Maximizing Fairness with Synthetic data: Access to Emergency Fund in Sub-Saharan Africa

Charavee Basnet Chettri, Vivian Wei, Betty Pu, Ziyue Yang [*]

Department of Statistical and Data Sciences, Smith College, Northampton MA

March 6, 2024

Abstract

Financial inclusion is paramount for economic stability and resilience, particularly in diverse regions like Sub-Saharan Africa, spanning low to high-income countries and encompassing both resource-intensive and non-resource-intensive economies. This study focuses on a crucial aspect of financial resilience: the accessibility of emergency funds, defined as having access to 1/20 of Gross National Income (GNI) per capita in local currency within 30 days. Leveraging previous colleagues' exploratory work on the Global Financial Inclusion Database 2021, our objective is to mitigate inherent gender biases in the dataset by rebalancing it with synthetic data, thereby enhancing fairness in predicting emergency fund accessibility. Through predictive machine learning modeling, we aim to contribute to the economic empowerment of individuals in Sub-Saharan Africa, ultimately fostering resilience and reducing disparities in access to essential financial resources.

Keywords: Fairness metrics; emergency funds; financial inclusion; gender-based bias & debiasing; Sub-Saharan Africa; predictive machine learning; synthetic data

1 Introduction

In the realm of financial inclusion, the accessibility of emergency funds plays a pivotal role in determining an individual's financial stability and resilience, especially in developing countries [1]. In this project, our goal is to predict the possibility for people in Sub-Saharan African countries to come up with emergency funds, defined as 1/20 of GNI per capita in local currency, within a 30-day period[2]. This prediction serves as a crucial factor for establishing future public financial policies, such as determining the eligibility of individuals for loans and financial assistance. The significance of this problem lies within its direct impact on the economic well-being and empowerment of individuals in developing regions. According to the Global Financial Inclusion (Global Findex) Database 2021 published by the World Bank, only a little over half of people over 15 years of age in developing economies could access extra funds within 30 days if faced with unexpected expenses [2]. Therefore, there is a pressing need to understand the factors influencing this accessibility and eliminate inherent bias in the dataset. By delving into this issue, we not only contribute to enhancing financial inclusion but also aid in mitigating the different effects of financial shocks on vulnerable populations.

Defining fairness is essential since the concept itself is relative among different people. In the data science discourse, fairness encompasses three key aspects: individual fairness, group fairness, and causal fairness[4]. Individual fairness focuses on preventing discrimination against individuals with similar relevant characteristics. This means ensuring that individuals in similar situations receive similar outcomes from the model, regardless of irrelevant factors[3]. Group fairness aims to prevent disparities in outcomes for different groups. This ensures equal opportunities for all groups, regardless of their membership[8]. Causal fairness goes beyond simply observing disparities and delves into understanding

their underlying causes. It seeks to mitigate these root causes to achieve fair outcomes within and across groups[6].

Our previous colleagues conducted analysis using the <AI & Equality> Human Rights Toolbox and used a Decision Tree Classifier machine learning model implemented via Python to predict access to emergency funds with 68% accuracy. Their work laid a solid foundation by exploring demographic and financial variables within the dataset, and assessed the fairness of the decision tree classifier, particularly concerning gender bias, and then applied various processing techniques to enhance the fairness of the model[7].

Based on their work, our group aims to incorporate synthetic data to rebalance the dataset, ensuring equitable representations amongst both genders in the dataset. Our approach aims to consider a broader selection of machine learning algorithms and mitigate the disparities the previous colleagues found in the decision tree's classifier's predictions and enhance fairness with synthetic data, ultimately better predict access to emergency funds in south-Saharan Africa countries[5].

2   Background on Sub-Saharan Region

Sub-Saharan Africa is a region characterized by a diverse economic landscape, encompassing low, lower-middle, and upper-middle-income countries. Demographically, Sub-Saharan Africa is marked by a rich tapestry of cultures and a population exceeding 1.2 billion people. This expansive and diverse demographic landscape includes 22 countries grappling with fragility or conflict, posing unique challenges to development efforts. Additionally, 13 small states within the region are characterized by limited human capital, modest populations, and constrained land areas [ADD CITATION].
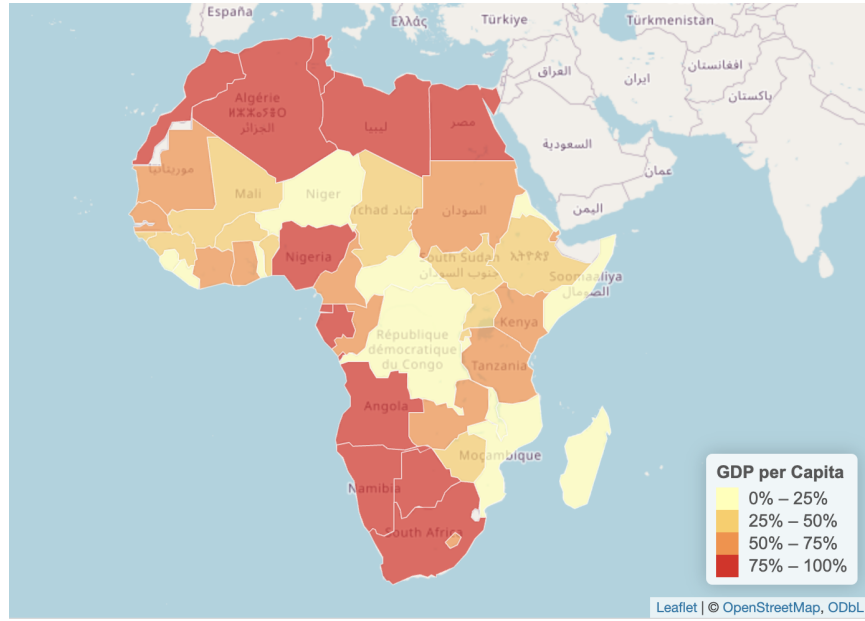
According to the World Bank's definition, middle-income countries had a per capita gross national income of more than US$995.00 in the years 2015–17 [ADD CITATION]. Among the 35 countries included in our dataset, 20 are classified as low-income countries, and 15 are classified as middle-income countries. Additionally, 11 are classified as countries in Fragile and Conflict-Affected Situations, which, by definition, have experienced a peacekeeping or peace-building mission within the last three years.

