School of Health Sciences
& School of Mathematics and Statistics

February 2024

Financial Inclusion Status for Women and Poor Adults During The COVID-19 Pandemic

Authors:

- Cathy Peng
- Chathurangi Godahewa Gamage
- Lizhu Dong

Financial Inclusion Status for Women and Poor Adults During The COVID-19 Pandemic

Authors:

Cathy Peng
School of Mathematics and Statistics
Email: cpe70@uclive.ac.nz

Chathurangi Godahewa Gamage
School of Mathematics and Statistics
Email: cgo82@uclive.ac.nz

Lizhu Dong
School of Mathematics and Statistics
Email: ldo61@uclive.ac.nz

Abstract:

This comprehensive project report delves into the critical role of financial inclusion amid the COVID-19 pandemic, focusing on its effects on marginalized socio-demographic groups, with a particular emphasis on women and the economically disadvantaged. Leveraging data from the Global Findex database, the study meticulously analyzes how enhanced access to financial services can bolster socio-economic resilience during times of crisis. Through the application of sophisticated statistical methodologies and cutting-edge machine learning techniques, the research not only uncovers pivotal factors driving financial inclusion but also offers profound policy recommendations aimed at mitigating inequality and reinforcing socio-economic stability. The investigation reveals that digital finance serves as a vital tool in empowering vulnerable populations, suggesting that customized financial inclusion strategies are essential for promoting economic empowerment and resilience. The report's findings advocate for a nuanced approach to financial inclusion, emphasizing the necessity of integrating digital finance into broader socio-economic development frameworks to ensure comprehensive support for those most in need during unprecedented challenges.

Keywords:

Financial Inclusion, COVID-19 Pandemic, Socio-Economic Resilience, Global Findex Database, Socio-Demographic Groups, Economic Empowerment, Policy Recommendations, Statistical Methodologies, Machine Learning Techniques, Economic Disadvantage, Financial Services Access, Inequality Mitigation, Vulnerable Populations, Economic Stability.

Table of Contents

1. Introduction 4
2. Data 5
    2.1 Descriptive Statistics 5
    2.2 Missing Data 6
    2.3 Outliers 7
    2.4 Imbalance 8
    2.5 Near-Zero-Variance 9
3. Methodology 9
    3.1 Data retrieval and cleaning 9
    3.2 Data Strategies 10
        - Outlier handling 10
        - Missing value imputation 10
        - Imbalance data treatment 10
        - Handling of high-cardinality columns 10
    3.3 Data Ethics 11
    3.4 Data modelling 11
4. Results 12
    4.1 Basics 12
        - Univariate Analysis 12
        - Bivariate Analysis 13
        - Trivariate Analysis 15
    4.2 Model 16
    4.3 Discussion 17
5. Conclusion 17
    5.1 Fit-for-Purpose 17
    5.2 Problems Throughout the project 18
    5.3 Bias 18
6. Future Work 18
7. Acknowledgements 19
References 19
Appendices 21

1. Introduction

The COVID-19 pandemic has brought to the forefront the critical role that financial inclusion plays in socio-economic resilience. Recognizing this, the Faculty of Health at the University of Canterbury has initiated an important research project entitled "Social Protection as a Social Determinant of Health," as part of the larger "Population Health 2" initiatives. This study is centered around understanding the impact of financial inclusion on the ability of different socio-demographic groups—especially women and those at an economic disadvantage—to withstand and recover from the economic disruptions caused by the pandemic.

Our aim is to dissect the relationship between access to financial services and socio-economic fortitude in the face of COVID-19. We seek to identify how financial tools and resources can support individuals in navigating the financial challenges posed by the pandemic. The focus is particularly on women and economically disadvantaged groups who are often the hardest hit during economic downturns. By shedding light on these relationships, we hope to contribute valuable insights into the role of financial inclusion in strengthening socio-economic stability.

The University of Canterbury's Faculty of Health, known for its multidisciplinary research and teaching, is leading this project with a clear objective: to advance our understanding of financial inclusion as a lever for socio-economic endurance. Our mission extends to addressing the changing socio-economic needs of communities, with a particular emphasis on those who have traditionally been marginalized. In the context of COVID-19, our project specifically examines how financial inclusion can serve as a catalyst for socio-economic recovery and empowerment.

By examining financial inclusion through the lens of gender and economic status, we intend to map out how equitable access to financial services can contribute to the socio-economic robustness of communities during crises. We aim to provide evidence that will guide policymakers in creating more resilient economic systems that can better support vulnerable populations. Our research will delve into the potential of financial inclusion as a strategy for building socio-economic resilience, with the hope of informing future efforts to create more inclusive and robust financial systems.

This project is centered around the primary research question: What is the relationship between financial inclusion indicators and demographic characteristics of individuals, particularly gender and poverty status? Our extensive goal is to dissect the complex interplay between financial behaviors and access, and socio-economic classifications, by analyzing patterns and correlations in financial inclusion metrics. Through this analysis, we aim to develop a predictive model using these financial indicators to identify an individual's gender and poverty status, thereby offering a deeper understanding of socio-economic factors influencing financial inclusion.

This research intends to provide crucial insights for empowerment and inequality reduction through financial services. It aims to guide policy and enhance understanding of financial behaviors in relation to gender and poverty, thereby contributing to socio-economic development and financial inequality discussions. The COVID-19 pandemic heightened the financial system exclusion for already vulnerable groups—such as the poor, elderly, unemployed, and women in specific countries—who, despite previously having limited access to offline financial services, found themselves further marginalized due to restricted access to traditional financial channels during the crisis (Dluhopolskyi et al, 2023).

A narrowing yet persistent gender gap in financial access, with women and poor facing unique barriers such as lack of identification, limited technology access, and distance from financial institutions, and these challenges delay their ability to open and effectively use financial

accounts. Policymakers are urged to intensify efforts to include these underserved groups in the financial sector's evolution (Asli Demirgüç et al., 2021). The COVID-19 crisis has underscored the vital role of digital finance for the poor, particularly in developing countries like Georgia, where remittances via digital payment systems are crucial for many impoverished families. However, the pandemic has led to a significant reduction in remittances due to job losses among migrant, impacting individual and national economic well-being (Tea, 2020).

Financial outreach plays a crucial role in mitigating poverty by counteracting the negative impact of inequality. By addressing disparities in financial services, it enables the poor to manage their finances more effectively, particularly in coping with significant shocks such as the pandemic (Roxana and Mostak, 2020). Financial inclusion's potential to reduce income inequality may differ by country, influenced by factors like economic development level, institutional quality, regulatory frameworks, the characteristics of financial institutions, markets, and instruments, and the specific financial inclusion strategies implemented (Ayse Damir et al, 2020)

Our project uses the Global Findex database (World Bank, 2021), adhering to the World's Bank's Open data terms, ensuring legal, fair, and accurate data use. We commit to non-discrimination and authenticity, prohibiting data alteration and misuse. The focus is on lawful, ethical research practices, forming the ethical foundation of our work. These principles constitute the foundation for the smooth progress of out project, ensuring the. responsible use of data.

