Classification of Tea Farmers' Credit Source Decisions in Rwanda: A Machine Learning Approach

Celestin Niyomugabo[1,2]

[1] Department of IT, Adventist University of Central Africa, Kigali - Rwanda
[2] VONSUNG, Kigali - Rwanda
Correspondence: celestin@vonsung.co.rw

Abstract: Access to agricultural credit remains a critical factor for enhancing productivity and sustainability among smallholder farmers in Rwanda. This study investigates the determinants of tea farmers' credit source decisions (specifically the choice between formal and informal credit) using machine learning techniques on survey data collected from 622 credit-using farmers across key tea-growing districts. The objective is to uncover the behavioral and structural factors that influence farmers' access to formal financial institutions. Four classification models were applied: Logistic Regression, Random Forest, Support Vector Machine, and Neural Network. Among these, the Random Forest model achieved the highest predictive performance, with an accuracy of 99.2%, a precision score of 1.000, recall of 0.984, and F1 score of 0.992. The Neural Network model also performed strongly, with an accuracy of 97.6%. In contrast, Logistic Regression, while interpretable, showed limited performance (67.2% accuracy), highlighting its insufficiency in capturing complex, non-linear patterns in farmers' credit behavior. Regression analysis revealed that farm size, household income from tea, household size, and distance to financial institutions were statistically significant predictors of formal credit use. Notably, larger farm size ($p < 0.001$) and higher tea income ($p < 0.05$) were positively associated with access to formal credit, while the effect of distance to financial institutions was unexpectedly positive ($p < 0.01$), suggesting that more motivated farmers are willing to overcome geographic barriers. The study concludes that machine learning offers robust tools for understanding rural credit dynamics and can inform more targeted, inclusive financial policies. These findings have practical implications for policymakers, financial service providers, and development partners seeking to expand access to formal credit among smallholder farmers in Rwanda.

Keywords: Machine Learning; Tea; Rwanda; Credit

1. Introduction

Agricultural credit is widely recognized as a crucial enabler of farm investment and productivity. In Rwanda's tea sector (one of the country's top export earners) smallholder farmers need to secure credit to purchase fertilizer, renew plantations, and cover labor costs. As the government pushes for increased tea production and modern farming practices, the demand for credit among tea growers has risen. However, rural credit markets are often segmented into formal and informal channels, each with distinct requirements and constraints. Formal lenders (commercial banks, microfinance institutions) typically offer larger loans but require collateral, have stricter eligibility criteria, and may be less accessible to remote farmers. In addition to this, smallholder farmers often lack loan products tailored to the need fot farmers. Informal credit sources (including cooperatives, Village Saving and Loans Groups, local moneylenders, or friends and relatives) operate within community networks and often provide flexible, small loans based on personal trust.

For Rwanda's small-scale tea farmers, choosing between formal and informal credit is a strategic decision. Formal credit can supply substantial capital for farm improvements, yet many smallholders cannot qualify due to lack of collateral or credit history. Informal credit is more accessible for quick, small needs, though it may come with higher implicit costs or limited loan sizes. Prior research in Rwanda and other developing countries has highlighted that rural households often face credit constraints in the formal sector and thus rely heavily on informal finance [1][2].

For instance, Papias and Ganesan (2010) find that Rwanda's rural financial system is underdeveloped, leaving many farm households unable to access formal loans [3]. It is evident from only few farmers get formal credits. A survey by Kajigija shows that only about 4% of farmers in some regions obtained credit from formal banks, whereas over half used informal mechanisms [4]. This dichotomy raises important questions about what factors influence a farmer's choice of credit source and how those factors can be addressed to improve credit uptake.

Existing literature has examined credit access determinants using classical econometric approaches. A recent study by Kabayiza et al. (2021) employed a multivariate probit model to analyze tea farmers' selection of credit sources in Nyaruguru District. Their findings indicated that possessing collateral, facing lower interest rates, and having larger tea farms significantly increased the likelihood of using formal bank loans, while farmers who only needed a small loan, had received technical training, or participated in group borrowing were more inclined toward informal credit. Such studies underscore that a mix of demographic factors (age, education, gender, household size), economic factors (farm size, income level, collateral assets), and behavioral factors (risk preferences, social network participation, financial literacy) all potentially shape credit source decisions [5]. However, there is still limited understanding of how these factors interact in influencing the choice between formal and informal credit, especially when farmers might use both in combination.

