JOURNAL OF
EMERGING INVESTIGATORS

Article

Model selection and optimization for poverty prediction on household data from Cambodia

Gabrielle Wong[1]**, Ahmed Shuaibi**[2]

[1] Wycombe Abbey, High Wycombe, England

[2] Department of Quantitative and Computational Biology, Princeton University, Princeton, New Jersey

SUMMARY

Addressing global poverty requires understanding of the most poverty-stricken regions. One approach towards achieving this is through poverty prediction, a task that entails classifying poverty levels of households using available data. While machine learning (ML) has been applied in numerous fields with considerable success, its application in poverty prediction using exclusively household survey data is yet to be thoroughly explored. Household survey data offers a detailed view into the living conditions, lifestyle, and socio-economic factors affecting households. Hence, we aim explore the use of this data type in predicting poverty levels. Our study primarily focuses on three ML models: softmax classification, random forest classification, and multilayer perceptron (MLP). We chose Cambodia for this study due to its unique socio-economic landscape and as a representative of developing nations struggling with poverty. This analysis will serve as the foundation for applying this approach to other nations. The analysis was based on a dataset consisting of 15,825 household samples and 1,873 features obtained through the Demographic and Health Surveys (DHS) program in Cambodia. The study's aim was to validate the effectiveness of ML in poverty prediction using household data and identify the best performing model among the selected three. We hypothesized that the MLP, due to its advanced neural network structure, would provide superior results compared to the softmax classification and random forest models. As anticipated, the multilayer perceptron outperformed the other models, achieving an accuracy of 87% against the 81% and 80% accuracy of the random forest and softmax classification models respectively.

INTRODUCTION

Economists and researchers have sought a holistic overview of poverty levels across countries to better understand economic landscapes around the world. The projected number of people living in extreme poverty is estimated to be around 700 million in 2022 (1). Accurately determining poverty levels enables countries to ascertain the effectiveness of their poverty reduction strategies. However, predicting poverty levels remains a challenging task due to gaps in the existing information. Specifically, remote, or conflict-ridden regions often face underrepresentation in conventional survey methodologies, leading to incomplete and sometimes biased poverty data.

There are several benefits to answering the challenging task of poverty prediction. For one, policymakers monitoring poverty reduction strategies can analyze trends across the countries to decipher where government aid or intervention should be directed (2). Moreover, philanthropists can determine the areas with the most urgent demand for their charitable donations (2). Additionally, understanding the distribution of poverty at a detailed level, such as at the household level, which is the focus of this study, can better focus a country's efforts for improving the poorest locations. It's crucial to note the granularity of the predictions. In this study, our predictive models are primarily focused on the household level. This approach provides a detailed mapping of poverty, enabling policy adjustments and interventions tailored to specific localities, rather than broader, less-targeted regional measures.

The challenge of poverty prediction arises from its multidimensionality, transient nature, issues of under-reporting in surveys, and geographical diversity. Traditional survey methods often focus on income or consumption measures, missing the full spectrum of poverty. To address these issues, researchers have increasingly turned to machine learning technologies for their computational and statistical analyses. With comprehensive survey data provided by programs like the Demographic and Health Surveys (DHS), which collects information regarding 5,000 to 30,000 households every 5 years for hundreds of countries (3), we can capture a holistic view of poverty and enhance prediction accuracy.

Our machine learning approach uses a wide array of DHS data, including educational attainments and the health of family members. The versatility of this approach allows it to adapt to changing patterns in data, accommodating the transiency of poverty. While it doesn't directly solve under-reporting, it mitigates its impact by making robust predictions even in the face of incomplete data.

One specific context where this machine learning approach can be applied is Cambodia, a country with a GDP per capita less than 2,000 US Dollars and acute poverty challenges (4). Its pronounced poverty levels, further exacerbated by the COVID-19 pandemic, underscore the necessity of accurate poverty prediction models (4). The country's unique sociopolitical history and economic conditions, as well as its primarily rural and agriculturally reliant population, make it an illustrative case study. The wealth of data available from international aid programs, coupled with the government's commitment to combating poverty, allow for the real-world application and evaluation of our predictive models (5). By focusing on Cambodia as a case study, we create a custom model for its specific context, which could potentially be replicated for other regions. This approach to poverty prediction could provide meaningful insights to improve poverty alleviation strategies and policy interventions.

Survey data-based models, such as those published by Alsharkawi et al., offer significant interpretability but may sacrifice accuracy (6). In contrast, models like Engstrom et al. and Jean et al.'s, which use satellite imagery and Rolf et al.'s, which indirectly employs satellite imagery, achieve high accuracy but struggle with interpretability, thus hindering our understanding of poverty's causes (5, 7, 2). Alsharkawi et al.'s gradient boosting model also suffers from limited interpretability despite high accuracy. Our approach aimed to bridge this gap by improving interpretability in standard machine learning models like Logistic Regression and Random Forest, even at a potential accuracy cost. The objective was to better comprehend the factors influencing poverty levels, guiding more effective poverty reduction strategies.

We predicted poverty in Cambodia with economic indicators obtained from wide-scale surveys using machine learning methods. The goal of this work was to evaluate the efficacy of household survey data for the prediction of poverty levels. Furthermore, we aimed to assess which machine learning method is best suited for the classification task with this input data. Given the richness and complexity of DHS data for Cambodia, our hypothesis was that a multilayer perceptron, a form of neural network, would yield superior accuracy in predicting poverty levels compared to less complex models such as softmax classification and random forest classification. Through more effectively and efficiently classifying poverty levels with machine learning, we established a foundation for researchers to utilize large-scale computational approaches for poverty reduction moving forward.

RESULTS

We used three machine learning models: the softmax classifier, random forest classifier, and MLP classifier, on the DHS data from Cambodia. We chose these models for their varying degrees of complexity and interpretability.

We considered wealth classification as both a 5-class problem and a simplified 3-class problem. In the 3-class problem, we grouped the upper two and lower two classes together, yielding classifications for *rich*, *middle*, and *poor*. We trained each of the three machine learning models on the training set and evaluated them on the testing set, using F1 score and accuracy score metrics.

With regards to the 5-class classification task, the accuracy score of the softmax classifier was the lowest at 57% **(Figure 1)**. The best performing model according to accuracy score alone was the MLP at 74% accuracy. The F1 score for each of these models was lower than the respective accuracy scores **(Figure 1, Figure 2)**. The random forest classifier has the lowest performance with respect to the F1 score **(Figure 1, Figure 2)**. We obtained similar results for the 3-class classification task, in which the MLP performs the best with respect to the accuracy score and F1-score while the other two models perform worse.

