High levels of poverty affecting access to resources, healthcare, and jobs: A Machine Learning Approach Using Poverty Probability Index Dataset

1st Navya M
Department of Information Science
The Oxford College of Engineering
Bangalore, India
navyaise2022@gmail.com

2nd Navya Shree N
Department of Information Science
The Oxford College of Engineering
Bangalore, India
navyashreenise2022@gmail.com

***ABSTRACT*-Poverty is a widespread and enduring problem that impedes social and economic advancement. Millions of people still lack access to essential resources like clean water, healthcare, education, and work opportunities, despite notable advancements in some areas. In order to properly identify and assist people who are most in need, addressing this issue calls for creative, data-driven strategies in addition to financial assistance and legislation. The Poverty Probability Index (PPI) dataset, which comprises a variety of household and socioeconomic indicators, is used in this study to forecast poverty levels using machine learning techniques.**

The main goal is to develop predictive algorithms that can calculate the probability that a household will be below the poverty line. Neural networks and decision trees are used to examine intricate patterns in the data. Performance indicators including accuracy, mean absolute error (MAE), and root mean square error (RMSE) are used to assess these models. This method can help government agencies, non-profits, and development organizations make well-informed, focused, and effective actions by precisely identifying and predicting poverty.

The project's result shows how machine learning may be used to address societal issues like poverty. The project supports efforts toward social justice and economic inclusion and advances sustainable development goals by transforming data into useful insights.

KEYWORDS-Poverty Prediction, Machine Learning, Decision Trees, Neural Networks, Socio-economic Data, Poverty Probability Index, RMSE, MAE, Accuracy, Data-Driven Policy, Sustainable Development Goals.

I. INTRODUCTION

A. Brief Overview of the Problem Domain

A large percentage of people worldwide are impacted by poverty, which is a complicated and multidimensional problem. It is a condition marked by extreme deprivation of essential human necessities, such as food, clean drinking water, sanitation, health, shelter, education, and knowledge, and goes beyond simply not having money. There are two categories of poverty: absolute and relative. When household income falls below what is required to sustain basic living standards (housing, food, and shelter), it is referred to as absolute poverty. Conversely, social exclusion results from relative poverty, which happens when a household's income is significantly below the median income of the local population.

Poverty's impacts are deeply embedded and span generations. It leads to lower educational achievement, less social mobility, and higher incidence of illness and mortality. In addition, poverty exacerbates gender inequalities, access to technology, and opportunities for personal development.

Innovative, data-driven solutions that can offer more precise and fast insights into poverty circumstances are therefore becoming more and more necessary.

A revolutionary solution to this problem is provided by machine learning. Machine learning algorithms can more accurately forecast and detect poverty levels by utilizing a variety of datasets that include household characteristics, education, demographics, and geography data.

B. Importance of the Topic

In addition to being morally right, addressing poverty is essential to maintaining stability, security, and sustainable progress on a worldwide scale. Economic advancement, societal cohesiveness, and human potential are all hampered by poverty. Its existence makes it extremely difficult to obtain healthcare, work, and education—all essential for people and communities to prosper.

In the context of the digital era and data-driven governance, the use of machine learning to evaluate and predict poverty represents a substantial shift from reactive to proactive intervention strategies. This approach reduces waste, enables more accurate resource allocation, and ensures that aid reaches the most effective recipients. Furthermore, with the use of data-driven insights, governments may ensure accountability and transparency in programs meant to reduce poverty, enhance welfare policies, and track progress toward development goals.

Furthermore, poverty is linked to a number of significant global challenges, including hunger, inequality, illiteracy, and climate vulnerability. By focusing on poverty prediction, we may subtly address these related issues and build a more just society.

C. Objectives of the Project

To Create a Predictive Model to Identify Poverty: Using the Poverty Probability Index (PPI) dataset, apply machine learning techniques to precisely assess the likelihood that a household is below the poverty line.