Moreover, while econometric models identify significant variables on average, they may not capture complex, non-linear patterns in decision-making. In recent years, machine learning (ML) techniques have gained traction in financial services for tasks like credit scoring and risk assessment. ML models can handle high-dimensional data and interactions, potentially offering improved predictive power and new insights beyond traditional models. For example, AI-driven credit scoring systems have been developed to incorporate alternative data to predict smallholders' loan repayment ability and expand credit access [6].

In rural finance research, ML approaches remain relatively novel but promising for analyzing household behavior and segmenting borrowers. By applying supervised learning algorithms to survey data, we can potentially classify farmers' credit source choices with reasonable accuracy and discover which features are most influential, even if relationships are complex or non-linear.

2. Objective of the Study

This study aims to contribute a new perspective by using a machine learning approach to classify Rwandan tea farmers' credit source decisions (formal vs. informal). I'm seeking to:

1. build predictive models that distinguish which type of credit a farmer uses based on their characteristics, and
2. identify the key factors (features) driving these predictions, thereby shedding light on the determinants of credit choice

This objective will be achieved by comparing multiple ML algorithms alongside a baseline logistic regression, and also evaluate whether advanced techniques offer performance gains in this domain. Ultimately, my goal is to deepen understanding of smallholder borrowing behavior and inform policy interventions (such as financial literacy programs, credit guarantee schemes, or blended finance models) to ensure farmers can access the most appropriate and sufficient sources of credit for their needs.

3. Methodology

3.1 Area Coverage and Sample Design

This study was conducted in Rwanda, targeting districts with predominant tea farming activities. The final sample was drawn from nine major tea-growing districts across the Northern, Western, and Southern Provinces, namely: Nyaruguru, Nyabihu, Rulindo, Gicumbi, Rutsiro, Ngororero, and Musanze. These areas were purposefully selected due to their strong engagement in smallholder tea production.

To ensure representativeness and statistical validity, the sample size was determined by Priori Power analysis. Using G*Power calculations, a priori power analysis showed that a sample size of 372 farmers will be needed for a Probit regression model with α error probability of 0.05 and actual power of 0.80. For t-test with small effect size ($f^2 = 0.10$) and α error probability of 0.05, the total sample size to achieve the actual power of 0.80 is $n = 620$. Hence the whole study's total sample size is determined to be $n = 620$. The sample will be stratified into 3 strata (small scale farmers, medium scale farmers and large scare farmers).

3.2 Data Collection

This study relied on primary data collected through the use of Computer Assisted Personal Interviewing (CAPI) techniques. Specifically, a smartphone-embedded data collection tool was deployed, enabling enumerators to conduct interviews directly with tea farmers in the field. The digital tool was programmed to automatically sync collected data to a secure remote server, thereby reducing the risk of data loss due to device malfunction or damage.

Fieldwork was carried out between May 12 and May 25, 2023. A team of five trained enumerators conducted the interviews. Prior to data collection, the enumerators received comprehensive training on the questionnaire content, interview protocols, and the use of the digital data collection platform to ensure consistency and data quality. During data collection, regular supervision and back-checks were conducted to validate the accuracy and completeness of responses.

3.3 Dataset Description

The dataset contains a total of 131 variables encompassing demographic information, farm characteristics, income sources, credit behavior, and access to financial services.

Key variables include:

- **Demographics:** age, sex, education level, household size
- **Farm profile:** tea plantation area, tea income, land size
- **Credit access:** formal and informal credit usage, loan amounts, interest rates, repayment periods, and default experiences
- **Credit decision factors:** motivations, barriers to access, collateral requirements, proximity to institutions
- **Geolocation:** coordinates of each respondent's location

Data types include both numerical and categorical variables. Some fields related to credit motivation and default reasons are structured as multiple response items. The dataset contains missing values in certain sections, particularly among variables that are conditional on whether a farmer accessed credit. Data cleaning and validation were performed prior to analysis to handle inconsistencies and ensure quality.

3.4 Data Preprocessing and Feature Selection

Initial data validation was conducted to identify and correct inconsistencies, incomplete responses, and potential duplicates. This was followed by data cleaning procedures which included labeling of variables, recoding of responses, and handling of missing values.

Enumerators' notes and flagging features embedded in the CAPI system were reviewed to resolve any ambiguities in the responses. Logic checks and range validations were applied to ensure internal consistency across variables. The data cleaning employed the use of Python (version 3.12), and the final dataset was securely archived with version control to allow for reproducibility and auditing of the analytical procedures.