One of the major challenges in this project is integrating and analysing diverse datasets to draw a meaningful insight. The technical capabilities required for handling such large and complex datasets are significant. We plan to utilize R and Python, known for their robust data analysis and visualization capabilities, for data cleaning, visualization, and modelling.

The constraints we face are not just in terms of data and technology but also in terms of the broader ethical and operational framework within which we must operate. Our adherence to ethical guidelines and data use policies forms the bedrock of our research methodology.

2. Data

2.1 Descriptive Statistics

This dataset originates from the Global Financial Database and is a comprehensive effort to measure financial inclusivity among women and impoverished adults during the COVID-19 pandemic. Data collection commenced on June 19, 2021, and concluded on February 26, 2023, spanning 123 economies. The dataset comprises responses from approximately 128,000 adults, making it extensive and representative of global financial conditions during this unprecedented period.

The dataset (micro_world_139countries.csv) consists of 128 variables within 143,887 observations and forms the core of comprehensive information. These variables encompass 90 floating-point variables, 35 integer variables, and three string variables, providing diverse data points, from quantitative aspects like age, income, and education to qualitative categorizations such as gender and regional classifications. This diversity is vital for painting a detailed picture of global financial inclusivity and resilience.

The dataset's quality is a prominent feature, encompassing formal and informal financial services and digital payment practices. It provides insights into utilizing these services and delves into behaviours that support financial resilience. This depth is precious for understanding the multifaceted nature of financial inclusivity. We are particularly interested in

variables related to gender, household income, and a range of financial indicators, including access to banking services, savings habits, credit utilization, and digital payment methods. However, it is essential to note that there are areas where data might need to be more robust, particularly in some financial indicators. Steps were taken to ensure data integrity, with rigorous cleaning and pre-processing protocols to address missing data, outliers, and inconsistencies. Despite these measures, users of this dataset should be aware of potential limitations in its scope and depth, which could influence the breadth of conclusions that can be drawn.

2.2 Missing Data

Among the 128 variables in the dataset, the prevalence of missing data varies from zero to high, reflecting numerous factors related to survey design, respondent behaviour, and regional differences. We conducted a missing data analysis and created a bar chart.

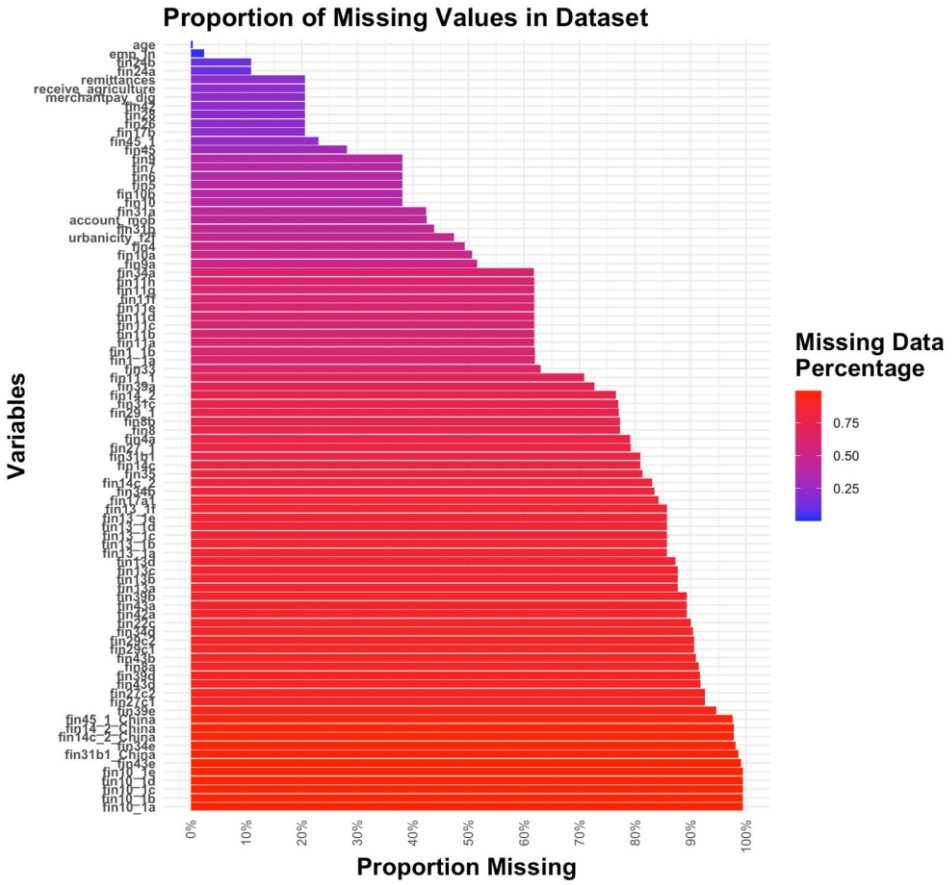

Figure 1: Proportion of Missing Values per Variable

No or low missing data (0% - 5% missing): This category includes essential basic variables and some fundamental financial access and usage variables, such as economic name, gender, education level, income, and more. The missing age rate is also relatively low, indicating occasional omissions during data collection or respondents choosing not to answer specific questions. This portion represents 33% of the total.

Moderate missing rate (5% - 50% missing): This group comprises variables more relevant to specific respondent groups or regional characteristics, such as employment status (emp_in), the use of mobile money accounts (account_mob), and questions related to financial behaviours like credit card and debit card usage. This portion represents 18% of the total.

High Missing Rate (Greater than 50% Missing): This category includes variables that are surveyed only in specific countries or are specific to China. This portion represents 49% of the total. The missing values for variables like fin1_1a, fin1_1b, and over 20 other financial inclusion indicators are not random. They happen because the surveys vary in scope and content from country to country. So, if a survey weren't done in a country, these variables wouldn't have data. It's worth noting that in China, there are also some variables like fin14_2, fin14c_2, fin31b1, and fin45_1, which have different survey question formats compared to other countries. Therefore, Chinese data represent separate variables (e.g., fin14_2_China, fin14c_2_China, fin31b1_China, and fin45_1_China).

The nature and impact of missing data are significant because they can lead to skewed analytical results and potential biases, necessitating special treatment for these missing values.

2.3 Outliers

We employed the boxplot method to identify univariate outliers in variables. Categorical variables such as gender, household income, employment status, account ownership, and mobile payment usage did not reveal any outliers.

