International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Poverty Prediction Using Machine Learning with Socio-Economic and Spatial Indicators

P. Nagashree Satya Haritha[1], Y. Siri Teja[2], S. Pranathi[3], Shaik Malik[4], K. Chaitanya[5]

CSE(AIML), Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, INDIA

Email: nagashree0627@gmail.com, yasasiri24@gmail.com, pranathisaragundla@gmail.com, shaikmalik1608@gmail.com, chaitanyakethavath@gmail.com

Abstract:

The issue of poverty is a multidimensional and complicated problem on the global level, which needs precise and scalable approaches to the identification and treatment. Conventional survey-based methodologies tend to be constrained on basis of high cost, low frequency of updates, and limited coverage. In this paper, a framework is suggested that employs machine learning to predict poverty, combining socioeconomic, demographic, and spatial measures, such as a nightlight intensity measured by satellites as a proxy of economic activity in a region. An extensive dataset that included household-based characteristics including income, education, ownership of assets, access to basic services and geographic accessibility was used. There were four supervised learning models, namely, Random Forest, Extreme Gradient Boosting (XGBoost), Logistic Regression, and Support Vector Machine (SVM), which were applied and measured on the basis of accuracy, precision, recall, and F1-score. Experimentation shows that the predictive capacity of ensemble models is much higher than traditional models, and XGBoost is the most predictive as it can describe complex non-linear relationships and the interactions of features. The importance of spatial indicators, especially distance to urban centers and intensity of night lights, as part of features and their importance and ablation analysis also underline the importance of the selected features in increasing predictions. The proposed system will be a scalable, data-driven way to classify poverty and will offer actionable insights into specific policy interventions to help Sustainable Development Goal 1 (No Poverty) be achieved.

Keywords— Poverty Prediction, Machine learning, XGBoost, Random Forest, Socio-economic Indicators, Nightlight Intensity, Classification, Data-driven Policy.

I. INTRODUCTION

Poverty is a multifaceted phenomenon which is not only about the absence of income. It also involves inadequate education, medical care, housing and primitive infrastructure. It would be essential to identify the economically vulnerable groups within a short time and precision to execute the welfare programs and allocate resources equally. Nevertheless, standard ways of measuring poverty, including household surveys and census statistics, are problematic. They are usually costly to perform, not readily scaled and updated very rarely. As more and more data and computing power became available, machine learning has proven to be a useful instrument in studying complex social and economic systems. In comparison to the conventional statistical procedures, machine learning algorithms are capable of identifying patterns and relationships in large data automatically. This results in more accurate and scaled predictions [5], [6].

The power of predictive models has increased recently because of the use of non-standard data sources such as satellite images and the intensity of nightlights. Such data as nightlight data is known to be one of the most effective indicators of economic activity, development of infrastructure, and the degree of urbanization. The objective of this study is to develop a powerful machine learning model to forecast poverty using a combination of conventional socio-economic variables with spatial proxy variables [11],[12]. The major contributions of this work are: Production of a multi-source integrated data. Comparison between different machine learning models. Examining the significance of the various characteristics in ascertaining poverty status. Demonstrating the ability of spatial indicators to improve the accuracy of prediction [11].

II. PROPOSED SYSTEM

A. System Overview

The proposed system is a data-based system that emphasizes on the forecast of poverty with the use of heterogeneous sources of data and the use of machine learning methods in classifying data as either poor or non-poor [1],[2]. The system also has the capability to process structured data, and also the spatial proxy data to increase the accuracy of the system.

B. Model Design

There were four models developed, which were the different types of learning paradigms. The developed models are as follows:

TABLE 1 MODEL TYPES

Model	Type	Strength
Random Forest	Ensemble	Handles non-linearity
XGBoost	Boosting	High accuracy
Logistic Regression	Linear	Baseline
SVM	Complex	Boundaries

C. System Architecture

The proposed system will have a flexible system architecture. The proposed system will have six stages in the workflow. The phases are linked with each other in a way that enables flow and transformation of data in a seamless manner.

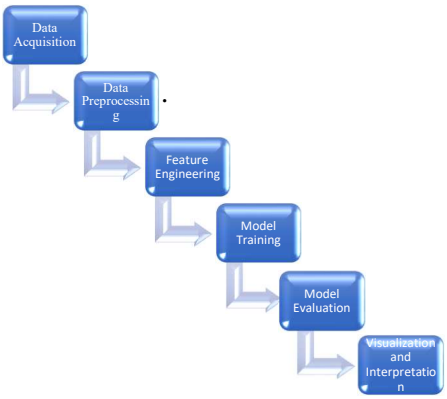

Figure 1 System architecture

D. System Modules

1) Data Collection Module : Data will be gathered through this module in various sources. These sources are socioeconomic data and indicators of development. The data gathered will consist of the demographic data, the economic data, and the basic services. Also, data will be collected using satellite imagery data on nightlight intensity. This information will be useful in the integration of spatial differences in economic activities. The inclusion of various sources of heterogeneous data will guarantee a holistic coverage of the issues of poverty. Dataset Description:

- Demographic Characteristics: Age, sex, household, education.
- Economic Characteristics: Income, employment, ownership of assets.
- Living Conditions: Type of housing, electricity, water,sanitation.
- Spatial Indicators: Intensity of nightlight, urban-rural.
- Social Indicators: Access to health care, nutrition, and education.

