ARTIFICIAL INTELLIGENCE AND FINANCIAL INCLUSION IN SELECTED EAST AFRICAN COUNTRIES

Adesanya, Adeniyi Olayinka[1] **Emilia, Vann Yaroson**[2]
[1]*School of Management, University of Bradford, Bradford, West Yorkshire, United Kingdom*
[2]*Operations and Supply Chain Management Sheffield University Management School*
***Emails**: adeniyigenius@gmail.com[1]; e.v.yaroson@sheffield.ac.uk[2]*

Abstract
The study addresses the use of artificial intelligence in predicting financial inclusion in East Africa, particularly the lack of bank accounts which negatively impact development and livelihood. To address this issue, the study proposes the use of machine learning techniques to predict individuals who own bank accounts and those who do not. The study employed various machine learning algorithms, including Logistic Regression, Naive Bayes, Random Forest Classifier, Decision Tree, Gradient Boosting, XGBoost, SVM, and K-Nearest Neighbours, to predict financial inclusion in East Africa. A public dataset obtained from Kaggle was used to predict whether an individual has a bank account or not. The study built a machine learning model that predicts financial inclusion in East Africa using individuals' demographic and economic information. The Gradient Boosting model outperformed other models, with a mean accuracy score of 0.89. Feature importance analysis revealed that Level of Education was the most significant predictor of financial inclusion, followed by Type of Job and the relationship with head. The study highlights the most essential factors in promoting financial inclusion in East Africa and provides insights for policymakers and financial institutions to improve access to financial services. Based on the results of the study, it was recommended that policymakers and financial institutions in East Africa should focus on improving financial inclusion for individuals with a higher level of education, especially those with vocational or specialized training and secondary education. Additionally, efforts should be made to improve access to cell phone technology, particularly in urban areas.

Key words: *Artificial Intelligence, Financial Inclusion, Random Forest Classifier, Gradient Boosting, XGBoost, K-Nearest Neighbours.*

Introduction

Artificial Intelligence (AI), first proposed by McCarthy in 1956, has grown into a major field of computer science focused on creating intelligent systems that mimic human reasoning. Its impact is evident in global finance and technology, where it enables faster, cheaper, and more efficient transactions. African countries, including those in East Africa, have increasingly adopted AI, a development especially important given that an estimated 1.7 billion people worldwide remain unbanked (Manyika et al., 2016). Reports from the McKinsey Global Institute highlight AI's potential to address key barriers to financial inclusion in East Africa, such as difficulties in identity verification and the lack of conventional data for underwriting marginalized populations. Fintech companies now use AI-driven alternative credit scoring, relying on digital footprints to expand financial access, as seen in countries like Bangladesh and Pakistan. AI also enhances customer service within financial institutions. For example, UBA's AI chatbot, Leo, provides instant assistance for transactions across platforms like WhatsApp and Facebook Messenger. More broadly, AI continues to influence financial inclusion and reporting by improving efficiency and accessibility (Mhlanga, 2020). Given the large unbanked population in East Africa, examining AI's role in financial inclusion is essential for tackling poverty and reducing economic inequality in the region.

Statement of the Problem

The advent of digital finance is gradually making traditional mode of financial transactions becoming obsolete. The emergence of digital finance and Fin Tech has been used to solve many financial payment problems as it is observed that the financial world faces the problem of ineffective and unsecure payment systems digitally (The World Bank Group, 2020). Financial inclusion is a need of the hour because the nature of business transactions has changed from the traditional form of payments to the digital form of payments (Demirgüç-Kunt and Singer 2017). Artificial Intelligence (AI) has been identified as a powerful tool for enhancing financial inclusion by enabling predictive analytics and reducing transaction costs, among other benefits (Fowler et al., 2020). Therefore, this study aims to fill a gap in existing literature by exploring how AI can help predict financial inclusion in East Africa and contribute to the advancement of financial inclusion.

Objectives

The purpose of this study is to develop an AI-based model that predicts financial inclusion in East Africa. The study involves preprocessing the data, balancing class distributions, splitting the data into training and testing sets, scaling the features, and evaluating multiple machine learning models to identify the best performer. After fine-tuning the selected model, its performance will be assessed on test data. The objective is to determine whether individuals have access to financial services using demographic, economic, and other relevant indicators. The study also aims to identify the key factors influencing financial inclusion, providing insights that can guide policymakers and financial institutions in improving financial access across the region.

1. To develop an Artificial Intelligence–based machine learning model capable of accurately predicting financial inclusion in East Africa using demographic and

economic factors, and to identify the key features that most significantly influence these predictions.

2. RQ2: To evaluate and compare the performance of various machine learning algorithms in predicting financial inclusion and determine the best-performing model for practical application in the region.

This study develops a machine learning model to predict financial inclusion in East Africa using demographic and economic data. Focusing on Kenya, Rwanda, Tanzania, and Uganda, it examines factors influencing financial inclusion and explores how AI can improve access to financial services. The process includes data preprocessing, balancing class distributions, selecting and tuning multiple machine learning models, and evaluating their performance. Algorithms such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, SVM, KNN, and Naive Bayes are compared to identify the most accurate model. Insights from the predictions aim to support policymakers and stakeholders in enhancing financial inclusion across the region. The study is limited by the scope of the available dataset and may not be fully generalizable beyond East Africa.