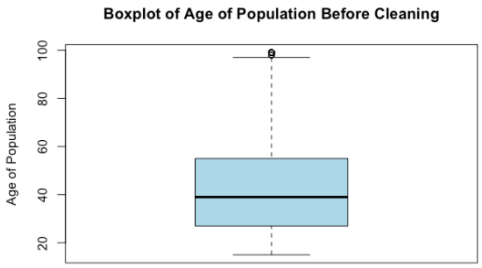

Figure 2: Boxplot of Age distribution (Before Cleaning)

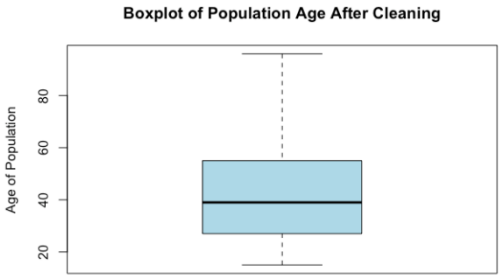

Figure 3: Boxplot of Age distribution (After Cleaning)

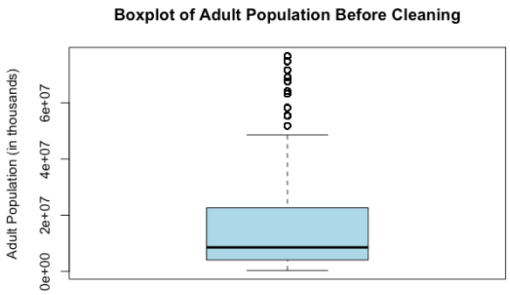

Figure 4: Boxplot of Population distribution (Before Cleaning)

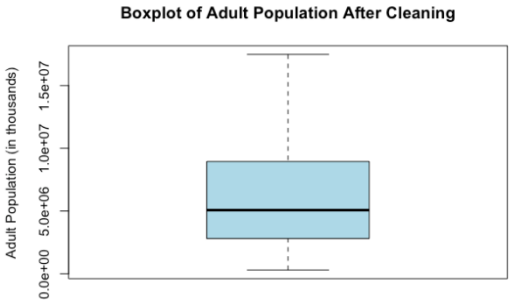

Figure 5: Boxplot of population distribution (After Cleaning)

Our dataset's age variable exhibits the presence of outliers (figure 1) when visualized through a boxplot. These data points, significantly higher than the upper quartile, might not be outliers in the traditional sense but could represent the natural tail of the distribution, indicating individuals with exceptional longevity. It highlights a segment of the population with vast ages, which is an essential aspect of the demographic diversity of our dataset. Alternatively, it is crucial to consider if these points could result from data entry errors, misreporting, or anomalies in the data collection process. A detailed analysis, with insights from domain experts,

is necessary to distinguish between these scenarios. Meanwhile, the boxplot without these extreme values (figure 2) offers a view of age distribution that is more typical for the bulk of the population.

The boxplot representing the adult population data with outliers (figure 3) indicates extreme values deviating from most of the data. These outliers are significantly more significant than the upper quartile, suggesting the existence of some economies with an adult population size that exceeds the norm. Outliers can dramatically impact mean values and variance calculation, resulting in a distorted view of the adult population's overall distribution. The boxplot without outliers (figure 4) provides a depiction of the population distribution that excludes these extremities, which may be more representative of most of the economies in the dataset.

In summary, most categorical variables did not display clear outliers in the boxplots, a positive sign for data quality. However, these data points, such as age, employment status, and other variables, require further data cleaning, validation, and processing to ensure data quality and credibility.

2.4 Imbalance

We have employed two main methods to assess the imbalance in categorical variables: calculating the percentage of each category and using bar charts. The bar chart provides a visual analysis, allowing us to observe the imbalance intuitively. Additionally, calculating the percentage of each category within the entire dataset offers a clearer understanding of the degree of imbalance.

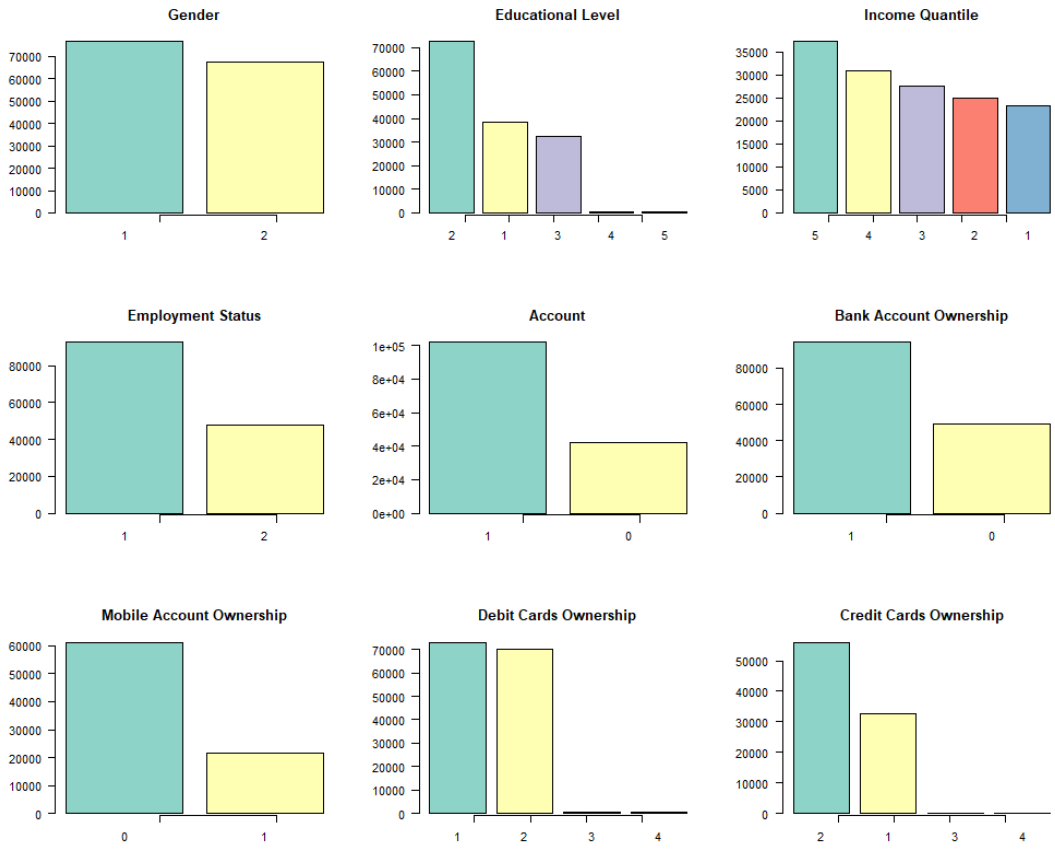

Figure 6: bar charts of key categorical variables

Firstly, for the "Gender" variable, the categories are balanced but slightly skewed towards females. However, the "Education Level" displays a pronounced imbalance, with most respondents falling into the middle education level category. The "Income Quantile" also exhibits a moderate level of imbalance, with a higher representation of respondents in the upper-income quantiles. Regarding employment status, having a job is more prevalent, indicating an imbalance.