2) Data Preprocessing Module : Preprocessing of data is a significant feature of the proposed system. The information gathered may be noisy and may have inconsistencies in the information. The data gathered may have such problems as missing data, repeated values and discrepancies. This might affect the performance of the model. In order to overcome these problems, the following measures will be undertaken: The median will be used to represent missing values in numeric features and mode in categorical features. Elimination of duplicate values will be done. Categorical values will be coded. Normalization will be performed for numerical values. The procedure will guarantee appropriate pre-data processing.

3) Feature Engineering Module : The module will make sure there is proper feature engineering. The model will be enhanced with the help of feature engineering. A aggregate poverty risk measure will be developed. The score will entail several socioeconomic characteristics such as income, education, and ownership of assets. Moreover, the model includes such spatial characteristics as the nightlight intensity to capture the regional economic features. The analysis of correlation is conducted to verify the relationships existing among the variables and eliminate redundant attributes. The composite risk score was established in the following way: Integrates several socio-economic data. Captures non-linear relationships Enhances the model sensitivity. Further, the following was accomplished: Implicit feature interaction. Integration of local and global spatial and economic relations.

4) Model Training Module : The model training module entails four supervised learning algorithms: the Random Forest, the XGBoost, the Logistic Regression and the Support Vector Machine (SVM) algorithms on the socio-economic data with complex, non-linear relationships.

Random Forest: Random Forest is an ensemble algorithm that integrates several decision trees with the help of a voting mechanism. It successfully models non-linear relationships between variables including income, education, and living conditions and overfitting is minimized by averaging. It was highly classified and offered feature importance (e.g., income and education). Nonetheless, it had slightly more false negatives than boosting models.

$$y = \sum_{t=1}^{T} h_t(x) \quad (1)$$

Where $h_t(x)$ represents predictions from individual decision trees and T is the number of trees.

XGBoost: XGBoost is a modern gradient boosting algorithm, which constructs a model one after another, with each model correcting the errors of the previous model. It is very efficient in modeling the non-linear feature interaction and makes accurate predictions. XGBoost was the most reliable model used in this study since it had the best accuracy and false negativity.

$$L = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \quad (2)$$

The objective function (2) consists of: Loss function l (prediction error) and Regularization term Ω to prevent overfitting.

Logistic Regression: Logistic Regression is a linear classifier and it is used to predict the probability of a binary result. Although easy and helpful to be used as a baseline model, it does not have good performance with large data and is not effective to represent non-linear relationships. Its linear character restricts the performance in socio-economic data that is multidimensional.

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (3)$$

This equation (3) represents the probability of a household being classified as poor (class 1).

Support Vector Machine (SVM): SVM recognizes an optimal hyperplane to distinguish between classes and is able to deal with non-linear and high-dimensional data with the use of kernel functions. It is also applicable to complex data and can be generalized to multi-class problems. But, it is computationally costly and parameter tuning sensitive.

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ subject to } y_i(w \cdot x_i + b) \geq 1 \quad (4)$$

This formulation (4) defines the objective of maximizing the margin between classes while minimizing classification error.

5) Evaluation Module : The model is evaluated using various evaluation parameters. The accuracy of the model is checked through different parameters: Accuracy, Precision, Recall, and F1 Score. Visual representation of the model classification performance is done by a confusion matrix. False-negative cases are considered particularly since such cases may be overlooked when running a welfare program. False positives are not of much concern in this context because they can receive benefits.

6) Visualization Module : The methods of data interpretation and model performance are done through visualization. The visual representation of the variable distribution is carried out with the help of histograms. Visual representation of variable relationships is done by Correlation heatmaps. Interpretation of feature importance is done using feature importance graphs.

E. Major Benefits of Proposed System

Multi-dimensional data handling, Incorporation of spatial intelligence, Scalability, and Improves the precision of the system using ensemble learning.