Receiver operating characteristic (ROC) curves depict the quality of the classifications for each model on a label-by-label level. The area under the curve (AUC) calculated indicates that there are more misclassifications especially with respect to predicting the *middle* class, with AUC values

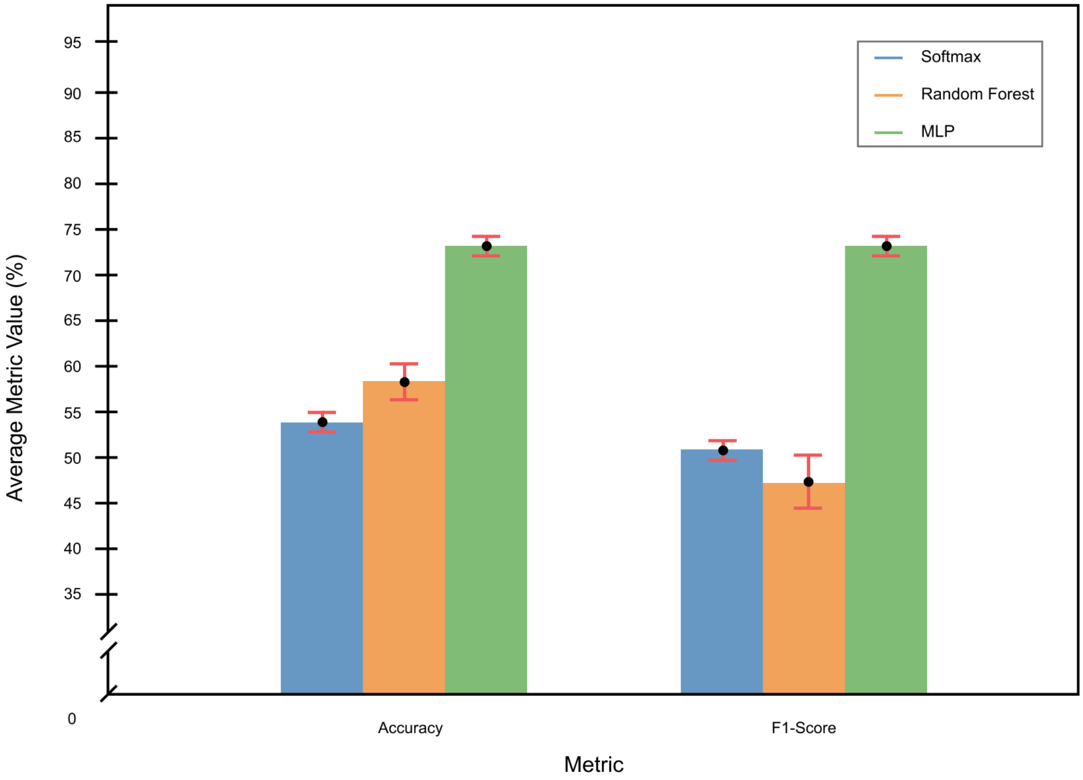


Figure 1: 5-Class classification task results for differing models. We evaluated the accuracy score and F1-score for each of the three models. The error bars indicate the standard deviation of the accuracies across the 5 model runs.

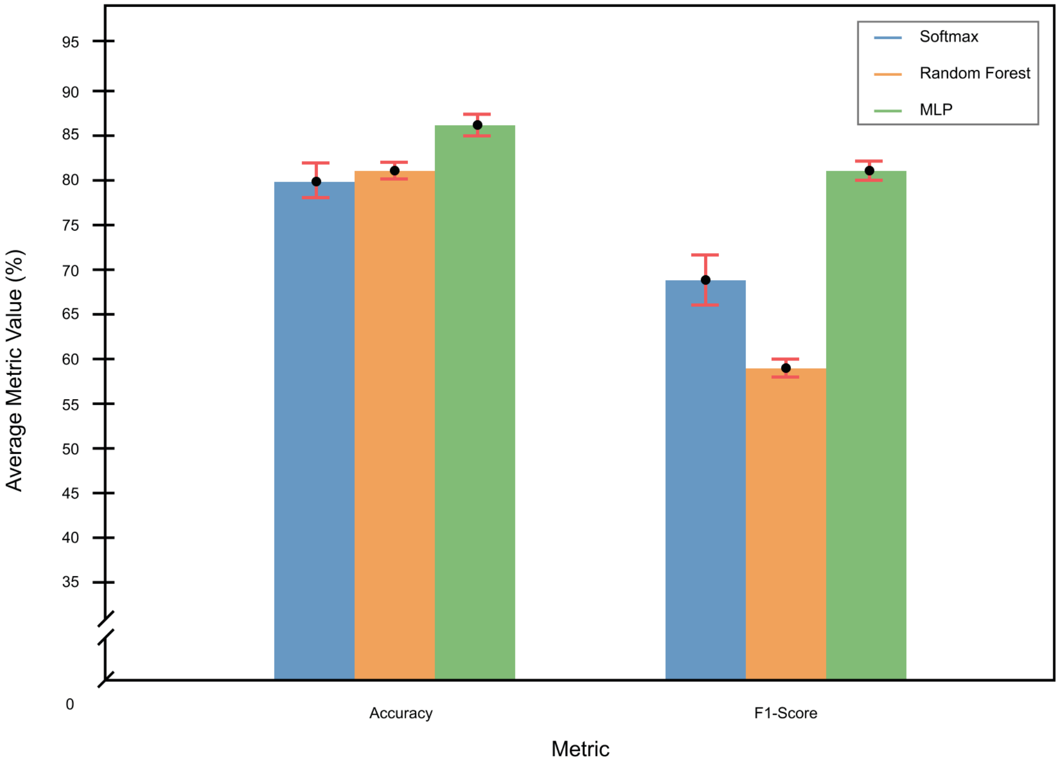


Figure 2: 3-Class classification task results for differing models. We evaluated the accuracy score and F1-score for each of the three models. The error bars indicate the standard deviation of the accuracies across the 5 model runs.

hovering around 0.85 across the different models **(Figure 3)**. In contrast, predictions of the rich or poor classes perform well at AUC values around 0.98 across the different models. As for the 5-class classification task, the predictions across each of the classes independently performed with an AUC value averaging 0.96 **(Figure 4)**. The classification accuracy was the most notable for the MLP classifier, at 87%. To highlight the most important features with regards to the classification, a plot of feature importance was generated from the random forest model. A small subset of the 40 features used compose the most significant feature values when measured with the Gini impurity values **(Figure 5)**.

The random forest model calculates Gini impurity values during training, which help identify the importance of each feature in the classification task (9). In our model, features such as electricity availability, toilet facility type, and cooking fuel type emerged as most significant in classifying wealth levels for Cambodian households in the 3-class task.

DISCUSSION

Our study aimed to develop a machine learning approach to predict household poverty levels based on survey data from Cambodia, with the intention of creating a tool to better allocate resources to those in need. The models used for this task were a softmax classifier, random forest classifier, and MLP each chosen for their varying degrees of complexity and interpretability.

In our findings, the MLP emerged as the top-performing model, demonstrating an average accuracy of 87% over five model runs. The high dimensionality and potential for nonlinear relationships between the 40 survey features seem to favor this more complex neural network model over simpler counterparts, such as the random forest and softmax classification models. These results align with existing literature that shows the proficiency of neural network models in managing high-dimensional, complex data (7).

Furthermore, our analysis of the classification model's errors highlighted a key issue: significant misclassification between the 'middle' class and the 'poor' or 'rich' classes in the 3-class task **(Figure 6)**. Similar patterns were observed in the 5-class task, with considerable errors when differentiating between adjacent classes. This observation suggests shared characteristics between these classes and potentially indicates the need for more discriminative features to improve classification accuracy, particularly for the 'middle' class.

Contextualizing our findings within the broader field, this study contributes to the evolving realm of poverty prediction using machine learning. While our focus was on Cambodia, an important point of discussion is the generalizability of our results to other countries. Cambodia was chosen due to its distinct wealth disparity, making it a suitable testbed for our models. However, it's crucial to note that socioeconomic factors and their influence on wealth disparity vary across nations. Hence, while our models could provide a baseline, feature selection and model tuning might be required when applying this approach to different countries. This stands as an area for future research.