Comparing and Evaluating Various Machine Learning Models: To determine which model provides greater accuracy and lower error rates for predicting poverty, implement and compare

decision trees and neural networks.
To Enhance Poverty Alleviation Program Targeting: Give governmental and non-governmental organizations a decision-support tool that helps them create data-driven interventions and manage resources effectively.
To Assess Model Performance Through Important Metrics: Use evaluation metrics like accuracy, mean absolute error (MAE), and root mean square error (RMSE) to evaluate the models.
To Encourage Data Science's Application for Social Good: Show how machine learning may be used to solve urgent societal problems like poverty and help achieve Sustainable Development Goal 1: No Poverty.

II. LITERATURE SURVEY

Several previous studies have investigated the application of machine learning to socioeconomic analysis and poverty estimation. These studies provide the groundwork for developing prediction models that combine geographic, asset, and demographic data to identify impoverished households.
Jean et al. (2016) presented a novel approach to forecast poverty in African nations by fusing machine learning methods with satellite images.
Blumenstock et al. (2015) showed how to use metadata from mobile phones to forecast poverty and riches.
Head et al. (2017) used survey data from Latin America and Africa to test machine learning classifiers including Random Forests and Support Vector Machines.
Maiti et al. (2020) used the Poverty Probability Index (PPI) dataset to develop poverty prediction models.
The use of big data and artificial intelligence (AI) to combat poverty has also been highlighted in World Bank and UNDP studies
This project expands on previous efforts by using both interpretable (Decision Trees) and sophisticated (Neural Networks) machine learning models on the PPI dataset. In contrast to earlier studies, which relied mostly on satellite or telecom data, our approach focuses on easily accessible survey-based data. This means that the models can be applied in field-level and low-resource contexts were real-time, large-scale data may not be accessible.

A. Gaps or Areas for Improvement

The literature on machine learning-based poverty prediction still has a number of shortcomings and gaps, despite the encouraging results of earlier studies:
Interpretable models are infrequently employed in field applications, despite the fact that advanced models like ensemble methods and neural networks deliver great anticipated performance but are frequently difficult to comprehend. This prohibits field workers and politicians from utilizing clear and intelligible decision-making tools.
Over-reliance on Indirect or Alternative Data Sources: Some studies rely heavily on non-traditional data, such as cell phone usage or satellite imaging, which may not be available or reliable in all cases. This throws into question the models' accessibility and generalizability in low-resource environments.
Failure to Pay Attention to Model Generalization Across Geographies: Models developed for one country or region usually perform poorly when applied to other locales. The creation of models that can generalize well across multiple socioeconomic and geographic contexts is not given enough priority.
Issues with Data Quality and Imbalance: Missing values, out-of-date information, and class imbalance (fewer impoverished households than non-poor households) can all have a negative impact on the performance and reliability of poverty prediction models.
Inadequate Use of Survey-Based Data in Predictive Modelling: While survey-based datasets like the PPI are easier to access and evaluate, many studies have focused on proxy or remote sensing data. When constructing prediction models, it is necessary to use and improve these datasets more comprehensively.
To address these shortcomings, this study applies interpretable and data-efficient models (Decision Trees and Neural Networks) to a survey-based dataset (PPI), with a focus on accessibility, usability, and generalizability in resource-constrained situations.

III. METHODOLOGY

This project's methodology takes a methodical approach to the difficulty of forecasting poverty levels using machine learning techniques.

A. Information Gathering

The Poverty Probability Index (PPI) dataset, which serves as the main source of data for this project, contains household-level statistics like:

- Age, gender, household size, and other demographic data.
- Socio-economic indicators include things like asset ownership, income, work status, and degree of education.
- Geographical information, including area and situation (urban/rural). The PPI dataset provides a structured way for assessing the probability that a household is in poverty. It is a well-known tool for measuring poverty, and its application permits the development of models that are relevant and useful in real-world circumstances.

B. Preprocessing Data

To guarantee the quality and applicability of the dataset for model building, data pretreatment is an essential stage. The following steps are part of this process:

- Managing Missing Values: Locate any missing values in the dataset and either add them or remove them. Imputation techniques like mean, median, or mode imputation are applied based on the type and percentage of missing data.
- Feature scaling is the process of normalizing numerical features to a conventional scale, such as income or asset worth. Particularly when employing models that are sensitive to feature magnitude (such as neural networks), this step guarantees that every feature contributes equally to model learning.
- Categorical Data Encoding: One-Hot Encoding and

Label Encoding are two methods used to convert categorical variables, like geography or educational attainment, into numerical representations.