III. EXPERIMENTAL SETUP

This experiment was written in Python, and run in a cloudbased Google Colab, which is a scalable and flexible platform to run machine learning workflows. Stratified sampling was used to preprocess and partition the dataset into training (80 percent) and testing (20 percent) sets in order to provide the same proportion of poor and non-poor households in each subset, and this way evaluate the models unbiasedly. Four machine learning models were applied and compared within similar conditions: Random Forest, XGBoost, Logistic Regression, and Support Vector Machine. Random Forest model was a model that used 100 decision trees and was used to capture complex patterns by using ensemble learning. XGBoost was used with default parameters, which allows effective optimization of structured data. Logistic Regression was set to a maximum of 1000 iterations to make sure that it would converge. The Support Vector machine used its default kernel to represent non-linear decision boundaries.

IV. RESULTS

The findings achieved on the intended poverty prediction system are discussed with regard to exploratory data analysis and model evaluation. The data set that is utilized in the current study is a collection of 12,000 records with each record describing a single household and a combination of socio-

economic, demographic, and spatial attributes. The dataset is structured and clean, and does not contain missing values, and can be directly used in machine learning pipelines and guarantees the same and reliable model performance.

A. Feature Distribution Analysis

The visualization of distribution of all features in the dataset was done through histograms. In such graphs, the x-axis will be the range of values of a given feature and the y-axis will be the frequency or the number of households of the value ranges.

TABLE 2 FEATURE DISTRIBUTION TABLE

Feature	Type	Encoding Format	Observation
Age	Continuous	Numeric (18–80)	Uniform
gender	Categorical	0=F, 1=M	Balanced
household size	Discrete	1–9	Even spread
education level	Ordinal	0–5	Higher values rare
monthly income	Continuous	Scaled numeric	Right-skewed
employment status	Categorical	0–3	Uneven
asset ownership	Continuous	0–1	Normal
access to credit	Binary	0/1	Slight imbalance
housing type	Categorical	0–3	Even
nightlight intensity	Continuous	0–1	right-skewed

B. Correlation Heatmap Analysis

The correlation heatmap gives a detailed visualization of how all the features in the dataset are related and thus an in-depth understanding of how various socio-economic and spatial variables relate to one another and to the target variable, poverty label. Poverty is correlated with education level and monthly income negatively, which implies that education level and monthly income decrease the risk of poverty. Space features are important as well. Poverty is negatively related with nightlight intensity, whereas distance to city center has a positive relationship.

C. Confusion Matrix Analysis

The confusion matrix is a valuable instrument in determining the effectiveness of machine learning models in classifying data.

TABLE 3 RF CONFUSION MATRIX VALUES

	Predicted Non-Poor	Predicted Poor
Actual Non-Poor	1512	48
Actual Poor	114	726

TABLE 4 XGB CONFUSION MATRIX VALUES

	Predicted Non-Poor	Predicted Poor
Actual Non-Poor	1528	32
Actual Poor	92	748

D. Feature Importance Analysis

To determine the significant variables that determine poverty prediction in the proposed model, feature importance analysis was conducted. The findings reveal that the most notable feature is the distance to city center, which means that the geographical accessibility is a key factor in determining poverty. The second feature with the highest importance is household size. Education level and monthly income are also significant predictors.

E. Model Comparison Analysis

The relative analysis of the Random Forest, XGBoost, Logistic Regression, and Support Vector Machine (SVM) results show that the ensemble learning approaches are much superior to the conventional models in poverty prediction activities. XGBoost was the most accurate and produced the best F1-score because of its gradient boosting algorithm that is able to effectively capture non-linear relationships.

V. APPLICATIONS

The suggested poverty forecasting system has a lot of practical implications especially in policy making, governance and planning social welfare. It can be used in one of the main ways of applying it in the targeted welfare distribution, in which the governments can precisely target the economically vulnerable families and make sure that the subsidies, financial aid, and social benefits go to the right people.

A. Ablation Study

The ablation experiment was carried out to measure the role of the various features groups to the overall model performance. Once the economic characteristics were eliminated, model performance declined drastically. Likewise, the deletion of spatial variables led to a significant reduction in accuracy, indicating the relevance of the geographic and infrastructural variable.

VI. CONCLUSION

This paper introduces a machine learning-based solution to poverty prediction on the basis of socio-economic,

demographic and spatial variables. The analysis shows that the factors that determine poverty are interdependent as they are income, education, ownership of assets and accessibility of these assets by geographic areas. The spatial characteristics of nightlights intensity and distance to urban centers are important additions to the model. XGBoost was found to be the most effective of the considered models. This work can help to reach the Sustainable Development Goal 1 (No Poverty) as it will help policy-makers to recognize and serve vulnerable populations more efficiently.

REFERENCES

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785– 794.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.

[4] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York, NY, USA: Wiley, 2000.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009.

[7] World Bank, "Poverty and Equity Data Portal," 2022.

[8] United Nations, "Transforming our world: The 2030 Agenda for Sustainable Development," 2015.

[9] World Bank, "World Development Indicators," 2020.

[10] United Nations Development Programme, "Human Development Report," 2021.

[11] N. Jean et al., "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[12] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