Additionally, while our study adds to the existing knowledge of using machine learning for poverty prediction, it stands out by exploring the trade-off between model complexity and interpretability in this specific context. Studies such as (7) have highlighted the importance of this balance, but few

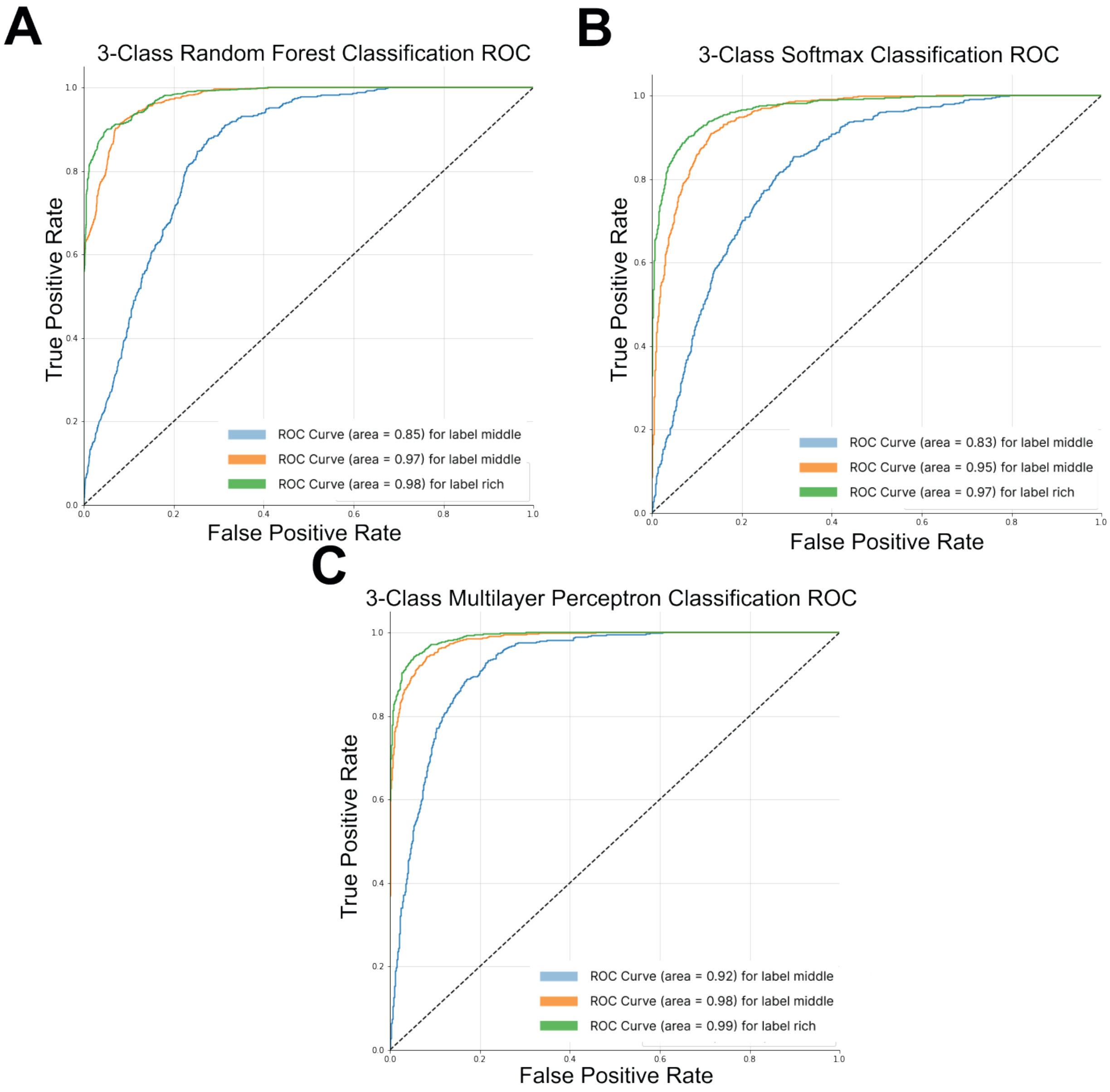


Figure 3: One versus rest ROC curves for 3-class task. ROC curves for each of the 3 different classes in the 3-class task. Each model is trained then evaluated on the testing dataset. The predicted values obtained for each class are used to generate the curves, in which each curve in a plot corresponds to the one-vs-rest task of classifying a label from the rest. Dotted lines are shown along the center of each plot that corresponds to and AUC of 0.5 as a baseline for reference. (A) Plot of ROC curves for each of the 3 classes for the random forest model. (B) Plot of ROC curves for each of the 3 classes for the softmax classification model. (C) Plot of ROC curves for each of the 3 classes for the MLP model.

have applied it in a similar context, making our study a novel contribution.

Taken together, our study demonstrated the effective use of machine learning models in predicting poverty levels and highlighted the challenges and potential improvements required for more accurate classifications. The findings offer insights not only on the technological front but also potentially inform policy decisions by identifying key areas of focus. The implications of this research are far-reaching, as the accurate prediction of poverty levels can aid in the precise allocation of resources, thus contributing to efforts towards reducing poverty globally.

Ultimately, we built a neural network model to accurately predict poverty levels using household survey data from Cambodia. It was able to predict poverty levels in a 3-class classification task with 87% average accuracy. These results indicate that taking real-world indicators and using them to predict poverty levels is not only feasible but easily scalable with high accuracy. It is more practical for the task of poverty prediction to use a subset of the survey features since they do not all contribute equally to indicating the poverty level. Through identifying the most discriminative features, future efforts for the collection of data that will be most beneficial for predicting wealth levels was greatly aided. An improvement in experimentation that can be made in subsequent rounds of analysis is the inclusion of these most significant features