Further examination of bank account ownership and account reveals a noticeable imbalance, with more respondents having accounts. Conversely, "Mobile Account Ownership" is heavily imbalanced, with more respondents not having mobile accounts. Ownership of "debit cards" and "credit cards" also shows a slight imbalance, with more respondents having debit cards than credit cards.

In the context of model training and evaluation, it is imperative to handle these imbalances carefully to ensure that the model performs well across all categories.

2.5 Near-Zero-Variance

In our analysis, we employed the "nearZeroVar" function from the R caret package, which serves as a tool for identifying variables in a dataset that exhibit near-zero variance. The "nearZeroVar" function employs the following criteria to identify such variables: a low percentage of unique values relative to the total sample size, indicating minimal variation within the variable, and a significantly high ratio between the frequency of the most common value and the second most common value, signifying the dominance of one category within the variable. Based on these criteria, the function determines which variables qualify as having near-zero variance.

The results obtained from the "nearZeroVar" function indicate that none of the variables meet the criteria for near-zero variance. All variables within the dataset exhibit sufficient variability and may contain meaningful information that could contribute to predictive modelling or analysis. Data quality is paramount for the effectiveness of statistical models, and as per the findings of this analysis, the dataset meets these quality requirements.

Data pre-processing plays a pivotal role in guaranteeing the quality and reliability of the data. Through the steps outlined above, we can accurately understand the dataset's characteristics and provide a reliable data foundation for subsequent analysis and modelling.

3. Methodology

3.1 Data retrieval and cleaning

Data retrieval and cleaning were fundamental steps in our study. We sourced the dataset from the Global Findex Database on the World Bank's website, focusing on the COVID-19 era by selecting the 2021 edition, which encompasses data from 2021 and 2022. Our primary interest lies in individual-level data to delve into the nuances of financial inclusivity during this period.

Upon analyzing the dataset, we identified many variables with extensive missing data. Specifically, 50 out of 128 variables exhibited more than 70% missing values. A deeper investigation, guided by the dataset's codebook, revealed that this was primarily due to certain survey questions targeted only at respondents from specific countries, coupled with instances where respondents opted not to answer. This methodological choice in the survey design inherently led to the observed missingness.

Given the categorical nature of these variables, traditional imputation methods like mean or mode substitution were deemed inappropriate and potentially misleading. As a result, we chose to omit these variables from any further analysis. This decision was informed by the understanding that preserving the dataset's integrity was paramount, and removing these variables would eliminate the risk of introducing bias or inaccuracies into our research findings.

For the initial phase of our analysis, we focused on variables with fewer than 10,000 missing values. This criterion led us to select 42 variables, encompassing a mix of string, continuous, and categorical data types. Among these, only two variables, 'age' and 'emp_in', had missing values, which we addressed with distinct imputation strategies due to their differing data types.

The remaining dataset, now refined to 143,887 records and 42 variables, provided a robust foundation for our analysis. This curated dataset enabled us to explore demographic and financial indicators, such as age, income, account ownership, and saving habits, with greater accuracy and reliability.

3.2 Data Strategies

- Outlier handling

In the two columns containing outliers, we only addressed the column 'age' and did not process 'pop_adult,' which means the number of adult populations from which the individual is from. As mentioned before, the country-level data is not our focus. Therefore, there is no need to deal with the outliers in the adult population column.

Those outliers may carry information for the 'age' column and, therefore, could be kept. Depending on the specific situation, the change could happen during the data modeling phase.

- Missing value imputation

In our case, the specific methods to deal with missing values are two different imputation strategies. For 'age,' the numerical variable, we use the mean of the whole column to replace the 467 missing values. The employment status variable – 'emp_in' with 3502 missing values is replaced by the column mode, 1. Furthermore, despite no missing value in the 'regionwb' column, which represents World Bank regional code (all types in this column see in the Appendix), based on the missing count table, the nonvalue does exist in this column in the original dataset. After a thorough examination, we have found that only records with 'Taiwan, China' in the 'economy' column have no value in the 'regions' column. Therefore, we filled all these missing values in the column with 'Taiwan.'

- Imbalance data treatment

We dealt with the problem of imbalance during the data modelling. Before fitting the model, we used several resampling methods such as down sampling with high count class and up sampling with low count class. These strategies would make the training data more balanced and therefore get better model performance.

- Handling of high-cardinality columns

There are only two columns with high cardinality: ' economy' and 'economycode'. Both have 139 unique string values. Since our research focus is not on the country/economy level, we will not use these two columns in our advanced analysis and drop these two columns for the dataset used for the study.

3.3 Data Ethics

In our research process, informed consent is a cornerstone of research ethics. We ensure that participants fully understand the purpose of the research, the nature of their participation, and their rights regarding data use. To achieve this, we provide clear and comprehensive information to guarantee that participants can make voluntary participation decisions, with the right to withdraw at any time without any conditions.

Our meticulously crafted process for participation prioritizes clear communication and participant independence, guaranteeing that everyone provides informed consent after fully comprehending the study's aims and content. In this way, we respect individuals' choices and privacy while strengthening the ethical foundation of our research.

Furthermore, our team adheres to high standards of data handling and protection principles to secure the safety and privacy of participant information. Our practices reflect a commitment to respect, fairness, and beneficence towards participants, aiming to maintain our research's high integrity and ethical standards.

3.4 Data modelling

In our dataset for advanced analysis, there are 30 financial indicators, such as financial account ownership, usage of any digital payment, and saving status. We chose account status as our first predicted variable. Regarding the selection of predictors, we started with the target population of our research - poor adults and women. Therefore, we formed our first model:

$$account = a0 + a1 * gender + a2 * income$$

In this model, the response variable is a categorical variable with two values, which makes the model a binary classification problem.
We adopted logistical regression as the first training method. Before fitting the model, we tried different resampling techniques, which are not sampling, down sampling, up sampling, hybrid, and observation reweighting.

Following that, we employed random forest and Gradient-Boosted Trees (GBTs) to train the data sequentially. When training by the random forest, resampling beforehand is not necessary since bootstrap sampling is included in this algorithm. Before implementing GBTs, resampling data could be beneficial especially when dealing with imbalanced data, which is our case.

We used precision, recall, accuracy, and especially the area under the receiver operating characteristic (auroc) as our performance metrics. As can be seen from Table 1 below, among all the above algorithms, GBTs had the best performance.