Literature Review

Conceptual Review:

This literature review uses the Demirgüç-Kunt and Singer (2017) framework, focusing on financial inclusion and mobile money in East Africa. Digital finance, driven by emerging technologies, has transformed the financial sector by enabling consumers and small businesses to access services through mobile and online platforms (Schuetz & Venkatesh, 2020). Innovations such as blockchain have further expanded financial opportunities for SMEs (Panova et al., 2020). Research consistently shows that mobile money significantly enhances financial inclusion across East Africa, particularly in remote areas lacking traditional banking services (Dinku et al., 2018; Awad et al., 2018; Ozili, 2021). Financial literacy also plays an essential role, as individuals with higher education and income levels are more likely to adopt mobile money (Ahamed & Mallick, 2019). Key enablers of financial inclusion include infrastructure, mobile money agents, and financial knowledge, though their influence varies across different demographic groups (Ahmad et al., 2020; Sha'ban et al., 2020).

Artificial intelligence further expands financial inclusion through FinTech innovations. FinTech development has evolved in three phases: infrastructure building (1886–1967), digital finance transformation marked by the first ATM (1967–2008), and post-2008 expansion driven by regulatory change and mistrust in traditional banks (Arner et al., 2015; Leong & Sung, 2018). These advancements continue to shape various financial sectors, particularly asset management.

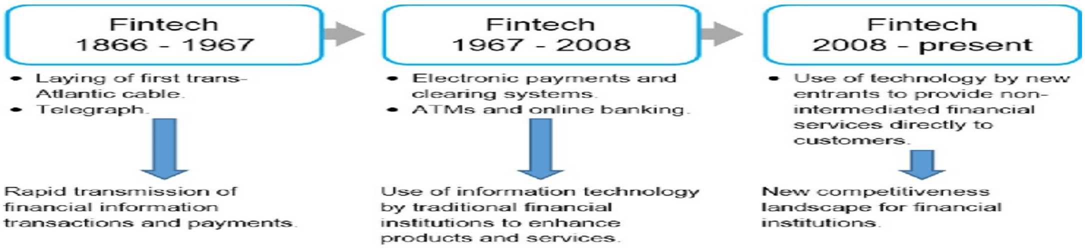

Figure 1: Three Phases of FinTech Development (Thakor 2020)

Artificial Intelligence and Machine Learning

Artificial intelligence (AI) and machine learning (ML) are transforming financial inclusion in developing regions by offering more efficient ways to serve low-income individuals and small businesses traditionally underserved by banks due to high transaction costs. AI-driven solutions such as digital lending, alternative credit scoring, and automated customer support help financial institutions assess creditworthiness using non-traditional data and deliver services through digital channels (Kshetri, 2019). This is especially valuable in developing economies where many lack formal credit histories.

Mobile money has become a major driver of financial inclusion in East Africa, providing access to financial services in remote areas (Dinku et al., 2018; Awad et al., 2018). Its uptake is strengthened by factors such as financial literacy, availability of agents, and infrastructure, though adoption varies across demographics, with urban areas and men showing higher usage (Ahamed & Mallick, 2019; Sha'ban et al., 2020). FinTech innovations, fuelled by AI, continue to expand digital finance, following global technological evolution from early infrastructure building to modern digital and post-crisis FinTech growth. However, challenges exist. AI systems require large, high-quality datasets, yet data gaps, hidden biases, and weak regulatory frameworks can undermine fairness and consumer protection (Sundblad, 2018; Harkut & Kasat, 2019). Financial institutions also hesitate to rely fully on automated systems without human oversight, creating tension between automation and risk management (Deloitte, 2018).

The theoretical foundation for analysing financial inclusion in East Africa can be explained through the Theory of Technological Diffusion, which suggests that new technologies spread more quickly in populations with stronger infrastructure and literacy (Lai, 2017). This aligns with empirical findings showing higher mobile money adoption in urban areas and among men in Kenya, Rwanda, Tanzania, and Uganda (Central Bank of Kenya, 2018; National Bank of Rwanda, 2019). Mobile money remains a critical driver of financial inclusion across the region (Jacobs et al., 2018).

