Hormuud University International Journal of Multidisciplinary Studies
https://journals.hu.edu.so/index.php/ihujms

Predicting Digital Inclusion in Somalia Using Machine Learning

Abdukadir Abdullahi Addan*, Ahmed Hassan Hussein
Department of Telecommunication Engineering, Faculty of Engineering, Hormuud University, Mogadishu, Somalia

*Corresponding author's email: engaddaani@gmail.com

Abstract

Digital inclusion affects participation in education, communication, financial services, and the wider digital economy, but individual-level evidence from Somalia remains limited. This study uses microdata from the 2022 Somalia Integrated Household Budget Survey to classify two outcomes among people aged 15 years and older: mobile-phone ownership and internet use during the previous three months. Individual records were merged with household, welfare, and asset information. Logistic regression and random forest classifiers were trained using an 80:20 stratified split and evaluated with accuracy, precision, recall, F1-score, receiver-operating-characteristic curves, and area under the curve (AUC). Random forest outperformed logistic regression for both outcomes. For mobile-phone ownership, random forest achieved accuracy of .841, an F1-score of .903, and an AUC of .860. For recent internet use, it achieved accuracy of .792, an F1-score of .723, and an AUC of .880. Feature-importance results associated mobile-phone ownership mainly with age, household position, assets, expenditure, and household size, whereas internet use was more strongly associated with literacy, education, electricity, assets, and household welfare. The models identify predictive associations rather than causal effects. The findings suggest that policies focused only on device access may be insufficient; broader digital inclusion also requires literacy, education, reliable electricity, affordable connectivity, and improved household economic conditions.

Keywords: Machine learning, Digital inclusion, Mobile-phone ownership, Internet use, Somalia

1. Introduction

Digital inclusion is now an important part of development because mobile phones, broadband, and internet-based services change how people learn, talk to each other, manage their money, and do business online. Previous research on developing nations indicated that simple access to digital technologies is insufficient. For people and businesses to get the most out of them, they also need the right institutions, skills, and support systems. Without these foundations, digital progress might only help a small group of people, leading to enclaves of development instead of more social and economic inclusion (Mansell, 2001).

In Africa, the digital divide has never just been about whether or not people can use technology. It also has to do with how people use digital tools, what skills they have, and what benefits they can get from them. The digital divide in Africa is closely linked to larger problems with income, education, and development (Fuchs & Horak, 2008). More recent studies make a similar argument about infrastructure. Limited access to cheap high-speed internet and low

levels of digital skills are still slowing down the digital transformation in developing countries, especially in rural areas (Kibinda *et al.*, 2025).

Studies have examined who gains access to digital technologies and which factors shape that access. Wesolowski et al. (2012) documented regional, gender, and socioeconomic disparities in mobile-phone ownership in Kenya, with poor rural populations among the least likely to own a phone. Haile et al. (2019) found that mobile-phone ownership improved farmers' access to information and was associated with fewer forecasting errors in Ethiopia. Broadband studies similarly link exclusion to geography, deployment cost, income, and education (Reddick et al., 2020).

Machine-learning methods are increasingly used in ICT and digital-divide research because they can process large datasets and identify nonlinear patterns (Dritsas & Trigka, 2024). Oughton and Mathur (2021) showed how machine learning and satellite imagery can estimate mobile adoption where direct survey data are sparse. Related research on financial technology in Sub-Saharan Africa also emphasizes literacy, infrastructure, and perceived usefulness, while noting the need for stronger field evidence (Hornuf et al., 2025).

Much existing research discusses the African digital divide in general terms, focuses on countries other than Somalia, examines infrastructure and affordability, or applies specialized machine-learning approaches such as satellite-imagery prediction. There is still insufficient evidence from nationally representative Somali microdata to look at individual-level digital inclusion outcomes like owning a mobile phone and using the internet.

This study addresses the gap by using the 2022 Somalia Integrated Household Budget Survey (SIHBS) to examine mobile-phone ownership and internet use during the previous three months. The objective is to classify current digital-inclusion outcomes from individual and household characteristics rather than to forecast future demand across locations. The analysis compares logistic regression with random forest and identifies the predictors that contribute most to classification. It provides national microdata evidence on patterns of digital inequality and highlights population groups that may face a greater risk of exclusion. Predictive importance is interpreted cautiously because it does not establish causality.

