Predicting Financial Inclusion in East Africa Using Machine Learning: A Comparative Study of Classification and Clustering Algorithms

Student Name

Faculty of Information Technology, Department of Big Data Analytics
Adventist University of Central Africa (AUCA), Kigali, Rwanda
Course: MSDA 9213 – Data Mining

February 2026

Abstract

Financial inclusion remains a critical challenge in East Africa, where only 13.9% of adults across Kenya, Rwanda, Tanzania, and Uganda have access to commercial bank accounts. This study applies machine learning techniques to predict individual bank account ownership using demographic and socioeconomic data from FinScope surveys (2016–2018) covering 23,524 individuals across four East African Community (EAC) countries. Four classification algorithms were implemented: Logistic Regression, Random Forest, Gradient Boosting, and a Multi-Layer Perceptron (MLP) neural network. Gradient Boosting achieved the best baseline performance with an AUC-ROC of 0.874 and F1-score of 0.491. To address the severe class imbalance (86% negative class), random oversampling and class weighting were applied, improving recall by up to 54.8 percentage points and F1-scores across all models. Additionally, K-Means and Hierarchical clustering identified four distinct demographic groups with varying financial inclusion levels. The results highlight that education level, job type, country, and age are the strongest predictors of bank account ownership, providing actionable insights for policymakers seeking to expand financial access in the EAC region.

Keywords: Financial Inclusion, Machine Learning, Classification, Clustering, East Africa, Data Mining

1 Introduction

Financial inclusion—defined as access to useful and affordable financial products and services—is widely recognized as a key enabler of economic development and poverty reduction (Demirguc-Kunt et al., 2018). In Sub-Saharan Africa, and particularly in the East African Community (EAC) countries of Kenya, Rwanda, Tanzania, and Uganda, financial exclusion remains a significant barrier to economic growth. Across these four nations, only approximately 9.1 million adults, representing 13.9% of the adult population, have access to or use a commercial bank account (Zindi, 2019).

Despite the rapid expansion of mobile money platforms such as M-Pesa, traditional banking services remain essential for enabling savings, facilitating payments, building creditworthiness, and accessing other financial services. Understanding the demographic and socioeconomic factors that determine whether an individual has a bank account is therefore critical for designing targeted interventions to improve financial access.

Machine learning offers powerful tools for analyzing large-scale survey data and identifying complex patterns that traditional statistical methods may overlook (James et al., 2023). Recent studies have demonstrated the effectiveness of various machine learning algorithms in predicting financial inclusion outcomes in developing countries (Alshebami & Seraj, 2021; Asongu & Odhiambo, 2020).

The objectives of this study are threefold: (1) to build and compare multiple classification models for predicting bank account ownership among individuals in four EAC countries; (2) to identify the most important demographic factors driv-

ing financial inclusion; and (3) to discover natural groupings within the population using clustering algorithms. This study employs four classification algorithms—Logistic Regression, Random Forest, Gradient Boosting, and a deep learning Multi-Layer Perceptron—and two clustering methods—K-Means and Hierarchical clustering—to achieve these objectives.

2 Methods

2.1 Dataset Description

The dataset used in this study was obtained from the Zindi Financial Inclusion in Africa competition, which compiled data from multiple FinScope surveys: FinAccess Kenya 2018, Finscope Rwanda 2016, Finscope Tanzania 2017, and Finscope Uganda 2018 (Zindi, 2019). The training dataset contains 23,524 observations with 13 variables. After removing the unique identifier, 11 features were used for modeling.

The feature set includes: *country* (Kenya, Rwanda, Tanzania, Uganda), *year* (survey year), *location_type* (Rural/Urban), *cellphone_access* (Yes/No), *household_size*, *age_of_respondent*, *gender_of_respondent*, *relationship_with_head*, *marital_status*, *education_level*, and *job_type*. The target variable is *bank_account* (Yes/No), indicating whether the individual has a bank account.

The dataset exhibits a significant class imbalance: 20,212 individuals (85.9%) do not have a bank account, while only 3,312 (14.1%) do. The country distribution shows Rwanda (8,735), Tanzania (6,620), Kenya (6,068), and Uganda (2,101). No missing values or duplicate records were found.