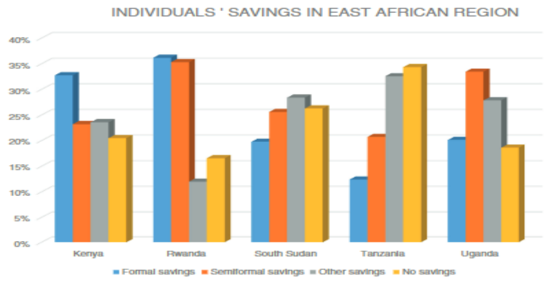

Figure.2: Individuals Saving in East Africa (Lotto 2018)

Financial inclusion has grown significantly across East Africa due to the rapid adoption of mobile money. In Kenya, access to financial services increased from 30% in 2006 to 86.2% in 2018, with over 90% of adults using mobile money. Rwanda shows similar trends, with financial inclusion rising to 60.9% in 2018 and nearly 90% of adults using mobile money. Tanzania and Uganda also report high mobile money usage over 85% and 80% respectively which has helped low-income households and small businesses access financial services. Recent empirical studies have explored machine learning as a tool for predicting financial inclusion. Algorithms such as decision trees, random forests, and SVMs have shown strong predictive performance, though limitations exist due to small datasets and varying country contexts. AI-based models have also been used successfully in other countries, suggesting opportunities for targeted financial interventions. Research further shows that mobile money adoption depends on infrastructure, agent networks, financial literacy, and demographic factors. Urban populations and men consistently exhibit higher adoption rates. Mobile money has improved access to savings, credit, and remittances, reduced transaction costs, and increased transaction speeds.

Literature also highlights the importance of digital financial inclusion, as traditional banking often excludes low-income populations. AI and big data enable alternative credit scoring using non-traditional information, helping expand access. However, challenges such as data quality, algorithmic bias, regulatory gaps, and reluctance to rely fully on automation still hinder AI’s full potential. Studies in other regions show that factors such as education, age, gender, income, and regulatory environments also influence financial inclusion. Machine learning models have proven effective for credit risk assessment, borrower classification, and predicting loan defaults, often outperforming traditional statistical methods.

Methodology

This study used Kaggle survey data from Kenya, Rwanda, Tanzania, and Uganda, covering 23,525 respondents and 13 attributes. Data preprocessing includes cleaning, feature selection, handling missing values, encoding categorical variables, balancing classes, splitting into training/testing sets, and scaling features. Several machine learning algorithms including Logistic Regression, Naive Bayes, Random Forest, Decision Tree, Gradient Boosting, XGBoost, SVM, KNN, and Multinomial Logistic Regression are developed and evaluated through cross-validation. The best-performing model is tuned using hyperparameter optimization and assessed using accuracy, precision, recall, and F1-score.

Data Analysis

This involves identifying key features that drive financial inclusion and evaluating model predictions. Descriptive statistics summarize demographic and socio-economic characteristics such as gender, education, age, location, and access to financial services.

```
financial_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23524 entries, 0 to 23523
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   country                  23510 non-null  object
 1   year                     23524 non-null  int64
 2   uniqueid                 23524 non-null  object
 3   Has a Bank account       23488 non-null  object
 4   Type of Location         23509 non-null  object
 5   Cell Phone Access        23513 non-null  object
 6   household_size           23496 non-null  float64
 7   Respondent Age           23490 non-null  float64
 8   gender_of_respondent     23490 non-null  object
 9   The relathip with head   23520 non-null  object
 10  marital_status           23492 non-null  object
 11  Level of Educuation      23495 non-null  object
 12  Type of Job              23494 non-null  object
dtypes: float64(2), int64(1), object(10)
memory usage: 2.3+ MB
```

```
financial_data.shape
```

```
(23524, 13)
```

Figure 3 Data information

Results

The output presented provides a summary of all columns in the dataset, detailing their respective data types and the count of non-null values for each column. Furthermore, I was able to determine the presence of null values within the dataset as we know that the total number of rows is 23,524, with a specific count of 13 rows containing null values.

Table 1 data showing total number of Columns and rows.

No.	Variable/Columns	Data Type	Context
0.	Country	Object	Uganda, Tanzania, Kenya, Rwanda
1.	Year	Int64	2016 – 2018
2.	Uniqueid	Object	Identifier specific to each interviewee
3.	Has a Bank account	Object	Yes, No
4.	Type of location	Object	Urban, rural
5.	Cell Phone Access	Object	If the interviewee is in possession of a cell phone: Yes/No
6.	Household size	Float64	number of residents per household
7.	Respondent Age	Float64	The interviewee's age
8.	Gender of respondent	Object	Male or Female
9.	The relationship with Head	Object	Relationship with Other non-relatives, Spouse, Head of Household, Child/Children, Parent, Other relative, Don't know
10.	Marital status	Object	Divorced/Separated, Single/Never Married, Widowed, Married/Living together, Don't know.
11.	Level of Education	Object	Vocational/Specialised training, no formal education, Primary education, Other/Dont know/RTA, Secondary education, Tertiary education.
12.	Type of job	Object	Government Dependent, Formally employed Private, Self-employed, Remittance Dependent, Formally employed Government,, Informally employed, Other Income, Farming and Fishing, No Income, Don't Know/Refuse to answer.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves visually examining a dataset to extract significant insights. It is a powerful approach that enables us to obtain the most comprehensive understanding of the dataset by utilizing diverse charts and graphs.

Building our model

```
# Split the dataset into features and target variable
X = xp2
y = y2

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 0

The dataset was splitted into features and target variables.