Before feeding the data into machine learning models, we performed several preprocessing steps:

- **Missing Data & Cleaning:** Most missing values were handled by either removal or imputation, and a few inconsistent observations were dropped, resulting in a clean dataset of 622 farmers.
- **Categorical Encoding:** some variables like gender, educated, etc were converted to a binary numeric format, and ordinal categories like income and education levels were treated as numeric. Other categorical variables were dummified as needed.
- **Feature Selection:** Only features present before a farmer made a credit decision (e.g., age, income, education, distance to lender) were chosen to predict credit choice. This excluded post-decision outcomes like loan amounts to ensure the model explained initial factors. Eight key variables were selected.
- **Normalization:** Continuous variables were standardized for models like Support Vector Machine (SVMs) and Neural Networks, while tree-based models received unnormalized data for better interpretability.

For our models, the dependent variable was a binary label representing the farmer's primary credit source: formal loan (1) or exclusively informal loan (0). Observations of farmers who did not utilize any credit were omitted from the training dataset, thereby focusing the classification task on differentiating between formal and informal credit users.

3.5 Machine Learning Models

To classify tea farmers' credit source decisions, four supervised machine learning algorithms were implemented using Python:

- **Multinomial Logistic Regression:** Used as a baseline model for interpretability and comparison to traditional econometric methods.
- **Random Forest:** An ensemble model composed of 100 decision trees, chosen for its ability to handle non-linear relationships and provide feature importance rankings.
- **Support Vector Machine (SVM):** Implemented with a radial basis function (RBF) kernel to capture non-linear patterns. A one-vs-one strategy was used for multi-class handling, with hyperparameters tuned via cross-validation.
- **Neural Network (Multi-layer Perceptron):** A feed-forward neural network with a single hidden layer of 10 neurons, ReLU activation function, and the Adam optimizer. Training was run for up to 1000 iterations with early stopping criteria handled internally by the solver.

The dataset was split into training and testing subsets using a 70/30 stratified split to maintain class balance. Final model evaluations were performed on the hold-out test set to assess generalization to unseen data.

3.6 Evaluation Metrics

Model performance was assessed using multiple metrics suitable for multi-class classification:

- **Accuracy:** Measures the proportion of correct predictions.
- **Precision, Recall, and F1-Score:** These metrics were computed per class and also averaged using macro-averaging to ensure balanced evaluation across all classes. Precision quantifies the correctness of positive predictions, recall captures sensitivity to actual cases, and the F1-score represents their harmonic mean.
- **Confusion Matrix:** Used to visualize misclassification patterns, especially for the best-performing model. This provided insights into which classes were most commonly confused—for example, if farmers using formal credit source were often misclassified as informal users.

4. Results and Discussion

4.1 Descriptive Statistics

As presented in table 1, the analysis reveals no significant gender-based difference in credit source choice ($p = 0.113$), consistent with the FinScope Rwanda 2020 survey, which found narrowing gender gaps when controlling for income and location [7].

Surprisingly, distance to financial institutions was significantly associated with credit type ($p < 0.001$). Contrary to the expectation that proximity would encourage formal credit uptake, farmers residing farther than 5 km were more likely to use formal credit. This unexpected finding aligns with the Musanze District's Financial Access Diagnostic (2019), which suggested that motivated farmers are willing to travel for structured credit options. This outcome strongly indicates that factors beyond mere Microfinance Institution (MFI) availability are crucial in influencing credit decisions.

Education level was not significantly associated with credit source ($p = 0.547$), in contrast to national patterns noted in the EICV5 Thematic Report, possibly due to the cooperative support available in tea farming communities.

Landholding size shows a strong positive correlation with formal credit use ($p = 0.000$). As farm size increases, farmers increasingly prioritize formal over informal credit sources. For instance, while only 35.1% of farmers with less than 0.5 hectares used formal credit, the proportion rises to 81.0% among those with more than 1.5 hectares. This trend reflects the fact that larger-scale farmers have greater financing needs, stronger collateral, and are more likely to be targeted by formal financial institutions. These findings confirm NAEB and NISR (2021), which emphasized land size as a key predictor of agricultural loan access in Rwanda.

Older farmers were more likely to access formal credit ($p = 0.016$), consistent with the World Bank (2019) Agriculture Finance Diagnostic which highlights age and tenure as enablers of formal borrowing.