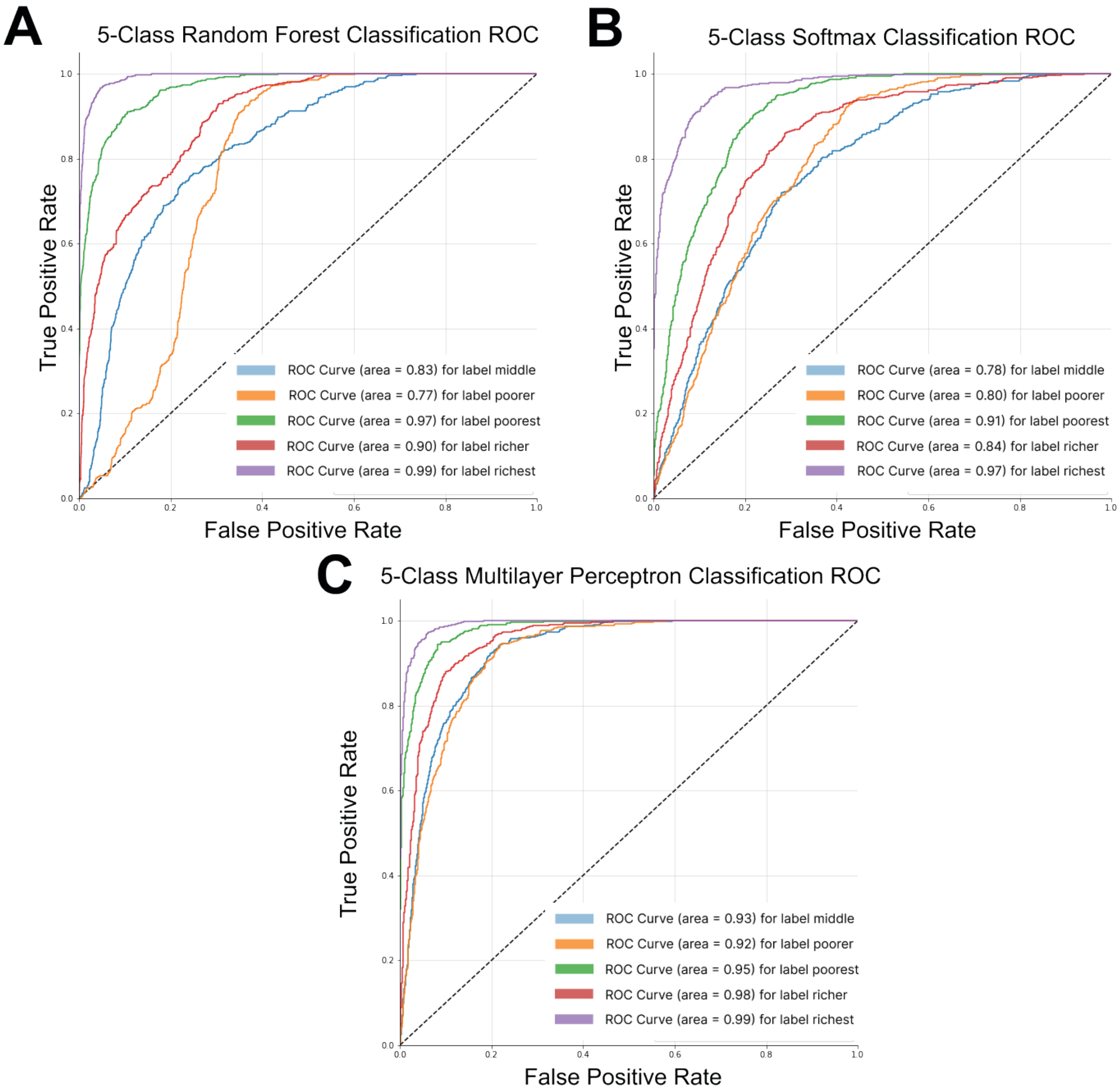


Figure 4: One vs. rest ROC curves for 5-class task. Plots showing ROC curves for each of the 5 different classes in the 5-class task. Each model is trained then evaluated on the testing dataset. The predicted values obtained for each class are used to generate the curves, in which each curve in a plot corresponds to the one-vs-rest task of classifying a label from the rest. Dotted lines are shown along the center of each plot that corresponds to and AUC of 0.5 as a baseline for reference. (A) Plot of ROC curves for each of the 5-classes for the random forest model. (B) Plot of ROC curves for each of the 5-classes for the softmax classification model. (C) Plot of ROC curves for each of the 5-classes for the MLP model.

in the training and evaluation. Beyond the selection of features, using more thorough criteria for the normalization and standardization of features may prove to be beneficial in some way. By focusing on a practical subset of the most discriminative features rather than a vast number of direct features, we streamlined the data retrieval process and maintained high accuracy in poverty prediction, illustrating the potential for impactful, data-driven change.

One future goal is to apply the model to similar datasets with existing classifications of poor/rich to allow us to evaluate its performance and generalizability across different contexts. This would help us understand if the model's accuracy in predicting poverty levels holds true in various populations and regions. However, in cases where explicit classifications are unavailable, the metrics used in the model can still be leveraged for predictions. Even without explicit classifications, analyzing the relationships between these metrics and poverty levels can inform predictive models. Additionally, semi-supervised learning approaches can be used when only some features are available, while advanced data imputation techniques address incomplete datasets. These strategies allow us to make predictions about poverty levels using available information, expanding the applicability of the model in situations where complete classifications or feature availability may be limited. This work also emphasizes the most impactful features in predicting poverty, such as electricity availability, toilet facility type, and cooking fuel type. This indirectly indicates potential areas for policy focus. Future

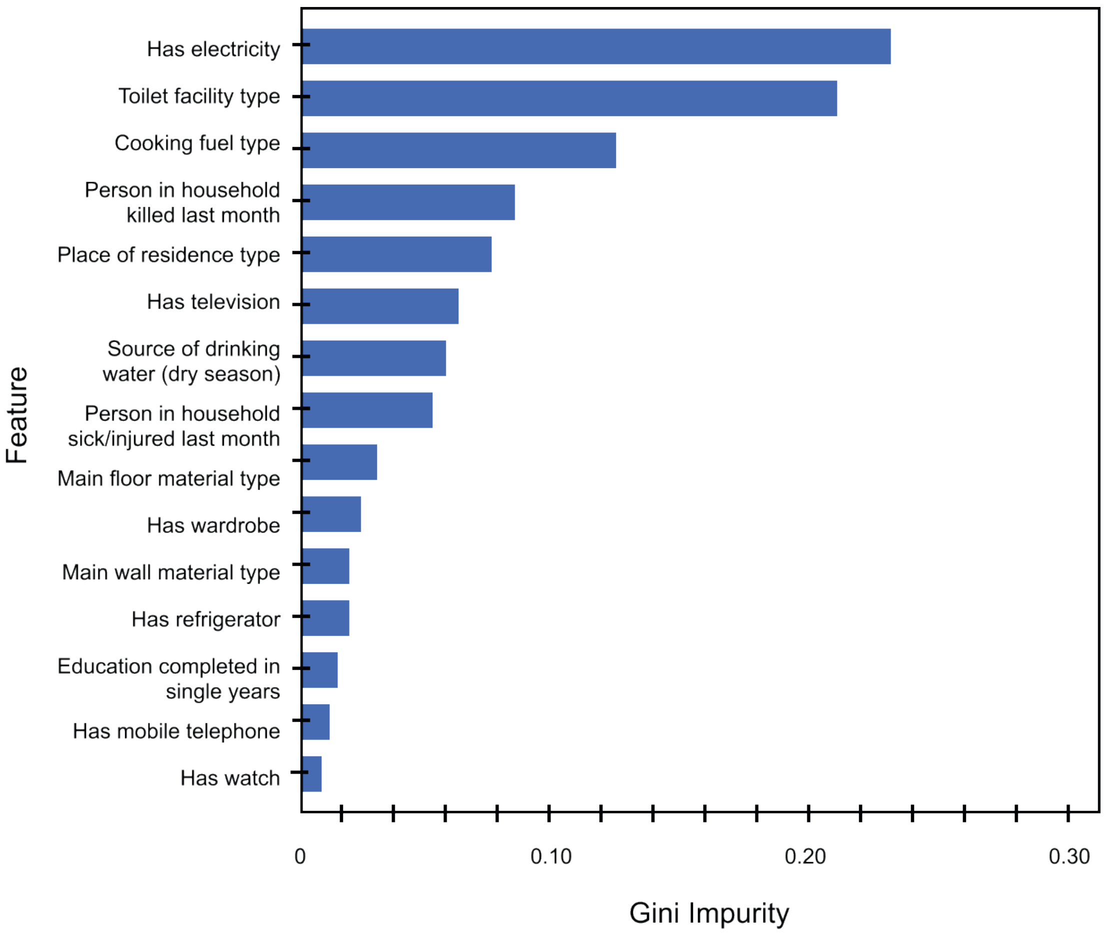


Figure 5: Random forest feature importance on the 3-class task. Bar plot showing the Gini impurity scores of the top 15 features in the evaluation of wealth levels for the 3-class task. These were calculated during the training and evaluation of the random forest classifier model and retrieval of the feature importance parameter values.