```
# Compare the algorithms using boxplots
import matplotlib.pyplot as plt
fig = plt.figure()
fig.suptitle('Algorithm Comparison on Financial Dataset')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

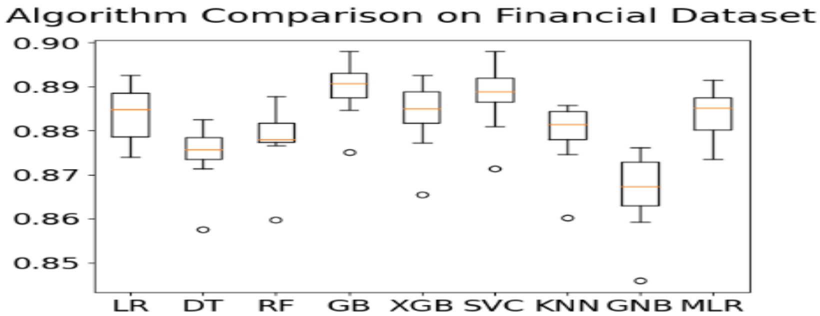

Figure 5 Boxplot Visual representation of Algorithm Performance on Financial Dataset.

Result.

The resulting visualization displays the distribution of results for each algorithm, allowing for a visual comparison of their performance on the financial dataset. The box plot shows that Gradient Boosting classifier has the best performance.

Precision: Precision measures how many of the instances predicted as positive are actually positive. High precision means few false positives**.**

Recall: Recall measures how many of the actual positive instances a model correctly identifies. High recall means the model misses very few positive cases**.**

F1 Score: The F1 score is the harmonic mean of precision and recall: F1 = 2 × (precision × recall) / (precision + recall)

It balances both metrics and is useful when precision and recall are equally important**.**

Accuracy: Accuracy is the proportion of all correct predictions (both positive and negative) out of all instances. Accuracy = (TP + TN) / (TP + FP + TN + FN)

It shows overall correctness but can be misleading when classes are imbalanced**.**
Support: Support is the number of data samples belonging to each class. It helps interpret evaluation metrics, especially in imbalanced datasets where a class with low support may result in unreliable performance scores**.**

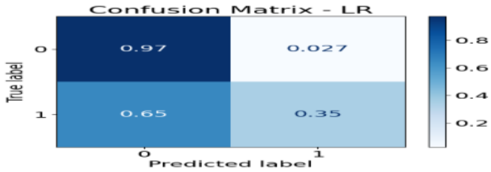

Figure 6 *Model confusion matrix showing target variable class performance.*

The model properly anticipated most of the data, as was demonstrated above. All indicators for evaluation are excellent. It was observed that false negative (FN), or the proportion of real values that are accurate but were projected to be negative. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.35, 0.97, 0.65, and 0.027, respectively.

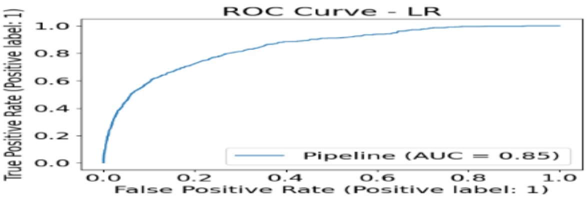

Figure 7 ROC curve for Logistic Regression.
The area under curve (AUC) value indicates how well a model performs and 85% of the respondent satisfaction level can be reliably predicted by this model.

LR

	precision	recall	f1-score	support
0	0.90	0.97	0.93	4033
1	0.68	0.35	0.46	672
accuracy			0.88	4705
macro avg	0.79	0.66	0.70	4705
weighted avg	0.87	0.88	0.87	4705

Figure 8 Accuracy performance for Logistic Regression.
The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.35, 0.97, 0.65, and

0.031, respectively.

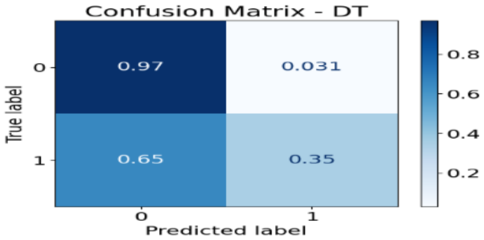

Figure 9 Model confusion matrix showing target variable class performance.

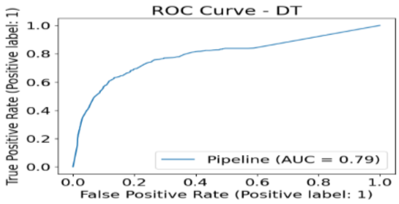

Figure 10 ROC Curve for Decision Tree

DT

	precision	recall	f1-score	support
0	0.90	0.97	0.93	4033
1	0.65	0.35	0.46	672
accuracy			0.88	4705
macro avg	0.78	0.66	0.69	4705
weighted avg	0.86	0.88	0.87	4705

Figure 11. Accuracy performance for Decision Tree

The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.37, 0.97, 0.63, and 0.033, respectively.

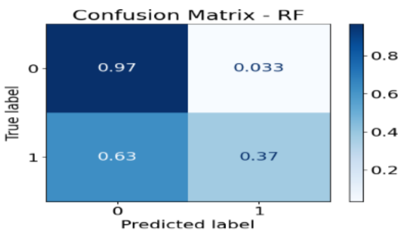

Figure 12 Model confusion matrix showing target variable class performance.