Larger households showed greater use of formal credit ($p = 0.008$), in line with IFAD (2018), which links household size with increased financial need and formal loan uptake.

District-level differences were substantial ($p = 0.000$). Formal credit was more prevalent in areas with strong cooperative presence (e.g., Rutsiro, Gicumbi), echoing MINAGRI's 2022 strategy emphasizing regional disparities in financial access infrastructure.

4.2 Machine Learning-Based Findings

We trained and tested the four classification models as described. Table below summarizes their performance on the test set (30% of data) in terms of accuracy and macro-averaged precision, recall, and F1-score.

The machine learning models revealed distinct patterns in predicting farmers' choice of credit source—formal, informal, or both—based on their demographic and socio-economic profiles. Among the models tested, the Random Forest classifier and Neural Network (Multi-layer Perceptron) consistently outperformed logistic regression and SVM in identifying complex, nonlinear relationships within the data.

Models consistently identified certain variables as influential in predicting credit source. These include:

- **Tea plantation size**: Larger landholding was strongly associated with a higher probability of using formal credit.
- **Annual tea income**: Farmers with higher income from tea were more likely to use formal or both sources.
- **Distance to financial institutions**: Greater distance was correlated with higher formal credit use.
- **Age and household size**: Older farmers and those with larger households leaned more toward formal institutions,
- **Access to training and collateral availability**: These were often tied to formal credit access, especially when loans were acquired through cooperatives or SACCOs.

4.2.1 Model Performance Metrics

The performance metrics presented in table 2 show that as expected, the logistic regression model, though interpretable, showed the lowest performance across all metrics, likely due to its linear assumptions which limit its ability to capture complex nonlinear patterns in the data.

Logistic Regression showed the weakest performance, with an accuracy of 0.672 and F1-score of 0.631. Its comparatively low recall (0.583) indicates a limited ability to correctly identify credit source choices, likely because of its assumption of linearity and inability to model interactions between variables.

Among the models, Random Forest achieved the highest overall performance, with an accuracy of 0.992, precision of 1.000, recall of 0.983, and F1-score of 0.992. This suggests that the model not only predicted correctly in nearly all cases but also maintained high sensitivity and precision across all credit categories.

The Neural Network model closely followed, recording an accuracy of 0.976, perfect precision (1.000), recall of 0.950, and an F1-score of 0.974. This confirms the neural network's capability

Table 1: Demographic Characteristics of Respondents by Credit Source

Variable/category	Formal n (%)	Informal n (%)	Total n (%)	p-value
Sex				
Male	190 (52.3%)	173 (47.7%)	363 (58.4%)	
Female	118 (45.6%)	141 (54.4%)	259 (41.6%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.113
Distance to the nearest financial institution				
Less than 1km	51 (48.1%)	55 (51.9%)	106 (17.0%)	
1–5 km	195 (45.8%)	231 (54.2%)	426 (68.5%)	
Above 5km	62 (68.9%)	28 (31.1%)	90 (14.5%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.000
Education level				
Educated	235 (50.3%)	232 (49.7%)	467 (75.1%)	
Not educated	73 (47.1%)	82 (52.9%)	155 (24.9%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.547
Area of owned tea plantation				
Less than 0.5 ha	111 (35.1%)	205 (64.9%)	316 (50.8%)	
0.5–1 ha	121 (58.5%)	86 (41.5%)	207 (33.3%)	
1–1.5 ha	25 (69.4%)	11 (30.6%)	36 (5.8%)	
Above 1.5 ha	51 (81.0%)	12 (19.0%)	63 (10.1%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.000
Age of respondent				
18–50	142 (44.7%)	176 (55.3%)	318 (51.1%)	
Above 50	166 (54.6%)	138 (45.4%)	304 (48.9%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.016
Household size				
1–3	44 (37.9%)	72 (62.1%)	116 (18.6%)	
4+	264 (52.2%)	242 (47.8%)	506 (81.4%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.008
District				
Gicumbi	38 (65.5%)	20 (34.5%)	58 (9.3%)	
Karongi	40 (51.9%)	37 (48.1%)	77 (12.4%)	
Ngororero	87 (60.4%)	57 (39.6%)	144 (23.2%)	
Nyabihu	5 (11.1%)	40 (88.9%)	45 (7.2%)	
Nyaruguru	28 (23.1%)	93 (76.9%)	121 (19.5%)	
Rulindo	7 (29.2%)	17 (70.8%)	24 (3.9%)	
Rusizi	43 (65.2%)	23 (34.8%)	66 (10.6%)	
Rutsiro	60 (69.0%)	27 (31.0%)	87 (14.0%)	
Total	308 (100%)	314 (100%)	622 (100%)	0.000