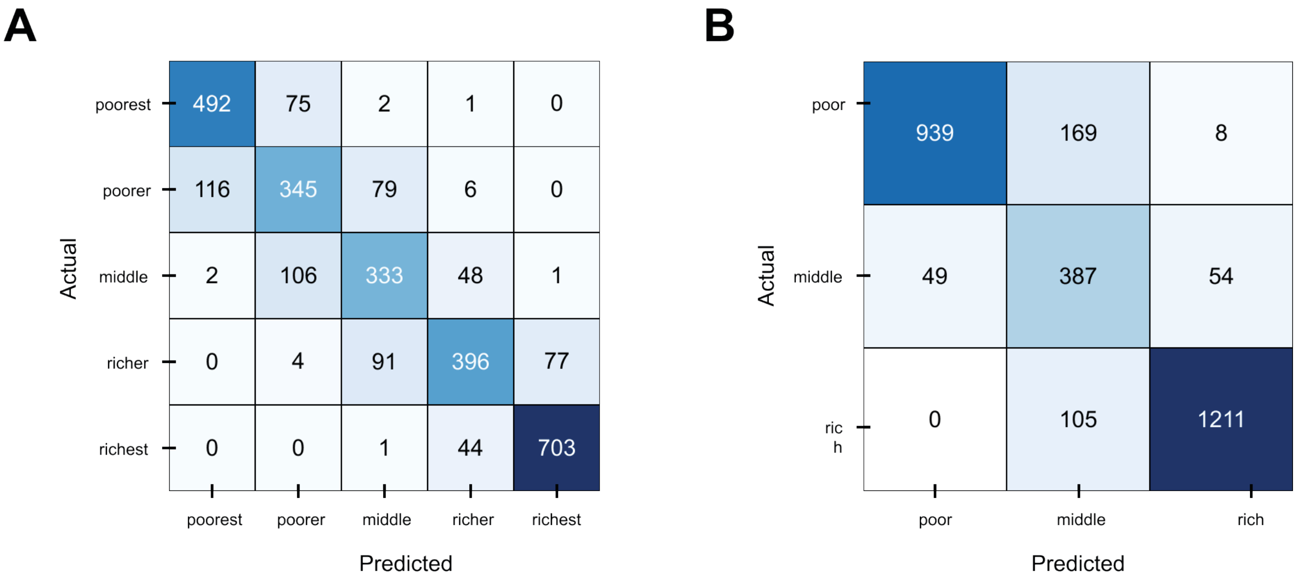
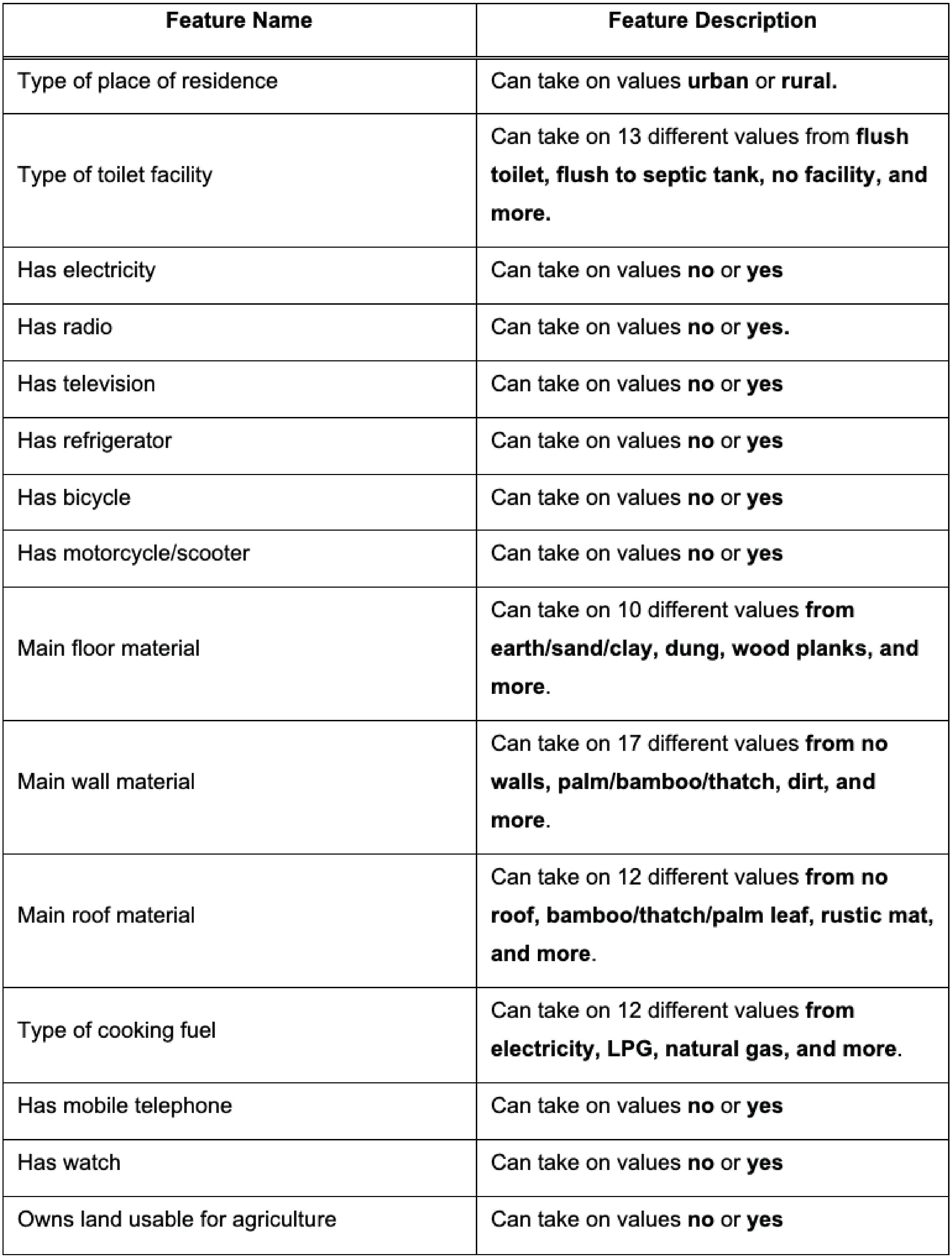
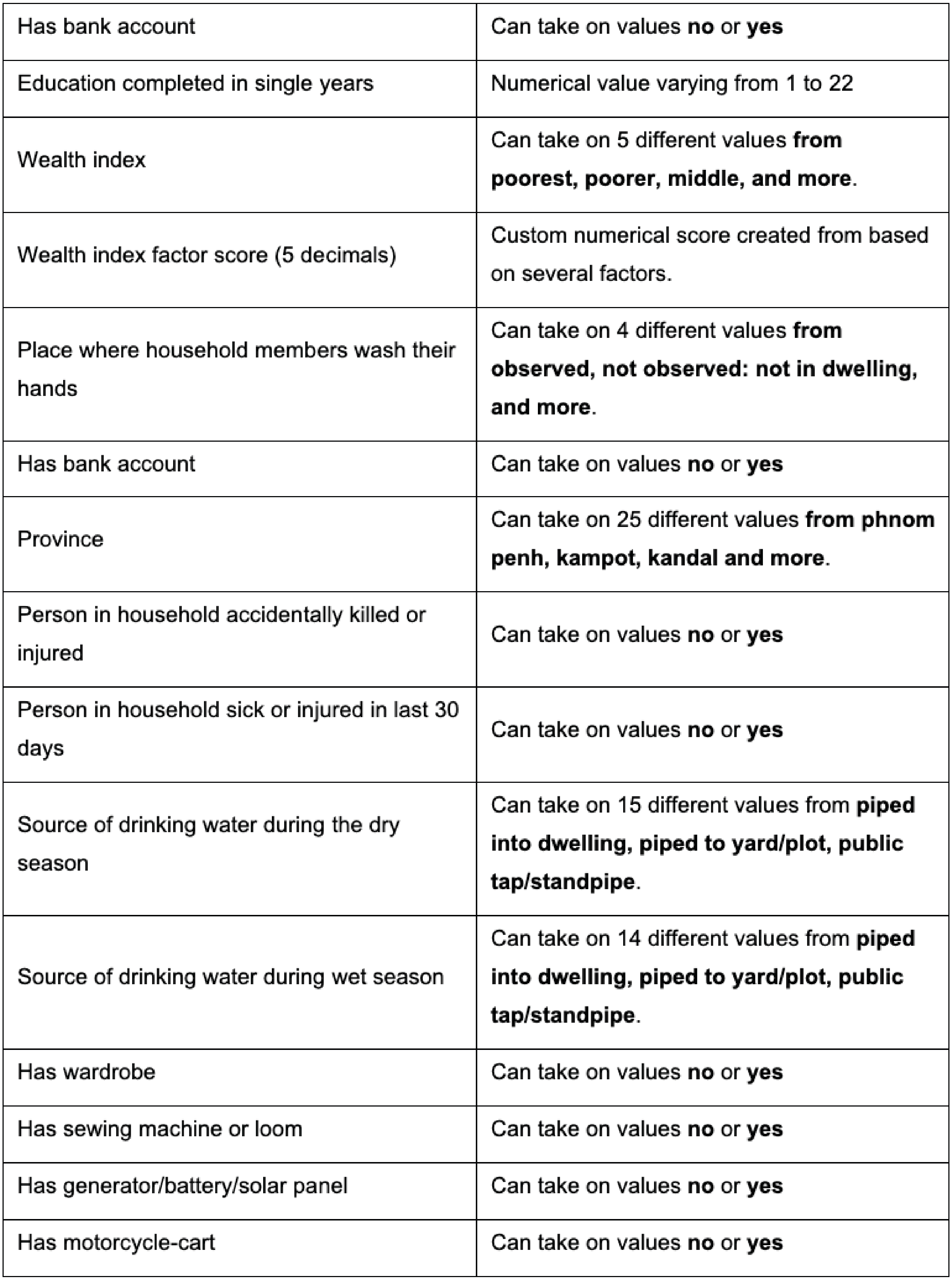