Table 1: the performance results of different algorithms for the model 1

algorithm	resampling	precision	recall	accuracy	auroc
Logistic Regression	Down sampling	0.73	0.83	0.67	0.60
Random Forest	No sampling	0.72	0.92	0.70	0.59
GBTs	Down sampling	0.72	0.92	0.70	0.60

However, 0.6 is not a satisfactory auroc score for an appropriate model. Thus, we considered adding an interaction term into the model, and got our second model:

$$account = a0 + a1 * gender + a2 * income + a3 * gender * income$$

The result was not as good as we expected (it is shown in Table 2), which has led us to consider whether we should add more predictors rather than just gender and income. To avoid neglecting any possible factors of influence, we incorporated all the remaining relevant demographic features into the model, which is as below (model 3):

$$account = a0 + a1 * gender + a2 * income + a3 * age + a4 * emp + a5 * educ + a6 * region$$

This time we finally got the best model performance (The outcome is in Table 2), and since the auroc is more than 0.8, we could carry out hyperparameter tuning to improve the model performance further instead of changing the model structure.

Table 2: the performance results of different models

model	algorithm	precision	recall	accuracy	auroc
1	GBTs	0.72	0.92	0.70	0.60
2	GBTs	0.72	0.93	0.70	0.61
3	GBTs	0.87	0.79	0.76	0.84

4. Results

4.1 Basics

- Univariate Analysis

In our research, using univariate analysis, we aim to uncover the various aspects of individual access to financial services and technology—including ownership of bank accounts, financial institution accounts, mobile phone ownership, and digital payment methods. By adopting this strategy, we can gain an in-depth understanding of each variable's effects on financial inclusion and technology access. It enables us to detect patterns and imbalances in utilizing available resources.

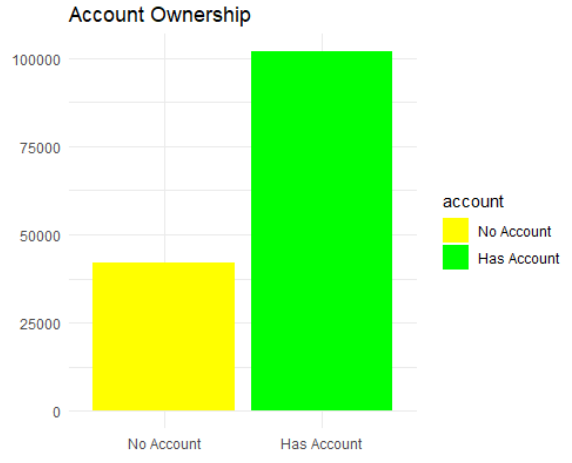

Figure 7: Account Ownership

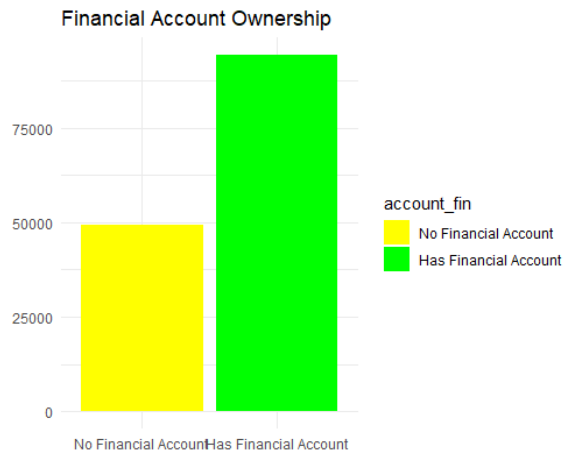

Figure 8: Financial Account Ownership

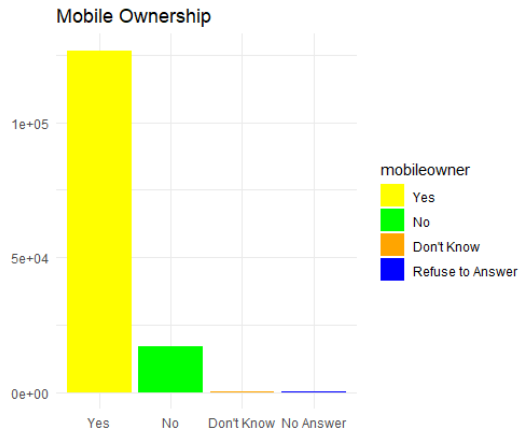

Figure 9: Mobile Ownership

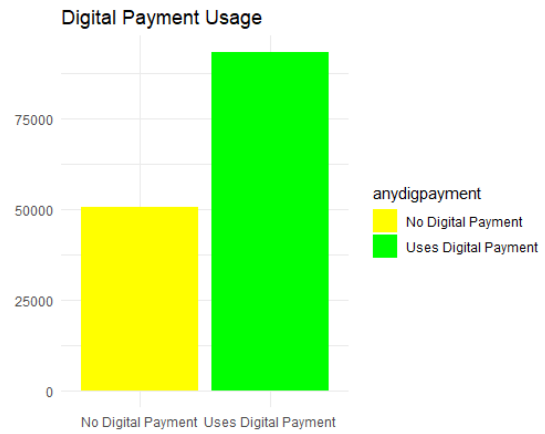

Figure 10: Digital Payment Usage Ownership

This set of four bar charts provides information on individual financial status and access to technology, specifically regarding the ownership of bank accounts, financial institution accounts, mobile phones, and digital payment methods. Overall, these charts are extracted from a survey on financial inclusion, aiming to showcase access to these essential financial and technological resources among people in one or more countries.

Account and Financial Account Ownership: Figure 7 and Figure 8 (Account Ownership and Financial Account Ownership) have a similar structure, each showing whether people have an account and a financial account (referring to specific financial services like savings accounts, investment accounts, etc.). In both charts, the green bars represent the number of people with accounts, while the yellow bars represent those without accounts. In both cases, the number of people with accounts significantly exceeds those without, indicating a high penetration of financial services among the surveyed population.

Mobile Phone Ownership: Figure 9 differs in structure from the first two, using a logarithmic scale, meaning the number of people owning mobile phones outnumbers the other categories. This chart shows the contrast between having and not having and includes those who are uncertain or refuse to answer whether they own a mobile phone. Using a logarithmic scale reveals a considerable imbalance in the proliferation of mobile phones.

Digital Payment Usage: Figure 10 focuses on whether people use digital payment methods, an essential aspect of financial technology. Like the first two charts, the green bar represents the number of people using digital payments, and the yellow bar represents non-users. The data indicates that the users of digital payments outnumber non-users, which may reflect the convenience and prevalence of digital payment tools.

In summary, this set of charts emphasizes the widespread use of financial services and mobile technology among the group surveyed. The prevalence of bank and financial account ownership might indicate financial stability and economic development levels. At the same time, the proliferation of mobile phones and digital payments reflects the level of technological access and digitalization. However, for mobile phone ownership data, the exceptionally high ownership rate underscores the importance of mobile phones as a point of access for communication and financial services, a key characteristic of modern society.