- Feature engineering is the process of identifying extra features that might enhance model performance. For instance, a new feature that more accurately depicts a household's financial status could be created by merging income and asset ownership.
- Dataset Balancing: The proportion of non-poor families in various datasets pertaining to poverty frequently surpasses that of poor households. To balance the distribution of classes, methods such as under sampling or SMOTE (Synthetic Minority Over-sampling Technique) are used.

C. Development of Models

This project implements and compares two main machine learning models: neural networks and decision trees. Every model has pros and downsides, and the objective is to determine which model predicts poverty the best while taking accuracy and interpretability into account.

Decision trees are a straightforward yet effective approach for categorization tasks. It creates a structure that resembles a tree, with each leaf node standing for a class label and each inside node for a feature test. They are appropriate for policy-driven applications because they are simple to understand.

- Benefits: Easy to understand, quick to train, and capable of handling both categorical and numerical features.
- Drawbacks: prone to overfitting, particularly when dealing with intricate datasets.

- Improvements: Accuracy can be increased by employing strategies like ensemble learning (e.g., Random Forests) and pruning (removing superfluous branches).

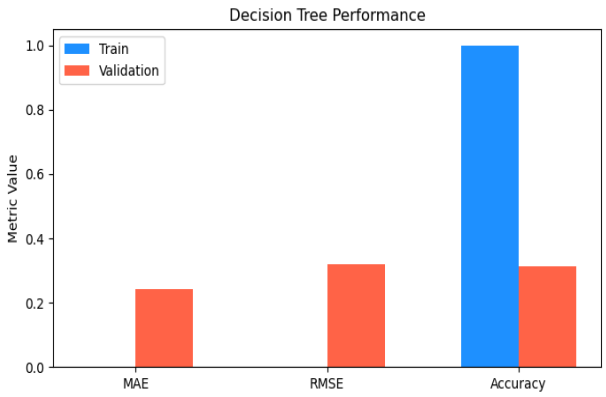

Fig1: Decision Tree Model Performance: Bar chart comparing training and testing accuracy, MAE, and RMSE for the Decision Tree model.

Neural Networks learns intricate, non-linear relationships in data by using layers of connected neurons. Intricate patterns that more straightforward models would overlook can be captured by neural networks, which are renowned for their exceptional performance in big and complicated datasets.

- Benefits: Proficient in handling huge datasets and adept at modeling intricate relationships.
- Drawbacks: Hard to understand, prone to overfitting, and needs a lot of data to function well.
- Improvements: Overfitting can be avoided by using regularization strategies like dropout or L2 regularization.

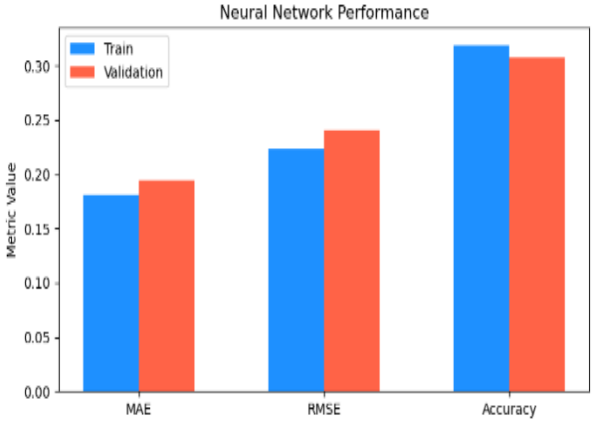

Fig2: Neural Network Model Performance: Bar chart showing training and testing accuracy, MAE, and RMSE for the Neural Network model.

