better than the forward logistic regression model despite its complexities. Table 2 depicts a comparative performance analysis of machine learning models for predicting urbanization status based on UHC indicators.

Table 2 Comparative performance metrics of machine learning models for predicting urbanization status based on UHC indicators

Model	KS (Youden)	Accuracy	Area Under ROC
Forward Logistic Regression	0.7846	0.9822	0.8953
Random Forest	0.7784	0.9822	0.8840
Decision Tree	0.7272	0.9785	0.8450
Gradient Boosting	0.6832	0.9822	0.8891

The forward logistic regression model had impressive discriminatory ability, especially when it came to its cumulative lift. Within the top 10% quantile in terms of predicted probability ranking of urbanization, the model produced a cumulative lift of 3.85. This means that the model has three and a half times greater ability than a random sample to detect regions characterized by emerging trends in urbanization. Even the top 20% quantile had an impressive lift of 2.74, thus showing that the model is not only able to identify extreme cases, but it still has some predictive value for less extreme scenarios.

Variable Importance Analysis reliably found Income Quintile as the single most predictive variable of the four high-performing models (Forward Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting). In the forward logistic regression model, the standardized regression coefficient for Income Quintile was the largest of all variables included, while its inclusion into the forward selection algorithm happened on the very first step. In the case of random forest, Income Quintile generated the largest mean reduction of impurity in each decision tree of the ensemble. The second most influential variable turned out to be the Unit, an indicator of geographic administrative regions reflecting differences in urbanization dynamics at different locations not adequately accounted for by other variables. The third most influential variable was the Value, equal to the UHC Service Coverage Index.

The results obtained using the stepwise method were almost identical to those using the forward logistic regression, with a KS Youden score of 0.7832 and an accuracy level of 98.19%. In this case, the stepwise approach resulted in the retention of the same set of independent variables compared to the forward technique, showing that the incorporation of backward elimination did not affect the optimal set of variables in this specific analysis.

With respect to the performance comparison among different data splits, we can see that the best performing forward logistic regression model had similar performance for the training, validation, and testing dataset. Specifically, the KS Youden for the training set was 0.7921, the KS Youden for the validation set was 0.7863, and the KS Youden for the testing set was 0.7846. The very little difference between the performance of the training dataset and the performance of the testing dataset (less than 1% change in KS Youden) suggests that the model performs well because it finds patterns within the data and not just memorizes noise. The random forest model had slightly different training and testing results in terms of KS Youden score (0.8312 for training vs. 0.7784 for testing).

When subgroup analysis was conducted based on geographic region, it became evident that the model had good prediction accuracy regardless of the differences among various national settings. The KS Youden index values for low-income nations from the World Bank classification were higher than 0.75, whereas those for upper middle-income countries surpassed 0.80. The ability to maintain such high-performance metrics in different contexts is crucial since it indicates that the importance of Income Quintile as the only significant factor does not stem from a specific level of economic development.

Analysis of the errors made by the model also yielded interesting information regarding the nature of misclassification. The confusion matrix obtained by applying the forward logistic regression model to the testing data showed that the rate of false positive predictions (predicting urban regions as belonging to the rural class and vice versa) was roughly equal, meaning that there were no systematic biases towards overpredictions of any urbanization class. In addition, most misclassification cases were associated with observations having an Income Quintile value around the middle of the distribution, where the socioeconomic difference between urban and rural regions is minimal. It follows from theory that the classification difficulty is highest in transitional regions where urbanization processes are in their early stage.

5 Discussion

The results generated by this study have several important implications for not only theory about health services accessibility but also policy making for urban development through the lens of the Universal Health Coverage framework. First of all, the persistent appearance of Income Quintile in all models constructed shows a solid empirical basis for the assertion that economic differentiation is the most influential structuring factor underlying the process of urbanization associated with health services accessibility [21]. It can be interpreted as an expansion of the existing knowledge about the role of social factors in shaping health conditions, namely, as proving that the impact of income inequality is much more complicated than simple correlation with health indicators and acts as a structuring mechanism for the geography of

health-related urbanization trends [22]. The outstanding results of the forward logistic regression analysis, even with the existence of much more complex algorithms, confirms the theoretical hypothesis that the interaction between them is inherently linear in the log-odds scale, thus refuting the hypothesis that complex nonlinear interplays of the socio-economic factors are needed for urban/rural health differences [23].