- Bivariate Analysis

For the two-variable analysis, we juxtaposed account ownership against various demographic factors such as income, region, gender, and education levels. This approach allowed us to discern how these demographic characteristics affect financial inclusion. By visually presenting this data in bar charts, we could effectively illustrate the significant impact that factors like income, geographical location, gender, and educational attainment have on an individual's likelihood of owning a financial account. In calculating the digital financial inclusion index, the "using an account" criterion—accounting for transactions made directly through financial institution accounts or via mobile phones—reflects variations in the primary banking methods across countries, distinguishing between those relying on institutional accounts, typical of markets with developed financial sectors like Europe, and those favoring mobile money, common in many African countries, to enable a nuanced cross-country comparative analysis (Dluhopolskyi et al, 2023).

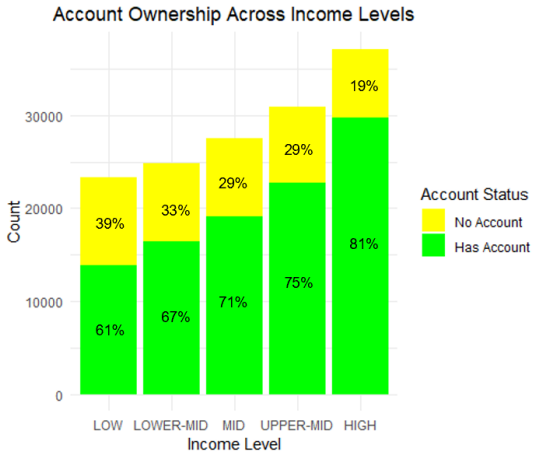

Figure 11: Account Ownership Across Income Levels

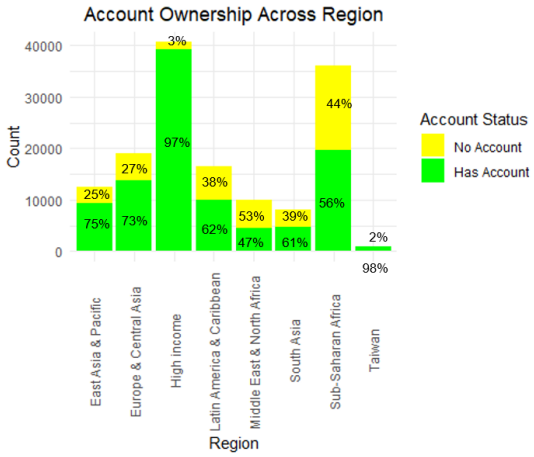

Figure 12: Account Ownership Across Region

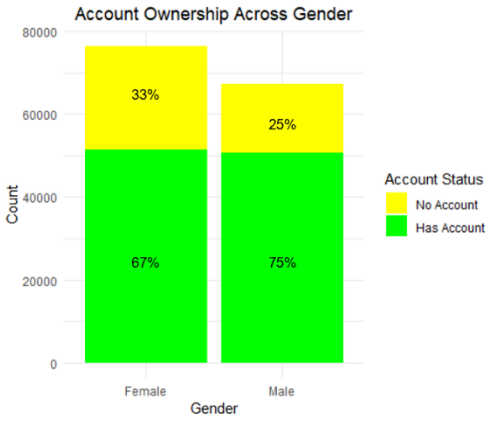

Figure 13: Account Ownership Across Gender

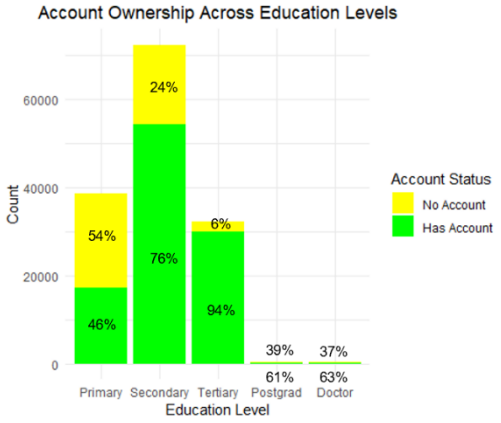

Figure 14: Account Ownership Across Education Levels

The bar charts depict the distribution of account ownership by various demographic factors.

Figure 11 illustrates the correlation between income levels and the likelihood of having an account. Five income categories represent: low, lower-middle, middle, upper-middle, and high. A clear positive trend is visible: higher income levels correspond to more individuals with an account. For instance, in the low-income group, 61% have accounts and 39% do not, while in

the high-income group, 81% have accounts and only 19% do not. It suggests that people are more likely to have an account as they move up the income ladder.

The second graph (Figure 12) provides a regional breakdown of account ownership, comparing various parts of the world. This stark contrast between different regions is evident. In high-income areas, 97% have accounts, while only 3% do not, indicating near-universal financial inclusion. Conversely, in the Middle East and North Africa, only 47% have accounts, with 53% having no, pointing towards significant financial exclusion. Other regions, like South Asia and Sub-Saharan Africa, also have percentages of the population without bank accounts, at 61% and 56%, respectively.

Gender disparity in account ownership is depicted in this graph (Figure 13), comparing males and females. While 67% of females reported having accounts, this figure rises to 75% for males. Conversely, 33% of females do not have accounts, compared to 25% of males. This data suggests a gender gap in financial inclusion, where men are more likely to have bank accounts than women.

The last graph (Figure 14) demonstrates the relationship between education levels and account ownership. A pronounced positive correlation is seen here; individuals with higher education levels are likelier to have an account. However, at the other end of the spectrum, a mere 6% still need an account among those with a doctorate, and a vast majority (94%) possess one. This trend is consistent across education levels, highlighting the role of education in financial inclusion.

Each graph not only charts the respective counts but also color-codes the bars to differentiate between those with and without accounts (green for account holders, yellow for non-account holders), making it visually clear that income, region, gender, and education are all significant factors in determining whether a person is likely to have an account.

- Trivariate Analysis

Our three-variable analysis combined demographic factors (income and gender) with financial inclusion (account ownership) in a mosaic plot, using statistical tools like Pearson residuals and significance tests. This approach highlighted the interplay between gender, income, and financial access, revealing gender disparities and the impact of income on account ownership. The clear, color-coded visualization and the statistically significant results provided a solid understanding of the complex relationships at play.