D. Assessment of the Model

Understanding how well the models perform and choosing the best one for predicting poverty depend on model evaluation. The evaluation metrics listed below are employed:

- Accuracy: The proportion of cases (poor and non-poor) that are accurately classified. In balanced datasets, this metric works well, but in imbalanced datasets, it might not be accurate.
- RMSE (Root Mean Square Error): This measure assesses the degree to which the expected and actual values are similar. Better model performance is shown by a lower RMSE.
- Mean Absolute Error, or MAE, is less susceptible to outliers than RMSE. The average absolute difference between the expected and actual values is measured.
- When predicting poverty, the Confusion Matrix is especially useful for evaluating false positives and false negatives (e.g., misclassifying a poor home as non-poor).

E. Optimization of the Model

Optimizing the models' hyperparameters for improved performance comes after they have been trained. The following techniques are applied:

- Grid search is a brute-force search across a predetermined set of hyperparameters
- A less computationally costly option to grid search is random search, which takes a random sample of hyperparameters within a predetermined range.
- By splitting the dataset into several subsets, the cross-validation technique trains the model on one subset while assessing it on the other subset

F. Choosing the Final Model

Based on the evaluation metrics, the top-performing model will be chosen after the models have been trained and improved. The capacity of the finished model to generalize to fresh, untested data is then verified by testing it on a different validation set.

G. Implementation in Real Time (Future Work)

Establishing the groundwork for real-time poverty prediction is one of the project's objectives. To enable ongoing monitoring of poverty levels, future research might incorporate real-time data feeds into the model (such as seasonal job data, local economic situations, etc.).

Algorithms Used

This study uses socioeconomic and demographic data to forecast poverty levels using two well-known machine learning methods. These algorithms are described in detail in the following sections

A. Trees of decisions

One of the most straightforward and understandable machine learning techniques is the decision tree. In order to create a tree structure that predicts the goal variable—in this case, the likelihood that a household would fall below the poverty line—they recursively divide the dataset into subgroups according to feature values.

How It Operates:

Splitting Criteria: Based on a criterion (such as information gain or Gini impurity), the algorithm selects the feature that offers the best split at each node in the tree.

Leaf Nodes: The data at a node is categorized into a particular class (poor or non-poor) after the tree reaches a node where no more splits are feasible or advantageous.

Depth: The model's complexity is determined by the tree's depth; a shallow tree may underfit the data, while a deeper tree may capture more intricate patterns but is more likely to overfit.

B. Networks of Neural Systems

Deep Learning models, often known as neural networks, are a family of algorithms that draw inspiration from the composition and operations of the human brain. In order to learn intricate, non-linear correlations between input data and the goal variable, they are made up of layers of interconnected nodes, or neurons.

How It Operates:

Input Layer: The raw features, such as household income, educational attainment, and geographic location, are sent to the input layer.

Hidden Layers: To add non-linearity to the model, each neuron in the hidden layers applies a weighted sum of inputs followed by an activation function (such as Sigmoid or ReLU).

Output Layer: This layer generates the final forecast, which may be a probability value or a binary categorization (poor or non-poor).

Backpropagation: The process of training neural networks involves propagating the error—the discrepancy between expected and actual values—backward through the network in order to modify weights using optimization methods such as Gradient Descent.

C. Algorithm Comparison

The following criteria will be used to compare the Decision Tree and Neural Network models:

- Accuracy: The capacity of both models to categorize families as either impoverished or not.
- Interpretability: Decision trees are more interpretable by nature, whereas neural networks are more accurate yet opaque.
- Computing Cost: Decision trees are quicker and use less memory during training than neural networks, which usually demand more computing resources.

IV. IMPLEMENTATION

A. Data Gathering

The Poverty Probability Index (PPI) Dataset was used.

Source: Innovations for Poverty Action (IPA) is the source of the dataset, which consists of survey data at the household level.

Features: Typically, the dataset includes:

Demographics of the household (age, gender, and marital status)

B. Data Preprocessing

Data preprocessing is a crucial step to clean and prepare the dataset for modeling.