2. Methodology

The analysis used microdata from the 2022 SIHBS. Individual records from the household-member file were merged with household, welfare, and asset files using the household identifier. Asset records were aggregated to create indicators for the number and total value of household assets. Figure 1 summarizes the methodological framework. Two binary outcomes were derived from the ICT module: mobile-phone ownership and internet use during the previous three months. Predictors covered age, sex, marital status, relationship to the household head, literacy, education, school attendance, household size, region, settlement type, electricity access, per-capita expenditure, poverty status, asset count, and asset value. The sample was restricted to people aged 15 years and older.

Logistic regression served as an interpretable baseline classifier, while random forest was used to capture nonlinear relationships and interactions. Each outcome dataset was divided into stratified training and test samples in an 80:20 ratio. A random seed of 42 was used for reproducibility. Performance was evaluated using accuracy, precision, recall, F1-score, and AUC. ROC curves and confusion matrices summarized discrimination and classification error. Random-forest feature-importance scores were used to rank predictors. Because impurity-based importance can favor variables with more categories or possible split points, the rankings are interpreted as indicative.

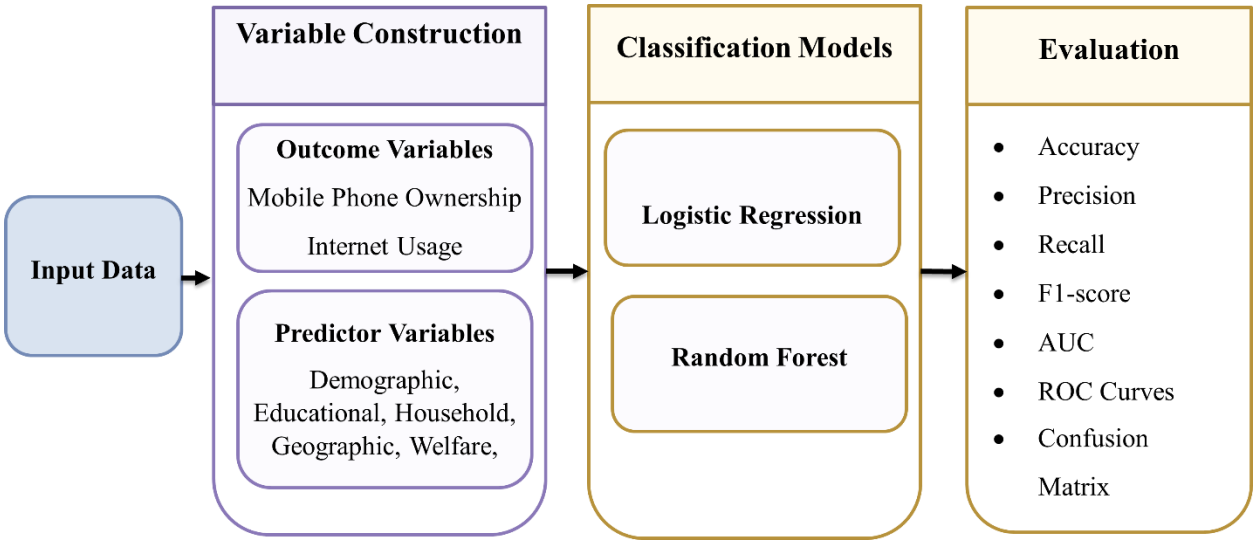


Figure 1: Methodological framework of the study.

3. Results

This section presents the predictive performance of the two classification models for mobile phone ownership and internet use in Somalia. It also discusses the most important predictors identified by the random forest model to explain patterns of digital inclusion.

3.1 Mobile-Phone Ownership

Random forest outperformed logistic regression in classifying mobile-phone ownership. Random forest achieved accuracy of .841, an F1-score of .903, and an AUC of .860, compared with .743, .831, and .823, respectively, for logistic regression. These results indicate stronger test-set performance for random forest and suggest that nonlinearities or interactions may contribute to the classification pattern.

Figure 2 shows the ROC curves for mobile phone ownership. The random forest model achieved a higher AUC than logistic regression, indicating better overall discrimination between phone owners and non-owners.