The area under curve (AUC) value indicates how well a model performs and 84% of the respondent satisfaction level can be reliably predicted by this model.

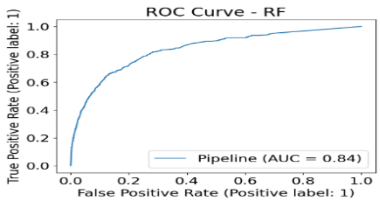

Figure 13 ROC for Random Forest Classier

RF

	precision	recall	f1-score	support
0	0.90	0.97	0.93	4033
1	0.65	0.36	0.47	672
accuracy			0.88	4705
macro avg	0.78	0.67	0.70	4705
weighted avg	0.87	0.88	0.87	4705

Figure 14 Accuracy performance for Random Forest Classier

The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.34, 0.98, 0.66, and 0.025, respectively.

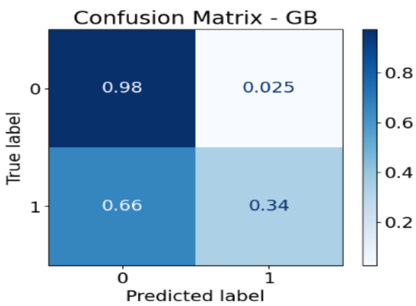

Figure 15 Model confusion matrix showing target variable class performance.

The area under curve (AUC) value indicates how well a model performs and 86% of the respondent satisfaction level can be reliably predicted by this model.

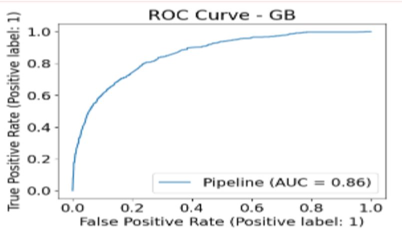

Figure 16 ROC curve for Gradient Boosting Classifier

GB

	precision	recall	f1-score	support
0	0.90	0.98	0.94	4033
1	0.69	0.34	0.45	672
accuracy			0.88	4705
macro avg	0.80	0.66	0.69	4705
weighted avg	0.87	0.88	0.87	4705

Figure 17 accuracy for Gradient Boosting Classifier

The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.37, 0.97, 0.63, and 0.03, respectively.

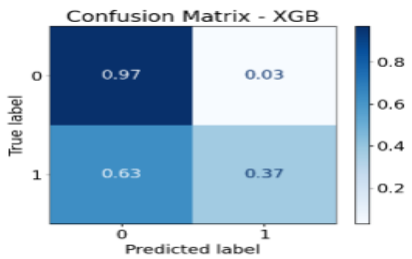

Figure 18 Model confusion matrix showing target variable class performance.

The area under curve (AUC) value indicates how well a model performs and 86% of the respondent satisfaction level can be reliably predicted by this model.

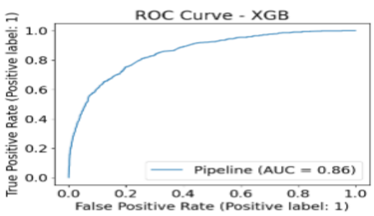

Figure 19 ROC Curve for XGBoosting classifier

XGB

	precision	recall	f1-score	support
0	0.90	0.97	0.93	4033
1	0.67	0.37	0.48	672
accuracy			0.88	4705
macro avg	0.79	0.67	0.71	4705
weighted avg	0.87	0.88	0.87	4705

Figure 20 Accuracy performance for XGBoosting classifier

The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.33, 0.98, 0.67, and 0.023, respectively.

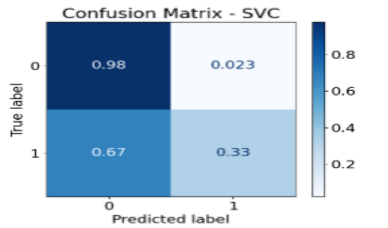

Figure 21 Model confusion matrix showing target variable class performance

The area under curve (AUC) value indicates how well a model performs and 79% of the respondent satisfaction level can be reliably predicted by this model.

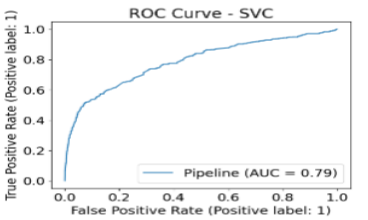

Figure 22 ROC Curve for Support Vector Classifier

SVC

	precision	recall	f1-score	support
0	0.90	0.98	0.94	4033
1	0.70	0.33	0.45	672
accuracy			0.88	4705
macro avg	0.80	0.65	0.69	4705
weighted avg	0.87	0.88	0.87	4705

Figure 23. Accuracy performance for Support Vector Classifier

The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.37, 0.96, 0.63, and 0.042, respectively.

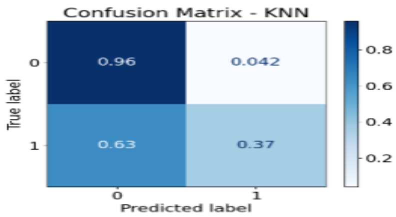

Figure 24 Model confusion matrix showing target variable class performance.