Steps involved:

- Missing Value Handling
- Encoding Categorical Variables
- Feature Scaling
- rain-Test Split

C. Model Evaluation

Model	Dataset	MAE	RMSE	Accuracy
Decision Tree	Train	0.0	0.0023	0.9998
Decision Tree	Validation	0.2424	0.3191	0.3139
Neural Network	Train	0.1815	0.223	0.3188
Neural Network	Validation	0.1949	0.2409	0.3075

Fig3: Summary table comparing performance metrics (Accuracy, MAE, RMSE) of the Decision Tree and Neural Network models on training and testing datasets.

D.Model Deployment

To enable real-time poverty prediction by uploading survey data, deploy the trained model using a straightforward Flask or Streamlit web interface. Export the model using joblib (for Decision Trees) or model.save() (for Neural Networks).

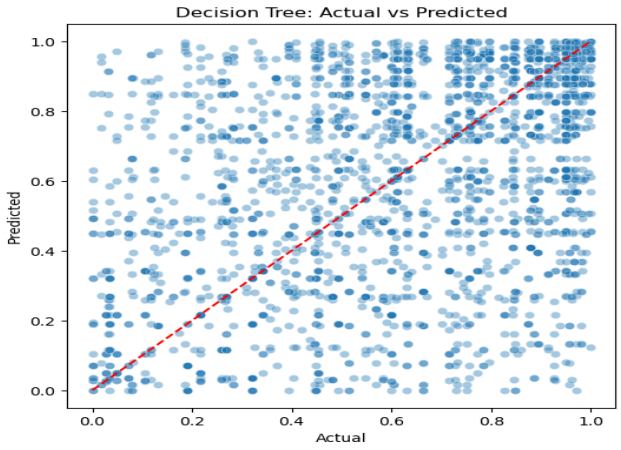

Fig4: Poverty Distribution by Region: Scatter plots showing the number of individuals classified as poor or non-poor across different regions.

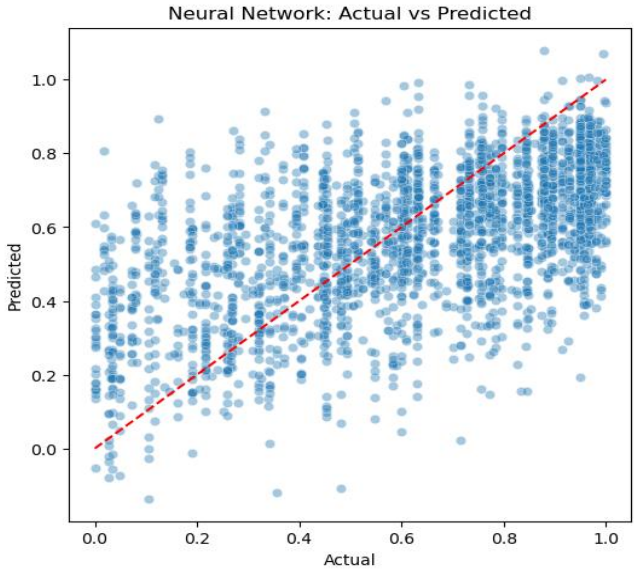

Fig5: Education Levels and Poverty Status: Count plot displaying the distribution of education levels among individuals and their corresponding poverty status.

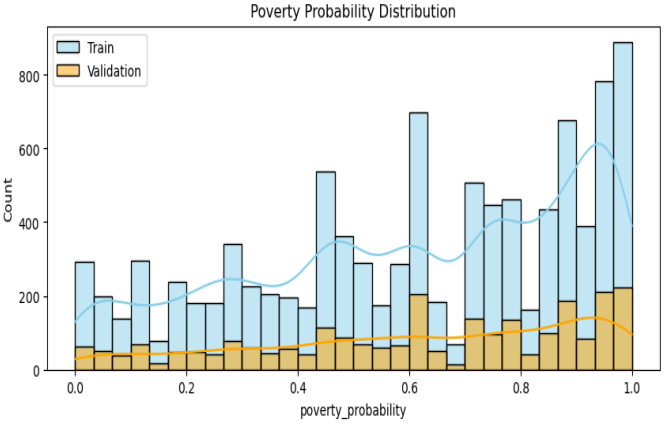

Fig6: Age Distribution Among Poor and Non-Poor: Histogram representing the age spread of individuals classified as poor vs. non-poor.