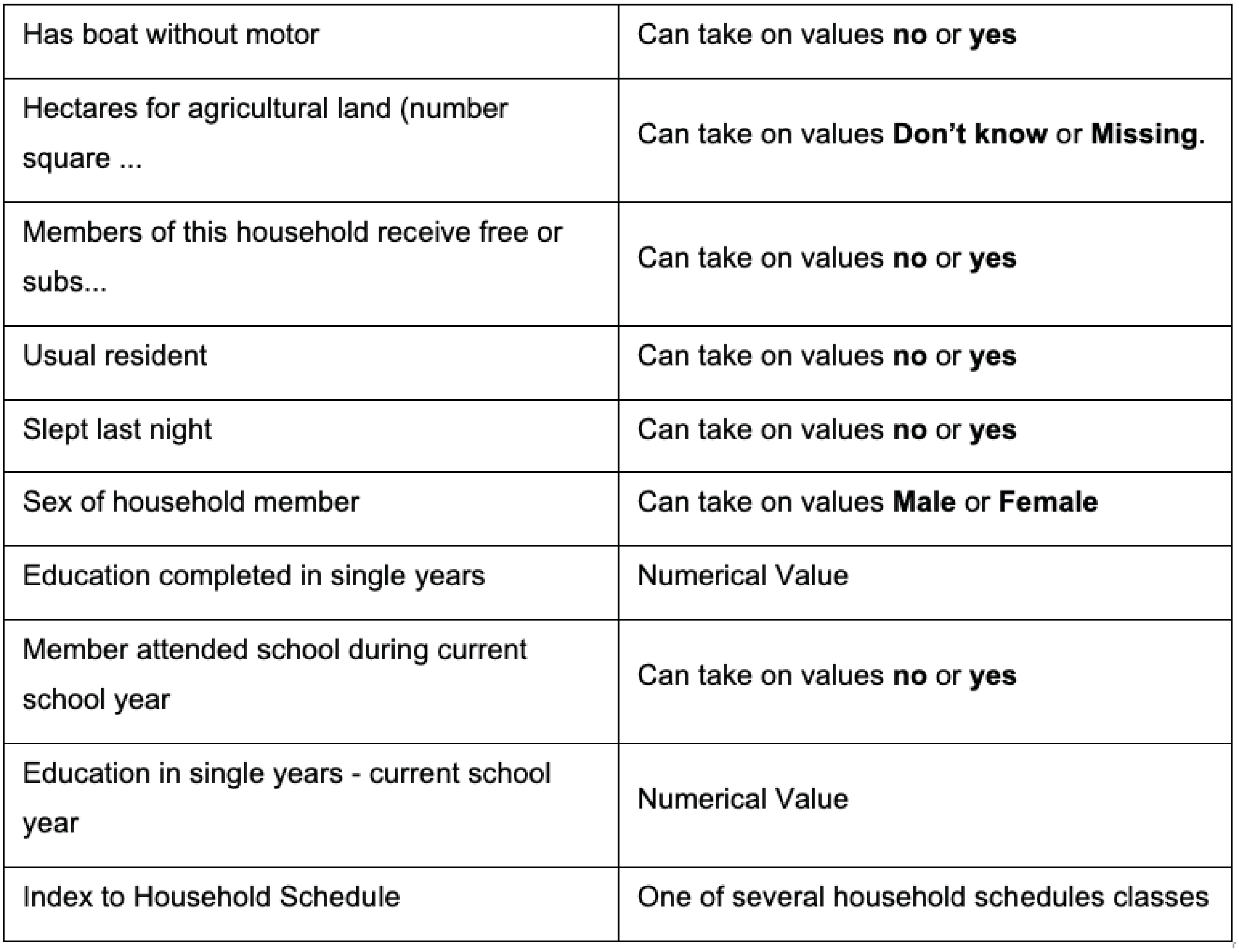
Figure 6: Multilayer perceptron classifier confusion matrices. (A) Confusion matrix table showing the misclassifications made with the model for the 5-class task. We assessed misclassification in a one-vs-all manner for each pair of classes. (B) Confusion matrix table showing the misclassifications made with the model for the 3-class task. We assessed misclassification in a one-vs-all manner for each pair of classes.