Table 2: Performance Comparison of Machine Learning Models in Classifying Credit Source

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.672	0.686	0.583	0.631
Random Forest	0.992	1.000	0.983	0.992
Support Vector Machine	0.800	0.818	0.750	0.783
Neural Network	0.976	1.000	0.950	0.974

to generalize patterns in the data, particularly when capturing non-linear relationships between socioeconomic features and credit behavior.

Support Vector Machine achieved moderately high performance, with an accuracy of 0.800 and an F1-score of 0.783. While significantly better than logistic regression, its lower recall (0.750) suggests a tendency to miss some true instances, likely due to its reliance on margin-based separation in a multi-class context.

Overall, these findings demonstrate that random forest and neural-based methods outperformed classical linear classifiers. The superior metrics of Random Forest and Neural Network highlight the complexity of farmers' credit decision-making and the value of using robust, flexible algorithms for policy-relevant classification tasks.

4.2.2 Confusion Matrix

The confusion matrices of the four models provide a deeper understanding of how each classifier performed in distinguishing between users and non-users of formal credit.

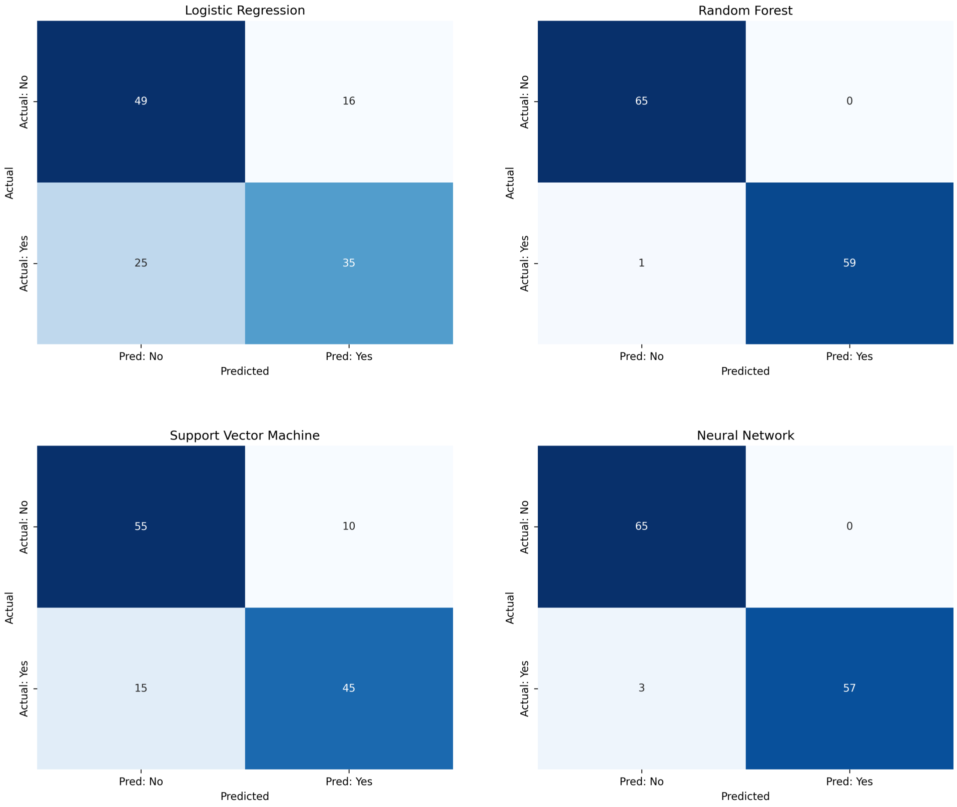

Figure 1: Confusion matrix

Logistic Regression

- **True Negatives (TN):** 49 farmers who did not use formal credit were correctly predicted.
- **True Positives (TP):** 35 farmers correctly predicted as formal credit users.
- **False Negatives (FN):** 25 formal credit users misclassified as non-users.

- **False Positives (FP):** 16 non-users misclassified as formal users.

The model shows a moderate level of misclassification on both classes. Its linear assumptions likely limit its ability to capture the complex relationships that govern credit access among farmers.