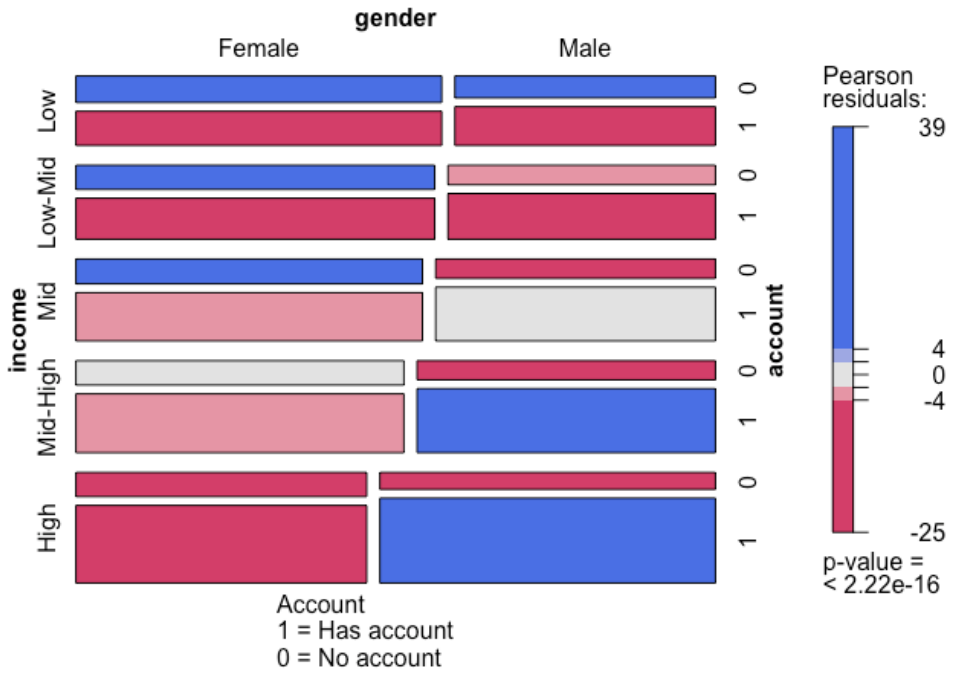

Figure 15: The mosaic plot associated with Having Account

The mosaic plot (figure 15) is associated with Pearson residuals and a significant test, often used in statistics to visualize the strength of the relationship between categorical variables and to identify patterns or associations within the data. The plot shows the distribution of account ownership within each gender and income level. The positive residuals (represented by the blue bar) suggest more observations in that cell than expected under independence, whereas harmful residuals (represented by the pink bar) suggest fewer than expected. The p-value indicated is less than 2.22e-16, a standard notation for a value of 0. This suggests that the association between the variables (gender, income, and account ownership) is statistically highly significant. By comparing the color saturation between genders within the same income level, we can assess whether there is a gender-based disparity in financial inclusion; if a particular income level shows a markedly distinct color saturation between the male and female segments, it suggests a gender disparity in account ownership within that income bracket. For example, if the tiles for females at a certain income level are consistently less saturated than those for males, it could indicate that females are less likely to own an account within that income bracket compared to males. An increasing trend of account ownership at higher income levels implies that a higher income correlates with greater financial inclusion, showing a positive link between the level of income and access to financial services.

4.2 Model

In Table 2 above, the highlighting data is the outcome of our final and best model. This demonstrated that for the prediction results of the test data, this model has the highest precision, accuracy, and auric, which means among all the predictive positive values, it has the highest prediction accuracy; in all predictive outcomes, it has the highest rate of correctness; for the majority of positive class samples, it gives the positive predictive value, and regarding most negative class instances, it gives the negative predictive value, therefore indicating that it has a better ability to distinguish between positive and negative classes. However, it has the lowest recall in all three models, which means it is more conservative than the other two models and would classify a record as positive only if it is highly confident.

When examining the k-fold cross-validation of this model, the model performance could still stabilize at the value in the table, which sufficiently illustrated the stability of this model's performance.

We also used grid search to implement hyperparameter tuning, but considering the cost of training time, we set two distinct values for each of the three parameters, which are the max depth of trees, max bins, and max number of iterations, and got the best values within them, which are 5, 32, and 30 respectively. The final metrics remain the same as in the table above.

4.3 Discussion

We examined why this method is appropriate from four key perspectives.

- Model Building

GBTs add trees sequentially, each trying to correct the residuals (errors) of the previous tree. This gradual optimization approach makes GBTs especially effective at reducing model error. It focuses on those samples that were wrongly predicted by the previous tree, making the model progressively focus on hard-to-predict cases.

- Overfitting Control

GBTs can finely control the model's complexity by adjusting parameters such as the number of trees, learning rate, and tree depth, effectively reducing the risk of overfitting.

- Model Complexity and Non-linear Relationships

As a tree-based model, GBTs can naturally capture non-linear relationships and complex interactions between variables without explicitly specifying these relationships in the model.

- Feature Handling and Selection

GBTs can automatically handle different types of features (numerical and categorical) and select important features during training, thereby enhancing the model's predictive power and interpretability.

In summary, GBTs, through gradual optimization and focusing on hard-to-predict samples, have the capability to capture complex data patterns and non-linear relationships, enabling them to outperform Random Forest and Logistic Regression in our case.

5. Conclusion

5.1 Fit-for-Purpose

The statistical models deployed in this study yielded results that are both statistically significant and fit-for-purpose. The models' precision, recall, and accuracy metrics indicate robust predictive capabilities. The application of Gradient-Boosted Trees (GBTs) has further enhanced the reliability of our findings. Relating these outcomes to our original objectives, it is evident that the models have successfully identified key socio-demographic characteristics that affect financial inclusion during the COVID-19 pandemic. Region, education, and income level have emerged as the most influential factors, aligning with our goal to illuminate the socio-economic dynamics affecting financial resilience.

5.2 Problems Throughout the project

We encountered challenges related to data availability and the intricacies of capturing the multi-dimensional aspects of financial inclusion. To overcome these, we implemented rigorous data cleaning, validation, and variable selection strategies that would maintain the integrity of our models. Moreover, we adapted our methodology to accommodate the limitations of our dataset, ensuring our results remained reliable and valid.

5.3 Bias

While we have made considerable efforts to ensure our models are unbiased, the potential for residual bias remains, especially in terms of data representativeness. There is a risk that certain populations may be underrepresented, which could lead to overgeneralized conclusions. This possibility underlines the importance of interpreting our findings within the context of known data limitations and applying them with caution. It is crucial for future research to continue to identify and minimize such biases, ensuring the equitable application of findings across all population segments.

Based on the findings from the report, a policy on financial literacy training should focus on developing and implementing comprehensive financial education programs aimed at enhancing financial inclusion. This policy would prioritize targeting key socio-demographic groups identified by your statistical models - such as regions with lower financial inclusion, individuals with varying education levels, and those across different income levels - to tailor financial literacy programs effectively. It would involve leveraging digital platforms to extend financial education's reach, ensuring that these initiatives are inclusive, accessible, and designed to equip individuals with the necessary skills to navigate financial systems, particularly in the context of recovering from crises like the COVID-19 pandemic. The policy should also encourage collaboration among governmental bodies, financial institutions, educational sectors, and community organizations to foster a supportive ecosystem for financial literacy and inclusion.