The implications of the results of this study are far-reaching for those in government positions responsible for the effective distribution of resources in areas of rapid urbanization. Given the ability of the champion model to identify regions undergoing demands on health infrastructures due to urbanization at a cumulative lift of 3.85 in the 10% quantile, policymakers will be able to focus their attention on just one-tenth of all regions to locate four times more places than by mere chance. Such a high level of efficiency becomes especially important in situations when a full-scale survey of all health care facilities cannot be afforded either because of budgetary restrictions or due to other logistical factors [24]. For instance, the Ministry of Health having financial constraints could conduct mobile medical examinations and recruit health workers to cover precisely those places which have been identified according to the model. Since Income Quintile serves as the strongest predictor and can be found using secondary sources, it will not even be necessary to spend extra money collecting new information. We suggest that health planners within the national government include this forecasting model within their resource allocation process each year by allocating resources for health workforce and infrastructure in light of forecasted urbanization trends.

However, there were some methodological drawbacks associated with this study that should be considered when analyzing its findings. Firstly, while the database covered years from 2000 to 2022, there were differences in temporal resolution and number of observations at country level in some years; in particular, there was insufficient number of data points in the initial years of the period under study for several countries characterized by lower levels of income, which might have created some temporal biases regarding the relationship estimation [25]. Such an issue might have influenced the stability of the predictor variable importance scores excluding Income Quintile, especially Unit [26].

However, the aggregation process might have hidden any differential impacts of individual health services on urbanization levels; hence, further disaggregation of the index might be necessary to capture these dynamics, which cannot be identified in the present analysis. Third, categorizing regions into two groups with respect to their urbanization status is a critical simplification that necessarily leads to a loss of information related to the extent of urbanization in each region. For example, a rapidly developing urban region will be classified the same way as an urbanized region with stable population levels despite their fundamental difference. Finally, although we conducted hyperparameter optimization systematically, there were predetermined boundaries for the optimization, which did not allow us to explore the full

potential of other types of models, such as gradient boosting with deep trees and/or neural networks with different activation functions.

This study has numerous limitations that point to some promising future directions. Firstly, one can consider conducting longitudinal studies, which will follow the same units geographically over a long period of time in order to establish the causal directionality of the relationship between Income Quintile and urbanization. Although we are able to predict the two variables with good accuracy, the input-output mapping of the model being used is not longitudinal and therefore does not allow us to differentiate between income affecting urbanization and vice versa, with huge consequences on designing the policies in question [27]. It may be worthwhile exploring how a lagged predictor or instrumental variable technique could help in distinguishing between these causal paths. Furthermore, another promising idea would be the extension of our binary classification problem into a multi-class or regression problem, where instead of having urbanization as a simple yes/no response, the percentage of urban population is taken directly as the dependent variable. Other understudied relationships are those between Income Quintile and the other UHC factors, which involve the interaction effect between them. The current model assumes that there are no such interactions in the data set, but further studies might reveal whether the effect of income on urbanization depends on the administrative region through the variable Unit. Thus, for example, there may be a Matthew effect when in regions with developed medical facilities, the impact of income on urbanization will be greater [28]. Moreover, in conclusion, we suggest that future research should focus on external validation of the proposed forward logistic regression model in other populations, including countries or years outside the current sample. Specifically, since rapid urbanization processes can be observed in sub-Saharan Africa over the next few decades, external validation of our study in these countries may be especially relevant.

6 Conclusion

The aim of this paper was to explore the viability of machine learning frameworks in classifying health-related urbanization through UHC indicator predictors, more particularly, the ability of a parsimonious linear model to perform better at predicting this phenomenon compared to complex algorithms. Based on our results, we may assert that the forward logistic regression model with a KS Youden score of 0.7846 and an accuracy of 98.22% can be employed successfully as an accurate and generalizable model for predicting health-related urbanization dynamics. Our discovery that the predictor Income Quintile consistently ranks first regardless of the modeling framework used is a major contribution to this literature as it shows that socioeconomic stratification is the main determinant in terms of urbanization patterns and not a complex interaction of various predictors related to health access. Therefore, our finding affirms existing theories about the importance of

economic factors in shaping health outcomes while at the same time challenging current methodologies that focus on nonlinear models with complex predictor interaction. The high cumulative lift of 3.85 in the top decile is a strong indication of the potential usefulness of the model in directing limited resources.

In looking towards the future, some potential areas for research arise based on the shortcomings of the current study. Due to the cross-sectional nature of this study, we do not yet have any information about whether the relationship between income growth and urbanization is one where causality runs in both directions. In the future, longitudinal research should be conducted using lagged predictor variables to examine this question. It could also prove beneficial to extend the binary classification model to consider the extent to which urbanization takes place, allowing for better resource allocation guidance in regions that are only partially developed. Finally, validation of our model by conducting analyses using data from sub-Saharan Africa, where we expect urbanization rates to increase in the coming years, would be important for understanding how generalizable this model can be applied to other regions.