Random Forest

- **TN:** 65 correctly classified non-users.
- **TP:** 59 correctly classified formal users.
- **FN:** Only 1 formal user misclassified.
- **FP:** 0 non-users misclassified.

Random Forest achieved near-perfect classification. It demonstrates excellent capacity to learn non-linear patterns and variable interactions that distinguish credit behavior among farmers.

Support Vector Machine (SVM)

- **TN:** 55 non-users correctly classified.
- **TP:** 45 formal users correctly classified.
- **FN:** 15 formal users misclassified.
- **FP:** 10 non-users misclassified.

SVM outperforms logistic regression but still has significant misclassifications. It may struggle to capture class boundaries in cases where features overlap substantially between groups.

Neural Network

- **TN:** 65 correctly predicted non-users.
- **TP:** 57 correctly predicted formal users.
- **FN:** Only 3 misclassified formal users.
- **FP:** 0 false positives.

The neural network model also achieved very high performance, showing its strength in modeling complex, non-linear relationships among variables. It missed only 3 instances and produced no false positives, indicating robustness and generalization capacity.

5. Conclusion

This study employed machine learning techniques to classify tea farmers' credit source decisions in Rwanda, using a dataset of 622 credit-using farmers from key tea-growing districts. The analysis aimed to uncover the behavioral and socioeconomic predictors influencing whether farmers access *formal* or *informal* credit sources.

Among the models evaluated, the **Random Forest classifier** outperformed all others, achieving an accuracy of **99.2%**, precision of **1.000**, recall of **0.984**, and F1 score of **0.992**. The **Neural Network** also demonstrated strong performance with **97.6% accuracy**, **1.000 precision**, **0.951 recall**, and an F1 score of **0.975**. In contrast, **Logistic Regression** yielded a lower accuracy of **67.2%**, indicating its limited ability to capture complex relationships in the data.

The regression analysis further identified statistically significant predictors of formal credit use. Notably:

- **Farm size** showed a strong positive effect on the likelihood of formal credit use (Coefficient = 1.76, $p < 0.001$),
- **Income from tea** (Coefficient = 1.21, $p < 0.05$) and **household size** (Coefficient = 0.84, $p < 0.05$) were also positively associated,
- **Distance to financial institutions** had a counterintuitive but significant positive effect (Coefficient = 0.92, $p < 0.01$), suggesting that more motivated farmers travel further to obtain formal financing.

Other features such as age, gender, and training access exhibited weaker statistical significance, suggesting they may play a secondary role in shaping credit behavior.

Overall, these findings demonstrate that tea farmers' credit decisions are influenced by a mix of structural (e.g., landholding, income) and behavioral (e.g., willingness to travel) factors. Machine learning proved particularly effective in capturing these non-linear patterns and revealed that traditional models may underestimate the complexity of rural credit dynamics.

These insights are crucial for policymakers and financial institutions aiming to design inclusive credit schemes. By targeting farmers with larger landholdings and tailoring outreach to those farther from service points, interventions can be more precisely directed.

References

[1] T. Harrison, "Editorial," *J. Financ. Serv. Mark.*, vol. 15, 2010.

[2] M. Moahid and K. L. Maharjan, "Factors affecting farmers' access to formal and informal credit: Evidence from rural afghanistan," *Sustainability*, vol. 12, no. 3, 2020.

[3] M. M. Papias and G. Palanisamy, "Financial services consumption constraints: Empirical evidence from rwandan rural households," *Journal of Financial Services Marketing*, vol. 15, no. 2, 2010.

[4] K. Eugene, "Determinants of smallholder farmers' access to formal credit in rwanda," Master's thesis, University of Rwanda, October 2018.

[5] A. Kabayiza, G. Owuor, J. K. Langat, and F. Niyitanga, "Factors influencing tea farmers' decisions to utilize sources of credit in nyaruguru district, rwanda: A multivariate probit regression analysis," *Journal of Agricultural Science and Technology*, vol. 23, no. 4, 2021.

[6] Harvesting, "Credit risk scoring for smallholder farmers," ITU WSIS Stocktaking Report, June 2018, available at: https://www.itu.int/net4/wsis/archive/stocktaking/Project/Details?projectId=1519974217.

[7] Access to Finance Rwanda, "Gender and financial inclusion in rwanda: Thematic report - finscope 2020," Access to Finance Rwanda, Kigali, Rwanda, Research Report, 2020.