The area under curve (AUC) value indicates how well a model performs and 79% of the respondent satisfaction level can be reliably predicted by this model.

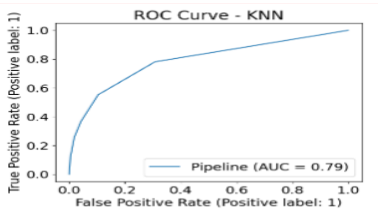

Figure 25 ROC Curve for K- Nearest Neighbour Classifier

KNN

	precision	recall	f1-score	support
0	0.90	0.96	0.93	4033
1	0.59	0.37	0.45	672
accuracy			0.87	4705
macro avg	0.75	0.66	0.69	4705
weighted avg	0.86	0.87	0.86	4705

Figure 26 Accuracy performance for K- Nearest Neighbour Classifier

The false negative (FN), or number of true actual values but were projected to be negative, can be seen in the confusion matrix below. A false positive (FP) occurs when a value is projected to be positive but really turns out to be negative. The true positive (TP) displays the positively predicted instances, whereas the true negative (TN) displays the negatively predicted instances. The values for True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.48, 0.93, 0.52, and 0.066, respectively.

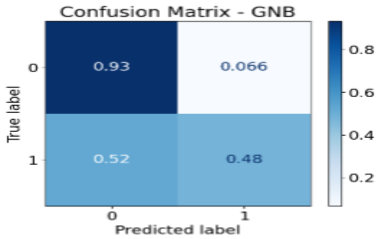

Figure 27 Model confusion matrix showing target variable class performance.

This model predicts 83% of respondent satisfaction using its AUC value.

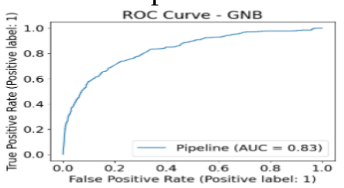

Figure 28 ROC Curve for Gaussian NB

GNB

	precision	recall	f1-score	support
0	0.91	0.93	0.92	4033
1	0.55	0.48	0.51	672
accuracy			0.87	4705
macro avg	0.73	0.71	0.72	4705
weighted avg	0.86	0.87	0.87	4705

Figure 29 Accuracy performance for Gaussian NB

The model anticipated most of the data properly. Starting with accuracy, precision, recall, and F1 score, all evaluation indicators are excellent.

False negatives (FN) are visible in the confusion matrix below. False positives (FPs) occur when the actual value is negative yet expected to be positive. True positive (TP) and true negative (TN) reflect correctly predicted positive and negative events, respectively. The values of True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are 0.33, 0.98, 0.67, and 0.023, respectively.

Figure 30 Model confusion matrix showing target variable class performance.

The area under curve (AUC) value indicates how well a model performs and 85% of the respondent satisfaction level can be reliably predicted by this model.

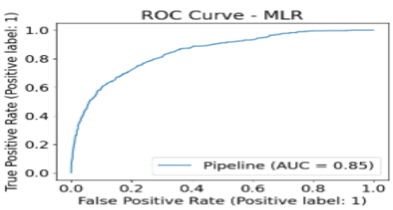

Figure 31 ROC Curve for Multinomial Logistic Regression

MLR

	precision	recall	f1-score	support
0	0.90	0.98	0.94	4033
1	0.70	0.33	0.45	672
accuracy			0.88	4705
macro avg	0.80	0.65	0.69	4705
weighted avg	0.87	0.88	0.87	4705

Figure 32 Accuracy performance for Multinomial Logistic Regression

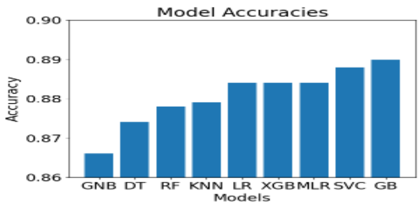

Figure 33 Model Accuracies performance

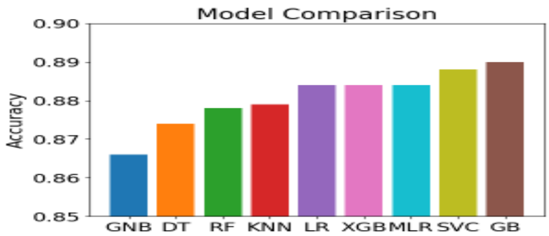

Figure 34 Model Comparison performance

As we can see from the Model Comparison and Model Accuracies, we have discovered that the Highest performed model is Gradient Boosting as Those that achieve 88.96% accuracy. We used machine learning techniques to estimate whether a person has a bank account. Some classifier algorithms are Logistic Regression, Naive Bayes, Random Forest Classifier, Decision Tree, Gradient Boosting, XGBoost, SVM, and K-Nearest Neighbours. The highest and outperformed model is Gradient Booting Model.

Figure 35 Correlation Chart of Performance.