2.2 Data Preprocessing

Data preprocessing involved the following steps: (1) removal of the `uniqueid` column, which serves only as an identifier; (2) encoding of the binary target variable (`Yes`→1, `No`→0); (3) label encoding of all categorical features using scikit-learn's `LabelEncoder`; and (4) feature standardization using `StandardScaler` for algorithms sensitive to feature scaling (Logistic Regression and MLP). The data was split into training (80%) and testing (20%) sets with stratified sampling to preserve the target class distribution.

2.3 Classification Algorithms

Four classification algorithms were selected to provide a diverse comparison:

Logistic Regression (LR): A linear model that estimates the probability of bank account ownership using a logistic function. Configured with `max_iter=1000` and default regularization.

Random Forest (RF): An ensemble method that builds 200 decision trees with a maximum depth of 15, aggregating their predictions through majority voting (Breiman, 2001).

Gradient Boosting (GB): A sequential ensemble method that builds 200 weak learners (depth 5) with a learning rate of 0.1, where each tree corrects the errors of the previous ones (Friedman, 2001).

Multi-Layer Perceptron (MLP): A deep neural network with three hidden layers (128, 64, 32 neurons), ReLU activation, Adam optimizer, and early stopping—representing the deep learning component of this study (Goodfellow et al., 2016).

2.4 Model Improvement

To address the class imbalance, two strategies were applied: (1) **Random Oversampling:** The minority class (bank account holders) was oversampled by randomly duplicating instances to match the majority class size, increasing the training set from 18,819 to 32,338 observations; (2) **Class Weighting:** For Logistic Regression and Random Forest, the `class_weight='balanced'` parameter was used to assign higher weights to the minority class. Additionally, hyperparameters were tuned: increased estimators (300), deeper trees, adjusted learning rates, and a deeper MLP architecture (256, 128, 64, 32 neurons).

2.5 Clustering Algorithms

Two clustering algorithms were applied to the encoded feature set:

K-Means Clustering: The optimal number of clusters was determined using the Elbow method and Silhouette analysis. K=4 was selected, corresponding to the four EAC countries.

Hierarchical (Agglomerative) Clustering: Ward's linkage was used to minimize within-cluster variance. Due to computational constraints, a random sample of 10,000 observations was used.

2.6 Evaluation Metrics

Classification models were evaluated using: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Given the class imbalance, F1-Score and AUC-ROC are prioritized as they better capture performance on the minority class. Clustering was evaluated using the Silhouette Score and cluster profile analysis.

3 Results and Interpretation

3.1 Exploratory Data Analysis

The EDA revealed several key patterns. Kenya has the highest financial inclusion rate (26.3% with bank accounts), followed by Uganda (13.1%), Tanzania (8.6%), and Rwanda (7.0%). Urban residents are approximately 3 times more likely to have bank accounts than rural residents. Individuals with tertiary education have significantly higher inclusion rates (over 50%) compared to those with no formal education (below 5%). Government and private formal employees show the highest bank account ownership rates, while farming/fishing workers and government dependents show the lowest.

Table 1: Baseline Classification Model Performance

Model	Acc.	Prec.	Rec.	F1	AUC
Log. Reg.	0.874	0.649	0.227	0.336	0.836
Rand. Forest	0.886	0.662	0.384	0.486	0.866
Grad. Boost.	0.887	0.672	0.387	0.491	0.874
MLP (DL)	0.885	0.669	0.363	0.470	0.864

3.2 Baseline Classification Results

Table 1 presents the baseline performance of all four models. Gradient Boosting achieved the highest AUC-ROC (0.874) and F1-Score (0.491), followed closely by Random Forest (AUC=0.866, F1=0.486). While all models achieved high accuracy (87–89%), this is misleading due to class imbalance—a model predicting all instances as "No" would achieve 85.9% accuracy. The low recall values (22.7–38.7%) indicate that all models struggle to identify bank account holders in the minority class.

The feature importance analysis from Random Forest revealed that *country*, *education_level*, *job_type*, and *age_of_respondent* are the most influential predictors, while *year* and *gender* contribute relatively less.

3.3 Improved Model Results

Table 2: Improved Classification Model Performance

Model	Acc.	Prec.	Rec.	F1	AUC
Log. Reg.	0.732	0.316	0.775	0.449	0.836
Rand. Forest	0.861	0.506	0.554	0.529	0.855
Grad. Boost.	0.814	0.410	0.736	0.527	0.867
MLP (DL)	0.817	0.402	0.622	0.489	0.816

Table 2 shows the results after applying class balancing and hyperparameter tuning. The most significant improvements were observed in recall: Logistic Regression's recall increased from 0.227 to 0.775 (+54.8 pp), and Gradient Boosting's recall increased from 0.387 to 0.736 (+34.9 pp). F1-Scores improved for all models: Random Forest achieved the best F1-Score of 0.529 (+4.4 pp) and Gradient Boosting reached 0.527 (+3.6 pp). The trade-off between precision and recall is expected when addressing class imbalance; the improved models sacrifice some precision to substantially improve their ability to identify bank account holders.