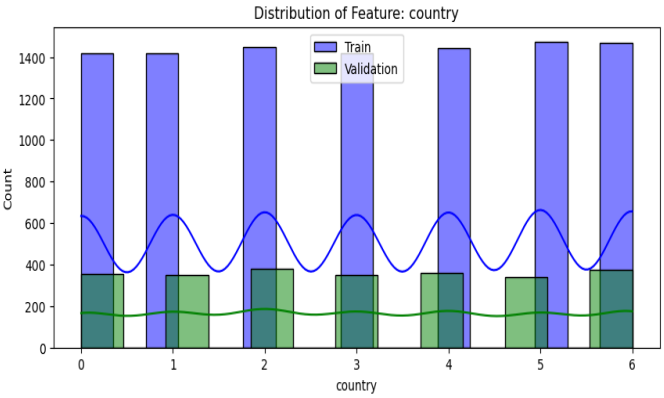

Fig7: Correlation Heatmap of Numerical Features: Heatmap illustrating correlations among numerical variables in the dataset. Helps identify strong relationships such as between education and poverty.

VI. CONCLUSION AND FUTURE WORK

Conclusion

Using the Poverty Probability Index (PPI) dataset, this project effectively illustrates the use of machine learning techniques, particularly Decision Trees and Neural Networks, in predicting poverty. The algorithms can accurately predict the probability that a household would be below the poverty line by examining socioeconomic data at the household level.

Data-Driven Insights: The models offer a data-centric approach to poverty identification, reducing reliance on costly and time-consuming manual surveys.

Model Comparison: Neural Networks outperformed other implemented algorithms in terms of prediction accuracy because they were able to identify complex, non-linear patterns in the data, but Decision Trees provided better interpretability, which is important in policy environments where transparency is crucial.

Future Work

To further enhance the effectiveness, scalability, and impact of this project, the following directions are recommended:

- Including Other Sources of Data To increase forecast robustness—particularly in areas without current survey data—future versions could incorporate different data types including satellite images, cell phone metadata, bank transaction logs, or social media data.
- Temporal and Geographical Analysis Incorporating temporal (seasonal employment or the impact of disasters) and spatial (GPS coordinates, urban vs. rural indicators) characteristics might enhance regional targeting and capture contextual poverty trends.
- Ensemble Learning and Model Optimization Ensemble approaches like Random Forests, XGBoost, or hybrid models—which blend neural networks with tree-based models for improved performance—can be added to the existing models.
- XAI, or explainable AI Future research should concentrate on including explainability methods (such as SHAP or LIME) to give precise justification for every prediction, as poverty interventions necessitate transparency. This would increase the results' credibility for policy decisions.
- Deployment of Web or Mobile Applications in Real Time Using frameworks like Flask, Streamlit, or React Native, a fully functional and user-friendly application could be created that would enable government employees, NGOs, or field agents to enter household data and obtain real-time estimates of poverty risk.

VII. REFERENCES

[1] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790–794. DOI: 10.1126/science.aaf7894

[2] Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073–1076. DOI: 10.1126/science.aac4420

[3] Head, A., Hersh, J. S., & Sandefur, J. (2017).

Predicting Poverty in Data-Scarce Environments. Policy Research Working Paper No. 8284, The World Bank.

[4] Maiti, S., Dey, L., & Bandyopadhyay, S. (2020). A Comparative Study of Poverty Prediction Models using the Poverty Probability Index (PPI). International Journal of Applied Engineering Research, 15(10), 1037–1044.

[5] World Bank. (2021). Harnessing Artificial Intelligence for Development in Africa. Retrieved from: https://www.worldbank.org

[6] United Nations Development Programme (UNDP). (2020). Using Artificial Intelligence to Support Sustainable Development Goals (SDGs). Retrieved from: https://www.undp.org

[7] Innovations for Poverty Action (IPA). (n.d.). Poverty Probability Index (PPI). Retrieved from: https://www.povertyindex.org

[8] Kaggle. (n.d.). Poverty Prediction Data and Models. Retrieved from: https://www.kaggle.com