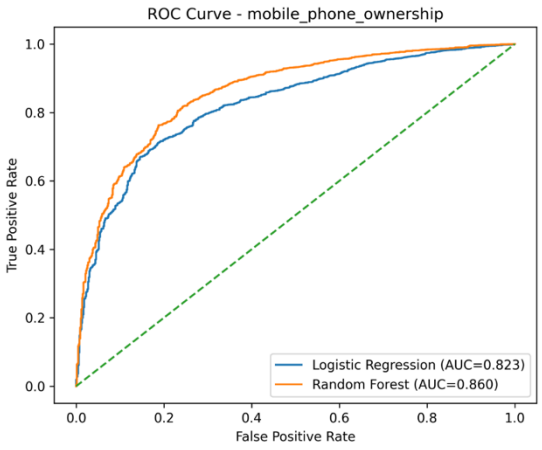


Figure 2: ROC curve for mobile phone ownership

The random-forest confusion matrix contained 3,348 correctly classified phone owners and 442 correctly classified nonowners. It also contained 498 false negatives and 219 false positives, for a total test set of 4,507 observations. Performance was therefore stronger for identifying owners (precision .939, recall .871) than nonowners (precision .470, recall .669), reflecting the high prevalence of phone ownership in the test set (85.3%). Age had the highest reported random-forest feature importance for mobile-phone ownership. Other prominent predictors included marital status, relationship to the household head, household asset value, per-capita expenditure, asset count, and household size. These rankings identify variables useful for prediction but do not measure causal effects or the direction of association. Overall, the results suggest that mobile phone ownership is relatively well structured by demographic and household characteristics, and that a flexible machine learning model is better able to capture this structure than a simple linear classifier.

3.2 Internet Use During the Previous Three Months

Random forest also outperformed logistic regression for recent internet use. It achieved accuracy of .792, an F1-score of .723, and an AUC of .880, compared with .772, .704, and .860, respectively, for logistic regression. Although accuracy and F1-score were lower for internet use than for phone ownership, the internet-use AUC was higher. This means that the model ranked internet users above nonusers effectively across classification thresholds, while classification at the selected threshold remained more difficult.

Figure 3 presents the ROC curves for internet use in the last three months.

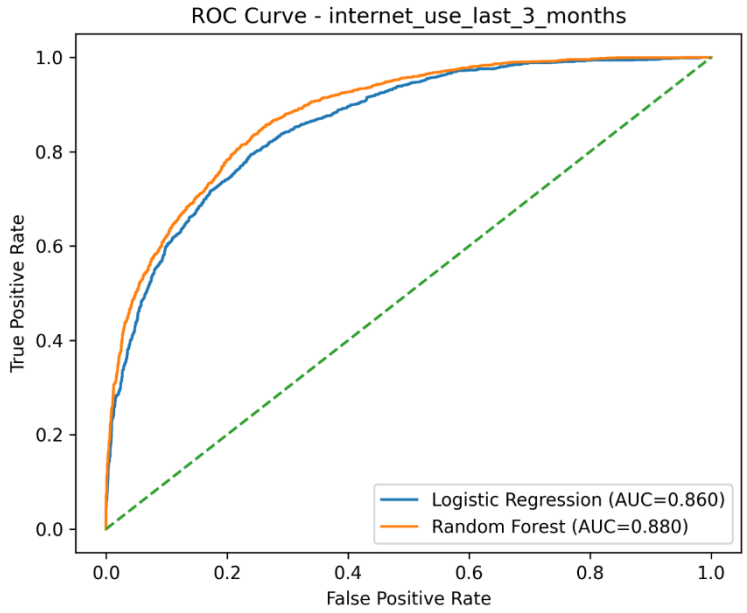


Figure 3: ROC curve for internet use

The random-forest confusion matrix (Figure 4) contained 2,345 correctly classified nonusers and 1,224 correctly classified users, with 603 false positives and 335 false negatives, again totaling 4,507 test observations. For internet users (34.6% of the test set), precision was .670 and recall .785; for nonusers, precision was .875 and recall .795.