We can see that the variables most highly correlated to our y- varaiable are:

```
Has a Bank account_Yes                                    1.000000
Level of Educuation                                       0.358824
Type of Job_Formally employed Private                     0.249308
Level of Education_Tertiary education                     0.241975
Type of Job_Formally employed Government                  0.236078
Level of Education_Vocational/Specialised training        0.232390
Cell Phone Access_Yes                                     0.209355
Level of Education_Primary education                      0.174249
Level of Education_No formal education                    0.142239
Level of Education_Secondary education                    0.123879
gender_of_respondent_Male                                 0.116338
The relathip with head_Head of Household                  0.114282
Type of Job_Informally employed                           0.098749
country_Tanzania                                          0.088456
Type of Location_Urban                                    0.087880
marital_status_Married/Living together                    0.086266
The relathip with head_Spouse                             0.060539
country_Rwanda                                            0.057519
Type of Job_No Income                                     0.057079
marital_status_Widowed                                    0.052162
The relathip with head_Parent                             0.051127
country_Uganda                                            0.049089
Type of Job_Remittance Dependent                          0.045382
marital_status_Single/Never Married                       0.040798
Type of Job_Farming and Fishing                           0.037750
Level of Education_Other/Dont know/RTA                    0.032331
Type of Job Other Income                                  0.025772
```

Figure 36 Correlation performance of dependent and independent variables

Result.

The above correlation matrix is visualized using a heatmap, with each cell color-coded based on its correlation value. The 'annot=True' parameter adds the numerical values to the heatmap cells. Finally, the title 'Correlation Matrix for Financial Inclusion in East Africa' is added to the heatmap, and it is displayed using the 'plt. show ()' command. The large figure size of (80,60) ensures that the heatmap is displayed in a clear and legible manner.

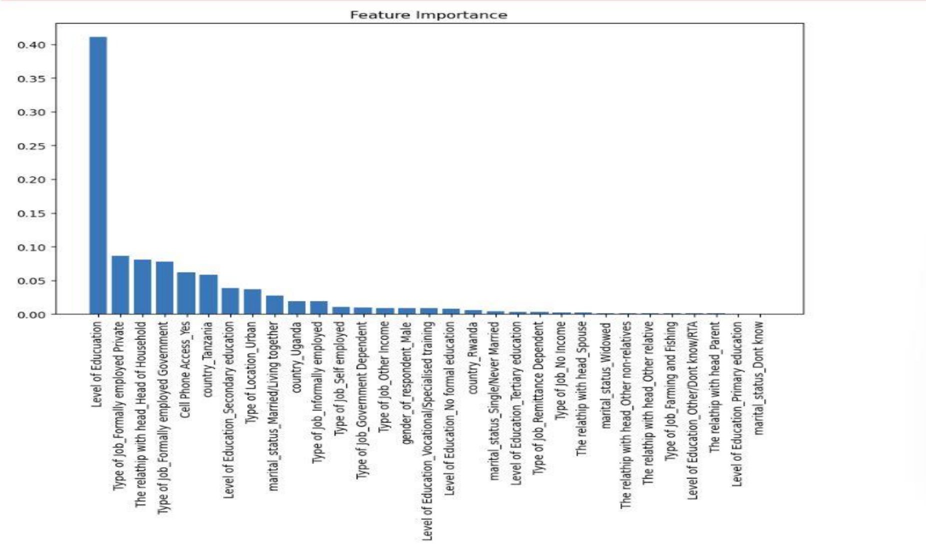

Figure 37 Feature Importance Chart indicators

Summary.

The above feature importance diagram shows that education level predicts financial inclusion.

Discussion and Conclusion

The study developed a machine learning model to predict financial inclusion in East Africa using demographic and economic data. After preprocessing, balancing the dataset, and testing nine machine learning algorithms, Gradient Boosting achieved the highest accuracy (0.889667). Key determinants of financial inclusion included education level, type of employment, cell phone access, and relationship to the household head.

The findings provide actionable insights for policymakers and financial institutions, emphasizing the need to improve access to education, expand mobile phone penetration, and design financial services tailored to different occupational groups. Although Gradient Boosting performed best, other models also showed strong accuracy and offer additional insights for enhancing financial inclusion across the region.

References

Apuke, O. D. (2017). Quantitative research methods: A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, *33*(5471), 1-8.

Arner, D.W., Barberis, J. and Buckley, R.P., 2015. The evolution of Fintech: A new post-crisis paradigm. *Geo. J. Int'l L.*, *47*, p.1271.

Azungah, T., 2018. Qualitative research: deductive and inductive approaches to data analysis. *Qualitative Research Journal*.

Basias, N. and Pollalis, Y., 2018. Quantitative and qualitative research in business & technology: Justifying a suitable research methodology. *Review of Integrative Business and Economics Research*, *7*, pp.91-105.

Bryman, A. and Bell, E. (2011) Business research methods. Oxford University Press, Oxford.