Feature Name	Feature Description
Type of place of residence	Can take on values **urban** or **rural.**
Type of toilet facility	Can take on 13 different values from **flush toilet, flush to septic tank, no facility, and more.**
Has electricity	Can take on values **no** or **yes**
Has radio	Can take on values **no** or **yes.**
Has television	Can take on values **no** or **yes**
Has refrigerator	Can take on values **no** or **yes**
Has bicycle	Can take on values **no** or **yes**
Has motorcycle/scooter	Can take on values **no** or **yes**
Main floor material	Can take on 10 different values **from earth/sand/clay, dung, wood planks, and more**.
Main wall material	Can take on 17 different values **from no walls, palm/bamboo/thatch, dirt, and more**.
Main roof material	Can take on 12 different values **from no roof, bamboo/thatch/palm leaf, rustic mat, and more**.
Type of cooking fuel	Can take on 12 different values **from electricity, LPG, natural gas, and more**.
Has mobile telephone	Can take on values **no** or **yes**
Has watch	Can take on values **no** or **yes**
Owns land usable for agriculture	Can take on values **no** or **yes**



Has bank account	Can take on values **no** or **yes**
Education completed in single years	Numerical value varying from 1 to 22
Wealth index	Can take on 5 different values **from poorest, poorer, middle, and more**.
Wealth index factor score (5 decimals)	Custom numerical score created from based on several factors.
Place where household members wash their hands	Can take on 4 different values **from observed, not observed: not in dwelling, and more**.
Has bank account	Can take on values **no** or **yes**
Province	Can take on 25 different values **from phnom penh, kampot, kandal and more**.
Person in household accidentally killed or injured	Can take on values **no** or **yes**
Person in household sick or injured in last 30 days	Can take on values **no** or **yes**
Source of drinking water during the dry season	Can take on 15 different values from **piped into dwelling, piped to yard/plot, public tap/standpipe**.
Source of drinking water during wet season	Can take on 14 different values from **piped into dwelling, piped to yard/plot, public tap/standpipe**.
Has wardrobe	Can take on values **no** or **yes**
Has sewing machine or loom	Can take on values **no** or **yes**
Has generator/battery/solar panel	Can take on values **no** or **yes**
Has motorcycle-cart	Can take on values **no** or **yes**



Has boat without motor	Can take on values **no** or **yes**
Hectares for agricultural land (number square ...	Can take on values **Don't know** or **Missing**.
Members of this household receive free or subs...	Can take on values **no** or **yes**
Usual resident	Can take on values **no** or **yes**
Slept last night	Can take on values **no** or **yes**
Sex of household member	Can take on values **Male** or **Female**
Education completed in single years	Numerical Value
Member attended school during current school year	Can take on values **no** or **yes**
Education in single years - current school year	Numerical Value
Index to Household Schedule	One of several household schedules classes

Table 1: Full set of features and their descriptions. Includes the description of all the possible values that the 40 features used can take. Used a machine learning model to identify the different values.

work could be designed to explicitly model how changes in these features could impact the predicted poverty classes, thus providing more direct policy guidance.

As another future goal, we would like to further fine-tune the subset of data we are working with and additionally incorporate more complex types of information for the analysis and use for poverty prediction. Most notably, we would like to obtain and include GPS data effectively and use it in combination with the survey data to better classify and automate the classification of poverty levels in Cambodia. Looking ahead, our aim is to enhance the data we are working with and incorporate more complex types of information for the purposes of analysis and poverty prediction. One significant aspect of this involves the inclusion of GPS data. When used effectively in conjunction with survey data, it could greatly improve the accuracy of poverty level classification in Cambodia.

In summary, our analyses using the softmax classifier, random forest classifier, and MLP models on household survey data from Cambodia revealed insights into the predictive power of machine learning models in poverty prediction. The features that held the highest importance in the models pointed to certain infrastructural and household variables as crucial markers in predicting wealth levels, notably electricity availability, toilet facility type, and cooking fuel type. The machine learning approaches we tested in this study show promising potential for application in poverty prediction, particularly when combined with comprehensive survey data. This can serve as a basis for more effective, data-driven poverty alleviation strategies. By harnessing machine learning technologies and utilizing comprehensive datasets, we offer a robust and adaptable tool for poverty assessment. This is instrumental in tailoring policies to effectively combat poverty and its consequences, contributing to a more equitable world.

MATERIALS AND METHODS

Dataset & Processing

We used the 2014 Cambodia data for the machine learning model to predict poverty values. For the machine learning models to train properly, we filtered it to remove features and samples with many missing values. We removed many household samples that had missing feature values for lack of important information for the machine learning process. We manually filtered down the 136 features down to 40 features for the remaining 14,607 household samples. This processing

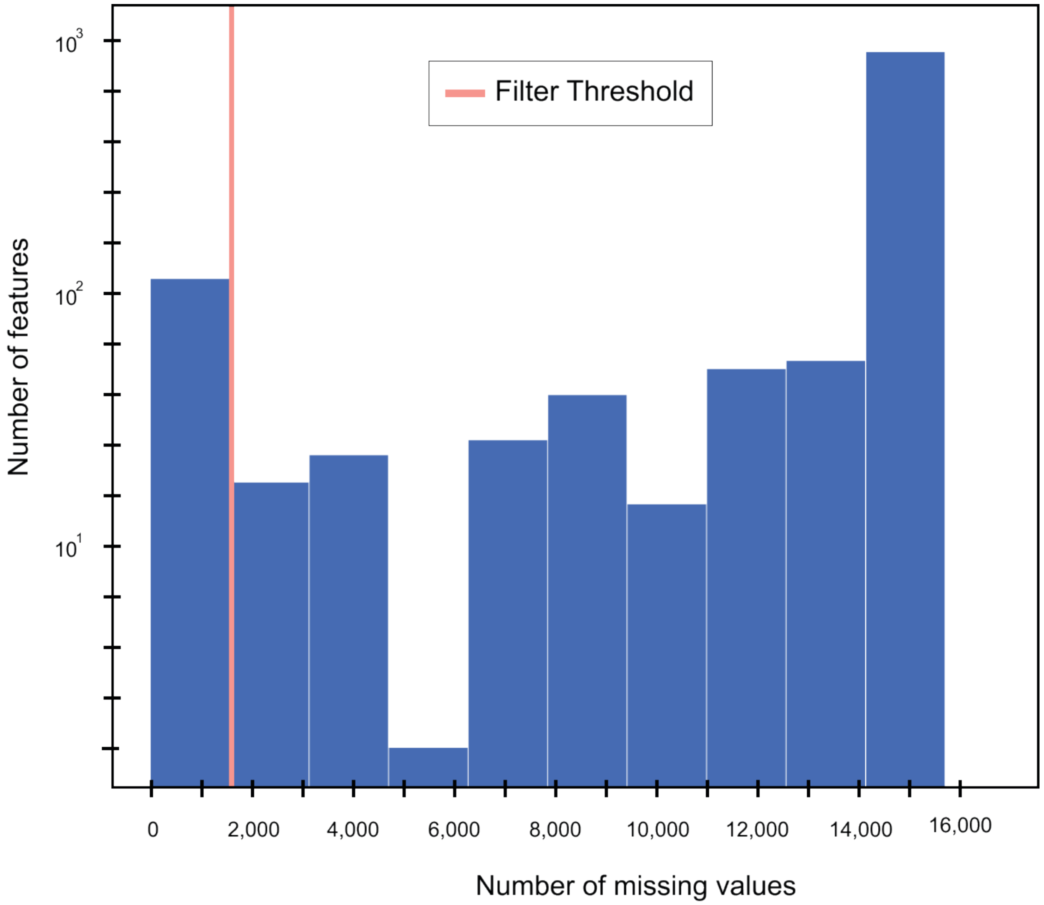


Figure 7: Distribution of missing values per feature. Histogram showing the number of features that possess distinct number of missing values in their columns. We calculated the number of null/missing values across all the samples for each feature and aggregated to obtain a single number of missing features. To inform the feature selection, we only selected features below the designated, manually selected threshold of ~1,750 (pink line).