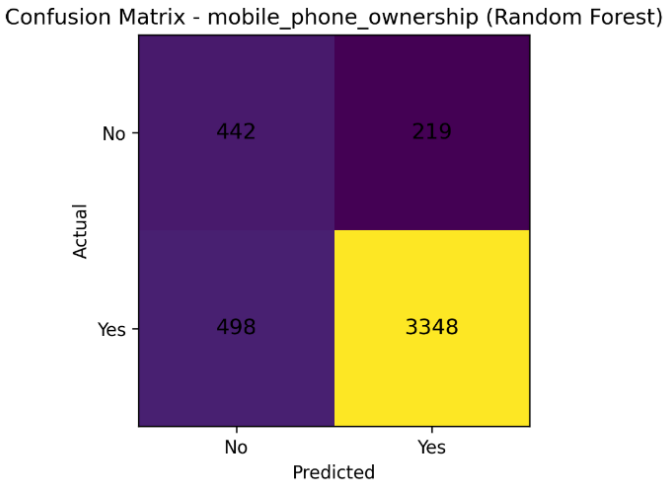


Figure 4: Confusion matrix for internet use

Literacy-related variables, household asset value, age, electricity access, asset count, school attendance, per-capita expenditure, and household size were among the most important predictors of internet use. Compared with mobile-phone ownership, internet use appeared more dependent on literacy, education, electricity, and household resources. These are predictive associations and should not be interpreted as causal determinants without additional analysis.

4. Discussion

Both outcomes could be classified from demographic, educational, and household characteristics, indicating that digital inclusion follows systematic socioeconomic patterns. Mobile-phone ownership was classified more accurately at the selected threshold, while internet use represented a more demanding form of participation associated with skills, electricity, and household resources. Device ownership alone therefore does not capture the full depth of digital inclusion. The consistent advantage of random forest over logistic regression suggests that nonlinearities and interactions may improve classification. However, the magnitude of the performance difference was modest for internet use, and feature importance does not reveal causal mechanisms. Policy analysis should therefore combine predictive modeling with interpretable effect estimates, survey-weighted descriptive statistics, subgroup fairness checks, and external validation. Policies to expand internet use should consider literacy, education, electricity, affordability, and household welfare alongside device access. The study is limited by its cross-sectional design, reliance on a single train–test split, and the absence of external validation. Survey weights and the complex SIHBS sampling design may also affect national interpretation. Future work should report calibration, subgroup performance, confidence intervals, sensitivity to alternative thresholds, and validation on a later survey or independent dataset.

5. Conclusion

Using 2022 SIHBS microdata, this study compared logistic regression and random forest for classifying mobile-phone ownership and recent internet use in Somalia. Random forest produced better test-set metrics for both outcomes, although the models and feature rankings identify predictive patterns rather than causal effects. Mobile-phone ownership was associated primarily with age, household position, and economic characteristics, whereas internet use was more closely associated with literacy, education, electricity, assets, and household welfare. These findings support a multidimensional approach to digital inclusion that combines device access with affordable connectivity, skills, education, electricity, and supportive household

conditions. Stronger model validation and survey-aware analysis are needed before the classifiers are used for national targeting.

Ethics Approval and Informed Consent

This study used de-identified secondary survey microdata; therefore, ethical approval was not required.

Acknowledgements

None

Conflict of Interest

The authors confirm that there is no conflict of interest related to the manuscript.

References

Dritsas, E., & Trigka, M. (2024). Machine learning in information and communications technology: A survey. *Information, 16*(1), 8.

Fuchs, C., & Horak, E. (2008). Africa and the digital divide. *Telematics and Informatics, 25(2),* 99-116.

Haile, M. G., Wossen, T., & Kalkuhl, M. (2019). Access to information, price expectations and welfare: The role of mobile phone adoption in Ethiopia. *Technological Forecasting and Social Change, 145,* 82-92.

Hornuf, L., Safari, K., & Voshaar, J. (2025). Mobile fintech adoption in Sub-Saharan Africa: A systematic literature review and meta-analysis. *Research in International Business and Finance, 73,* 102529.

Kibinda, N., Shao, D., Mwogosi, A., & Mambile, C. (2025). Broadband infrastructure sharing as a catalyst for rural digital economy: A systematic review for developing countries. *Telecommunications Policy*, 103028.

Mansell, R. (2001). Digital opportunities and the missing link for developing countries. *Oxford Review of Economic Policy, 17*(2), 282-295.

Oughton, E. J., & Mathur, J. (2021). Predicting cell phone adoption metrics using machine learning and satellite imagery. *Telematics and Informatics, 62,* 101622.

Reddick, C. G., Enriquez, R., Harris, R. J., & Sharma, B. (2020). Determinants of broadband access and affordability: An analysis of a community survey on the digital divide. *Cities, 106,* 102904.

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. (2012). Heterogeneous mobile phone ownership and usage patterns in Kenya. *PLoS ONE, 7(4),* e35319.