3.4 Clustering Results

Table 3: Clustering Performance

Algorithm	Silhouette Score
K-Means (K=4)	0.159
Hierarchical (K=4)	0.158

3.5 Predictions on Unseen Test Data

The best performing improved model (Gradient Boosting with oversampling) was applied to the unseen Test dataset (10,086 individuals) which does not contain the target variable. The model predicted that approximately 30% of individuals in the test set have a bank account, with Kenya showing the highest predicted inclusion rate, consistent with the training data patterns. These predictions were formatted and exported in the Zindi competition submission format. The predicted probability distribution shows a clear separation between the two classes, with most probabilities concentrated near 0 (no account) and a smaller peak near 1 (has

account), confirming the model's confidence in its predictions.

3.6 Clustering Results

Both K-Means and Hierarchical clustering produced similar results (Table 3). The moderate silhouette scores suggest overlapping clusters, which is expected given the mixed nature of demographic data. The four identified clusters correspond to distinct demographic profiles: (1) young Tanzanian/Ugandan rural residents with low education; (2) elderly individuals with limited education and low cellphone access; (3) Kenyan respondents with higher education and cellphone access; and (4) rural married individuals in farming/fishing occupations.

4 Discussion and Recommendations

4.1 Discussion

This study demonstrates that machine learning algorithms can effectively predict bank account ownership in East African countries, with Gradient Boosting and Random Forest consistently achieving the best performance. The findings align with prior literature suggesting that ensemble methods outperform linear models and simple neural networks on tabular demographic data (James et al., 2023).

The severe class imbalance (86% vs 14%) posed a significant challenge. Baseline models achieved high accuracy but poor minority class detection. Applying random oversampling and class weighting substantially improved recall and F1-scores, demonstrating the importance of addressing class imbalance in real-world datasets (Chawla et al., 2002).

The feature importance analysis provides actionable insights: education level, job type, and country are the strongest determinants of financial inclusion. This confirms findings from economic research that education and formal employment are key drivers of financial access in developing countries (Demirguc-Kunt et al., 2018).

The clustering analysis revealed meaningful population segments, though the moderate silhouette scores indicate that financial inclusion is not determined by a single clear-cut demographic divide but rather by a complex interplay of multiple factors.

A limitation of this study is the use of label encoding for ordinal and nominal categorical variables, which may impose artificial ordinal relationships. Future work could explore one-hot encoding or target encoding. Additionally, the MLP neural network did not significantly outperform traditional ensemble methods, consistent with observations that deep learning provides less advantage on smaller tabular datasets (James et al., 2023).

4.2 Recommendations

Based on the findings, we recommend:

1. **Targeted education programs:** Individuals with no formal or primary education show the lowest inclusion rates. Financial literacy programs should be integrated with basic education initiatives.
2. **Rural outreach:** Urban-rural disparity in bank account ownership remains significant. Mobile banking solutions and agent banking networks should be expanded in rural areas.
3. **Leveraging cellphone access:** The strong association between cellphone access and bank account ownership suggests that mobile-first banking solutions could bridge the financial inclusion gap.
4. **Country-specific strategies:** Rwanda and Tanzania lag behind Kenya and Uganda in bank account penetration, requiring tailored policy interventions.
5. **Class-balanced modeling:** Future predictive models for financial inclusion should incorporate class balancing techniques to ensure accurate identification of underserved populations.

References

Alshebami, A. S. & Seraj, A. H. A. (2021). Financial inclusion and economic growth nexus: Evidence from Sub-Saharan Africa. *African Journal of Economic and Management Studies*, 12(4), 480–498.

Asongu, S. A. & Odhiambo, N. M. (2020). Financial access, governance and insurance sector development in Sub-Saharan Africa. *Journal of Economic Studies*, 47(4), 849–875.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic

minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution*. World Bank Publications.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer.

Zindi (2019). Financial Inclusion in Africa Competition. `https://zindi.africa/competitions/financial-inclusion-in-africa`.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.