Chu, A. B. 2018. Mobile Technology and Financial Inclusion. In Handbook of Blockchain, Digital Finance, and Inclusion, Volume 1: Cryptocurrency, FinTech, InsurTech, and Regulation. Cambridge: Academic Press, pp. 131–44

Dawei, Liu, Hu Anzi, and Li Gen. 2018. Big Data Technology: Application and Cases. In Handbook of Blockchain, Digital Finance, and Inclusion. Amsterdam: Elsevier Inc., pp. 65–82.

Deloitte. 2018b. Challenges Artificial Intelligence in Financial Services—Deloitte Forward, Deloitte. Available online: https://www.deloitteforward.nl/en/artificial-intelligence/challenges-to-widespreadartificial-intelligence-in-financial-services/ (accessed on 2 June 2020)

Demirgüç-Kunt, A. and Singer, D., 2017. Financial inclusion and inclusive growth: A review of recent empirical evidence. *World Bank Policy Research Working*

Paper, (8040).

Fryer, L.K., Larson-Hall, J. and Stewart, J., 2018. Quantitative methodology. In *The Palgrave handbook of applied linguistics research methodology* (pp. 55-77). Palgrave Macmillan, London.

Harris, C.G., 2018, April. The Risks and Dangers of Relying on Blockchain Technology in Underdeveloped Countries. In *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium* (Pp. 1-4). IEEE.

Hassani, Hossein, Emmanuel Sirimal Silva, Stephane Unger, Maedeh TajMazinani, and Stephen Mac Feely. 2020. Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? Ai 1: 143–55.

Hotchkiss, Griffin, and David Lee Kuo Chuen. 2018. From the Ground Up: The Financial Inclusion Frontier. In Handbook of Blockchain, Digital Finance, and Inclusion. Amsterdam: Elsevier Inc., pp. 405–29.

Hurley, M. and Adebayo, J., 2016. Credit scoring in the era of big data. *Yale JL & Tech.*, *18*, p.148.

Johnston, M.P., 2017. Secondary data analysis: A method of which the time has come. *Qualitative and quantitative methods in libraries*, *3*(3), pp.619-626.

Küfeoglu, S., Liu, G., Anaya, K. and Pollitt, M., 2019. Digitalisation and new business models in energy sector.

Leong, K. and Sung, A., 2018. FinTech (Financial Technology): what is it and how to use technologies to create business value in fintech way? *International Journal of Innovation, Management and Technology*, *9*(2), pp.74-78.

Levin, Richard B., Peter Waltz, and Holly LaCount. 2018. Betting Blockchain Will Change Everything-SEC and CFTC Regulation of Blockchain Technology. In Handbook of Blockchain, Digital Finance, and Inclusion. Amsterdam: Elsevier Inc., pp. 187–212.

Liao, Gaoke, Dequan Yao, and Zhihao Hu. 2020. The Spatial Effect of the Efficiency of Regional Financial Resource Allocation from the Perspective of Internet Finance: Evidence from Chinese Provinces. Emerging Markets Finance and Trade 56: 1211–23.

Li, M., 2015. Moving beyond the linear regression model: Advantages of the quantile regression model. *Journal of Management*, *41*(1), pp.71-98.

Mader, P., 2018. Contesting financial inclusion. *Development and change*, *49*(2), pp.461-483.

McCaffrey, M. and Schiff, A., 2017. Finclusion to fintech: Fintech product development for low-income markets. *Available at SSRN 3034175*.

Mhlanga, D., 2020. Industry 4.0 in finance: the impact of artificial intelligence (ai) on digital financial inclusion. *International Journal of Financial Studies*, *8*(3), p.45.

Ozili, Peterson K. 2018. Impact of digital finance on financial inclusion and stability. Borsa Istanbul Review 18: 329–40

Ozili, P.K., 2020. Theories of financial inclusion. In *Uncertainty and challenges in contemporary economic behaviour*. Emerald Publishing Limited.

Panova, E.A., Kozhanova, A.V. and Tilov, A.A., 2020. The Influence of Digital Technologies' Development on Financing of Small Companies and Startups. *Socio-economic Systems: Paradigms for the Future*, *2*, p.23.

Sanderson, O., 2018. How to Trust Green Bonds: Blockchain, Climate, and the Institutional Bond Markets. In *Transforming Climate Finance and Green Investment with Blockchains* (Pp. 273-288). Academic Press.

Sapovadia, Vrajlal. 2018. Financial Inclusion, Digital Currency, and Mobile Technology. In Handbook of Blockchain, Digital Finance, and Inclusion. Amsterdam: Elsevier Inc., pp. 361–85.

Schuetz, S. and Venkatesh, V., 2020. Blockchain, adoption, and financial inclusion in India: Research opportunities. *International journal of information management*, *52*.

Terrell, S.R., 2022. *Writing a proposal for your dissertation: Guidelines and examples*. Guilford Publications.

Thakor, AV, 2020. Fintech and banking: What do we know? *Journal of Financial Intermediation*, *41*, pp 1-13.

Van den Akker, O.R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J. and Ritchie, S., 2021. Preregistration of secondary data analysis: A template and tutorial. *Meta-psychology*, *5*.

Walter, M. and Andersen, C., 2016. *Indigenous statistics: A quantitative research methodology*. Routledge.