6. Future Work

For those interested in building upon this research, it would be advantageous to consider additional variables that may affect financial inclusion, such as access to digital platforms, financial literacy, and the impact of COVID-19-specific economic policies. Longitudinal data could also reveal how financial inclusion changes over time, helping to distinguish between short-term pandemic effects and longer-term trends. In the forthcoming evolution of the financial sector, it is anticipated that there will be a pronounced trend towards the digitization of services, alongside a push for tailoring these services more closely to the individual preferences and requirements of consumers, with an emphasis on delivering an expansive array of these services through a singular, streamlined digital platform; concurrently, there will be a strategic shift in identifying financial accounts by mobile phone numbers rather than traditional account numbers to broaden the scope of financial inclusion, paired with a concerted effort to elevate the financial status of women to foster independence and empowerment, while governments are expected to step up their direct involvement in ensuring the provision of essential financial services to the economically marginalized sectors of society, all amid the rise of pioneering financial technologies designed to systematically lower the costs associated with transactions. (Ozili, P.K. (2023).

To enhance our study, we plan to refine our model's performance by incorporating advanced machine learning techniques tailored to our data's unique traits, like ensemble learning, and by adding variables such as personal spending habits. We'll also explore the combined effect of gender and income on financial account ownership and extend our investigation into how

age and education level interact with financial inclusion using statistical methods. Additionally, we aim to conduct a detailed regional and demographic analysis using GIS technology to uncover variations in financial access across different locations. Expanding our research to include other financial services like digital payments and microloans will also be a priority, utilizing surveys and public data to gather information on these services.

7. Acknowledgements

We are deeply thankful to our supervisors, Professor Arindam Basu, Phil Davies, and Nicky Cartlidge, for their invaluable advice and constant encouragement throughout this project. Their insight and guidance were valuable in shaping both the direction and execution of our work.

We are also immensely grateful to our team members: Cathy Peng, Chathurangi Godahewa Gamage, and Lizhu Dong. Our dedication, expertise, and collaborative spirit were fundamental to the success of this project. Each member brought unique strengths and perspectives, enriching the quality of our work.

Our sincere appreciation extends to the Faculty of Health at the University of Canterbury for their association and support. This collaboration enriched our understanding and added depth to our research.

We also acknowledge the Global Financial Index for providing crucial data in our analysis and findings. Access to these data sources was instrumental in achieving the objectives of our project.

Although this project did not require financial support, the resources and environment provided by the University of Canterbury significantly facilitated our project. We are thankful that these provisions enabled us to conduct our study effectively.

Special thanks are due to the advanced AI assistance of ChatGPT, which provided valuable insights and suggestions, and to Grammarly for its aid in ensuring the clarity and correctness of our written materials. Both tools were indispensable in enhancing the quality and coherence of our research communication.

Finally, we extend our heartfelt thanks to all those who offered encouragement and moral support throughout this endeavor. Your belief in our work has been a source of motivation and strength.

References

1. Demirgüç-Kunt, A. et al. The Global FINDEX Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19. *World Bank Group*.
2. Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). *The Global Findex Database 2021*. https://doi.org/10.1596/978-1-4648-1897-4

3. The World Bank. (2022). *The Little Data Book on Financial Inclusion 2022*. http://hdl.handle.net/10986/38148 License: CC BY 3.0 IGO

4. Welcome to Python.org. (2024, February 8). Python.org. https://python.org/

5. RStudio Desktop. (2024, January 11). Posit. https://posit.co/download/rstudio-desktop/

6. Hess, J., Klapper, L., & Beegle, K. (2021). Financial inclusion, women, and Building Back Better. *World Bank Publications - Reports*. https://ideas.repec.org/p/wbk/wboper/35870.html

7. Kasradze, T. (2020). Challenges facing financial inclusion due to the COVID-19 pandemic. *European Journal of Marketing and Economics*, *3*(2), 50. https://doi.org/10.26417/523jma34n

8. Demir, A., Pesqué-Cela, V., Altunbaş, Y., & Murinde, V. (2020b). Fintech, financial inclusion and income inequality: a quantile regression approach. *The European Journal of Finance*, *28*(1), 86–107. https://doi.org/10.1080/1351847x.2020.1772335

9. Gutiérrez-Romero, R., & Ahamed, M. M. (2021). COVID-19 response needs to broaden financial inclusion to curb the rise in poverty. *World Development*, *138*, 105229. https://doi.org/10.1016/j.worlddev.2020.105229

10. Ozili, P.K. (2023). The Future of Financial Inclusion. In: Mhlanga, D., Ndhlovu, E. (eds) *Economic Inclusion in Post-Independence Africa. Advances in African Economic, Social and Political Development.* Springer, Cham.

11. Dluhopolskyi, O., Пахненко, О., Lyeonov, S., Semenog, A., Artyukhova, N., Cholewa-Wiktor, M., & Jastrzębski, W. (2023). Digital Financial Inclusion: COVID-19 Impacts and opportunities. *Sustainability*, *15*(3), 2383. https://doi.org/10.3390/su15032383

Appendices

Appendix A: Dataset Documentation

Dataset Name: **micro_world.csv**
Source: Global Financial Database
Description: This dataset is a comprehensive collection of financial inclusivity metrics among women and impoverished adults during the COVID-19 pandemic, spanning 123 economies with approximately 128,000 adult respondents. The dataset comprises 128 variables over 143,887 records, encompassing a broad spectrum of quantitative and qualitative data points, including age, income, educational levels, gender, and geographic classification.
Data Processing: Initial data cleaning and preprocessing involved handling missing data, identifying and treating outliers, and addressing data imbalances. Detailed steps include applying imputation techniques of disappeared values, boxplot methods for outlier identification, and strategies to mitigate data imbalance.

Appendix B: Variable Descriptions

Source: **GlobalFindex2021-MicrodataCodebook.pdf**
Content: Detailed explanations of each variable included in the dataset, specifying the type of data (e.g., floating-point, integer, string) and the relevance of each variable to the study's objectives. This section would also include any specific categorizations or coding schemes used to classify responses.

Appendix C: Notebooks of Code

R Notebook for Methodology & Analysis:
Filename: Methodology&Analysis.R
Description: The code for data retrieval, cleaning, and preliminary analysis. This notebook includes scripts for handling missing data, outlier identification, and initial exploratory data analysis to understand the dataset's structure and main characteristics.
Python Notebook for Modeling:
Filename: 601project.html.ipynb
Description: Presents the development and validation of predictive models using Python. It uses machine learning algorithms, such as Gradient Boosted Trees, to estimate an individual's gender and poverty status based on financial indicators. The notebook includes code annotations explaining the choice of models, feature selection, model tuning, and evaluation metrics.

Appendix D: Report Charts and Visualizations

Filename: Supplementary material
Content: Includes all charts and visualizations referenced in the report, offering a visual representation of key findings. This section would encompass:

- Boxplots illustrate the distribution of variables such as age and income before and after data cleaning.
- Bar charts show the proportion of missing values per variable, highlighting data completeness across dimensions.
- Mosaic plots and other visual analyses depict the relationship between financial inclusion indicators and demographic characteristics like gender, income level, and region.