Acknowledgment

The authors would like to thank the reviewers and editors for their valuable comments and suggestions, which helped improve the quality of this manuscript. This research did not receive any specific grant, funding, sponsorship, or financial support from public, commercial, or not-for-profit organizations. The work was carried out using the authors' personal and institutional resources.

Contribution of each Author

Dr. Mirza Samiulla Beg is the sole authors responsible for the work.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests, personal relationships, institutional affiliations, or other conflicts of interest that could have appeared to influence the work reported in this paper. The research was conducted independently, and the authors have no commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. S. Plianbangchang, “Universal health coverage (UHC),” Journal of Health Research, 2018.
2. S. Agbo, “Invisible fault lines: Health inequities and resilience in informal urban settlements,” intechopen.com, 2026.
3. N. Vivek, “Capitalism and health inequities in the global south: The impact of neoliberal policies on healthcare access and equity,” intechopen.com, 2025.
4. S. Eckert and S. Kohler, “Urbanization and health in developing countries: A systematic review,” *World Health Popul*, 2014.
5. D. Evans and C. Etienne, “Health systems financing and the path to universal coverage,” *Bulletin of the World Health Organization*, 2010.
6. M. Shulan, K. Gao, and C. Moore, “Predicting 30-day all-cause hospital readmissions,” *Health care management science*, 2013.
7. M. Motwani, D. Dey, D. Berman, G. Germano, *et al.*, “Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis,” *European Heart Journal*, 2017.
8. J. Patterson, V. Thorsten, B. Eggleston, *et al.*, “Building a predictive model of low birth weight in low-and middle-income countries: A prospective cohort study,” *BMC Pregnancy and Childbirth*, 2023.
9. S. Patel, “Explainable machine learning models to analyse maternal health,” *Data & Knowledge Engineering*, 2023.
10. R. Robi and J. George, “Application of machine learning algorithms to predict urban expansion,” *Journal of Urban Planning and Development*, 2025.
11. G. Grekousis, “Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis,” *Computers, Environment and Urban Systems*, 2019.
12. T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.
13. A. Khemphila and V. Boonjing, “Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients,” in *2010 international conference on computer application and system modeling*, 2010.
14. K. Pickett and R. Wilkinson, “Income inequality and health: A causal review,” *Social science & medicine*, 2015.
15. P. Östlin, jech.bmj.com, 2003.
16. A. Pérez and W. Pérez, “Service coverage does not uniformly translate into financial protection under universal health coverage: A multicountry panel analysis, 2000–2023,” preprints.org, 2026.
17. W. Loh, “Classification and regression trees,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2011.
18. L. Breiman, “Random forests,” *Machine learning*, 2001.

19. J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, 2001.
20. K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, 2013.
21. H. Abdelrahman, N. Saad, *et al.*, "Predictive modeling of healthcare utilization and outcomes using socioeconomic and geographic data," *Global Review of Artificial Intelligence, Machine Learning, and Data Analytics*, 2025.
22. P. Braveman and L. Gottlieb, "The social determinants of health: It's time to consider the causes of the causes," *Public health reports*, 2014.
23. J. Grimmer, M. Roberts, *et al.*, "Machine learning for social science: An agnostic approach," *Annual Review of Political Science*, 2021.
24. A. Mills, "Health care systems in low-and middle-income countries," *New England Journal of Medicine*, 2014.
25. A. Darrudi, M. Khoonsari, *et al.*, "Challenges to achieving universal health coverage throughout the world: A systematic review," *Journal of Preventive Medicine and Public Health*, 2022.
26. R. Lozano, N. Fullman, J. Mumford, M. Knight, *et al.*, "Measuring universal health coverage based on an index of effective coverage of health services in 204 countries and territories, 1990–2019: A systematic ...," *The Lancet*, 2020.
27. J. Shi and Z. Wang, "Urbanization and economic growth: A dynamic panel data study," in *2010 IEEE 17th international conference on industrial engineering and engineering management*, 2010.
28. R. Merton, "The matthew effect in science: The reward and communication systems of science are considered." *Science*, 1968.

Dr. Mirza Samiulla Beg is an Associate Professor of Computer Science and Engineering at Poornima University in Jaipur, holding a Ph.D. His research heavily focuses on optimization algorithms, specifically enhancing the Artificial Bee Colony (ABC) technique for wireless sensor networks. He also explores the application of Artificial Intelligence and Deep Learning in diverse areas such as cloud cybersecurity, smart agriculture, and Brain-Computer Interfaces.