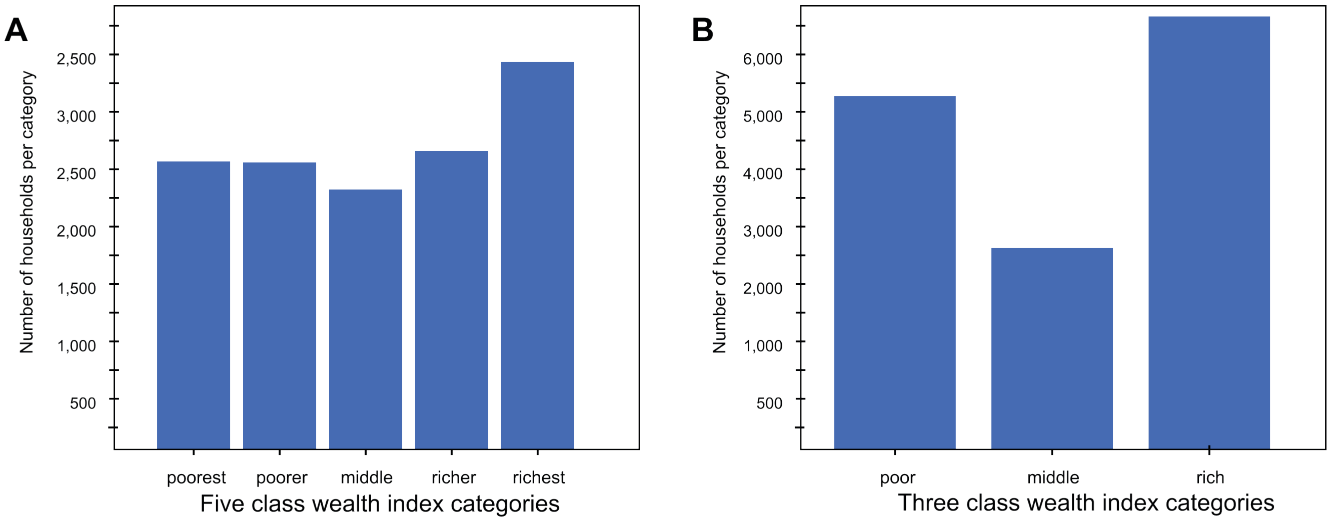


Figure 8: Distribution of the samples among the classes. Histogram plot to depict the relatively balanced distribution of the DHS Cambodia household survey dataset according to the 3-class and 5-class classification tasks. (A) Bar plot showing the number of samples (households) with each of the 5 different wealth labels: poorest, poorer, middle, richer, richest. (B) Bar plot showing the number of samples (households) with each of the 3 different wealth labels: poor, middle, rich.

yielded numerical and categorical features, with a small subset of the 40 features shown **(Table 1)**. Of these features is the wealth index (poor, middle, rich), the feature indicating the poverty level of the household. We separated the filtered dataset into training and testing sets for model training and evaluation with an 80/20 split.

We split the dataset of 14,607 household samples and 40 features into training and testing sets, following an 80/20 split ratio. We filtered the features according to the number of missing values each feature had in the dataset **(Figure 7)**. We evaluated the distribution of the classes to ensure that there was proper balance in the dataset before proceeding with the analysis. There is a balance between the classes very clearly for the 5-class classification task **(Figure 8)**. We used an 80% subset of the samples for training and the remaining 20% for testing. We performed 5-fold cross validation, repeating this procedure five times for each distinct split of the dataset.

Software Versions & Package Information & Computational Methodology

We evaluated training and testing sets for model training and evaluation with an 80/20 split.

The code leverages various Python libraries and software for data analysis and machine learning tasks. These include NumPy and pandas for data manipulation and processing, seaborn and matplotlib for data visualization, and scikit-learn for machine learning modeling and evaluation. A complete list of software and package version information is provided: Python version: 3.9.12, NumPy version: 1.21.5, pandas version: 1.4.2, seaborn version: 0.11.2, matplotlib version: 3.5.1, scikit-learn version: 1.0.2.

First, we pull our data from the Stata file obtained from DHS containing survey data on Cambodia. Since many columns had null values, we dropped columns and rows that had any null entries. We linked the feature values with the feature names given a separate .DO file

We carry out exploratory data analysis to further select a subset of the feature columns and the sample rows. We modify the label feature to take on a form more suitable for the classification task. We specifically also converted the 5-class wealth index labels to 3-class labels for the 3-class classification problem. We format the data as an input matrix and a label vector and thereafter split the data into training and testing sets. We train the three distinct ML models on the data and additionally visualize the important features using the random forest output tree.

Data & Code Availability

All training, evaluation, and analysis for this project is available at: github.com/GW819/poverty-prediction

Received: December 21, 2022
Accepted: May 24, 2023
Published: September 29, 2023

REFERENCES

1. Mahler, Daniel Gerszon, *et al.* "Pandemic, Prices, and Poverty." World Bank Blogs, Apr. 2022, blogs.worldbank.org/opendata/pandemic-prices-and-poverty.
2. Rolf, E., Proctor, J., Carleton, T. *et al.* A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* 12, 4392 (2021). https://doi.org/10.1038/s41467-021-24638-z.
3. ICF. "Cambodia Demographic and Health Surveys." The DHS Program Website. Funded by USAID., Aug. 2022, www.dhsprogram.com.
4. Nations, United. "Ending Poverty." United-Nations., United Nations, 2022, www.un.org/en/global-issues/ending-poverty.
5. Engstrom, Ryan, *et al.* "Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-Being." *The World Bank Economic Review,* vol. 36, no. 2, 2021, pp. 382–412, https://doi.org/10.1093/wber/lhab015.
6. Alsharkawi, Adham, *et al.* "Poverty Classification Using Machine Learning: The Case of Jordan." *Sustainability,* vol. 13, no. 3, 2021, pp. 1412, https://doi.org/10.3390/su13031412.
7. Jean, Neal, *et al.* "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science,* vol. 353, no. 6301, 2016, pp. 790-794, https://doi.org/10.1126/science.aaf7894.
8. Källestål, C., *et al.* "Predicting Poverty. Data Mining Approaches to the Health and Demographic Surveillance System in Cuatro Santos, Nicaragua." I*nternational Journal for Equity in Health,* vol. 18, no. 1, Oct. 2019, https://doi.org/10.1186/s12939-019-1054-7.
9. Yuan, Ye, *et al.* "Gini-Impurity Index Analysis." *IEEE Transactions on Information Forensics and Security,* vol. 16, 2021, pp. 3154–69, http://dx.doi.org/10.1109/TIFS.2021.3076932.