Sub-Saharan African countries not only differ in terms of economic prosperity but also in economic structure and resource intensity. Resource-intensive countries include both oil-exporting nations, where net oil exports make up 30 percent or more of total exports, and commodity exporters, where nonrenewable natural resources represent 25 percent or more of total exports. The divergence between resource-intensive and non-resource-intensive countries became more entrenched following the commodity price shock of 2015 [ADD CITATION]. Non-resource-intensive countries have proven more resilient, supported by their more diversified economies. On the other hand, resource-intensive economies generally have a less diversified structure, making them more susceptible to external shocks. In our

dataset, 6 of the countries are oil exporters, and 11 export other commodities such as iron ore, copper, cotton, coffee, and sugar. The remaining 18 countries are non-resource-intensive and their economies are not reliant on exports.

In recent years, the Sub-Saharan Africa region has grappled with significant economic challenges, including soaring inflation, pronounced exchange rate pressures, debt vulnerabilities, and widening economic disparities within the region. Therefore, addressing these structural and economic disparities is imperative for tackling developmental issues in the region.

3 Data

In order to continue assessing and enhancing the fairness of the machine learning models, we use the same Global Findex Database as our colleagues before us. The Global Findex database was first launched in 2011 by the World Bank—with funding from the Bill & Melinda Gates Foundation. It is the world's most comprehensive data set on how adults save, borrow, make payments, and manage risk. The dataset contains over 200 indicators including account ownership, payments, savings, credit, and financial resilience and has coverage over 140 nations, representing 97% of the world's population for year 2017, 2014, and 2011.

The data was constructed through a series of surveys carried out by Gallup, Inc in association with the annual Gallup World Poll. They randomly sampled 1000 individuals from each country and asked them to respond to a survey either over the phone or in-person. The target population is the civilian, non-institutionalized population 15 years and above. In consistency with the former colleagues, our sample covers only the Sub-Saharan region (35 countries total), thus there are 35,000 total observations and 105 total variables. Since

the sampling was random and the sample size is large, we can assume that our sample is representative of the total population of people living in the 35 Sub-Saharan countries in the data. The data was collected directly from individuals over the 2017 calendar year and is self-contained, meaning it does not rely on any external resources.

3.1 Variable Definition and Descriptions

After basic data cleaning, we have created various visualizations to help understand who is in the data and acknowledge potential sources of bias within the data.

	All	Female	Male
gender	0.50936	1.00000	0.00000
emp_in	0.70035	0.63793	0.76515
fin2	0.19301	0.16791	0.21907
fin7	0.04627	0.03890	0.05393
fin15	0.21585	0.18921	0.24351
fin16	0.11821	0.10018	0.13693
fin19	0.06083	0.04980	0.07228
fin20	0.17252	0.16766	0.17757
fin21	0.11813	0.10530	0.13144
has_access	0.43911	0.37033	0.51050
fin26	0.33015	0.28279	0.37932
fin28	0.37033	0.36059	0.38043
fin30	0.30133	0.28956	0.31356
fin32	0.25506	0.20036	0.31184
fin37	0.08334	0.08606	0.08051
fin38	0.04564	0.04476	0.04656
fin42	0.25670	0.22687	0.28766
fin48	0.71945	0.68558	0.75461
mobileowner	0.70321	0.65395	0.75435
educ_2	0.43490	0.39635	0.47492
educ_3	0.03828	0.02800	0.04896
inc_q_2	0.17214	0.19062	0.15296
inc_q_3	0.18640	0.19541	0.17706
inc_q_4	0.21097	0.20614	0.21598
inc_q_5	0.27100	0.22489	0.31887
economy_Botswana	0.03769	0.05129	0.02358
economy_Burkina Faso	0.03546	0.02230	0.04913
economy_Cameroon	0.03849	0.03989	0.03704
economy_Chad	0.03601	0.02412	0.04836
economy_Congo, Rep.	0.03420	0.03229	0.03618
economy_Cote d'Ivoire	0.03816	0.02313	0.05376
economy_Ethiopia	0.03639	0.04179	0.03078
economy_Gabon	0.03584	0.03271	0.03910
economy_Guinea	0.03206	0.02957	0.03464
economy_Kenya	0.04114	0.04229	0.03996
economy_Lesotho	0.03710	0.04220	0.03181
economy_Madagascar	0.04114	0.04889	0.03310
economy_Malawi	0.03896	0.05104	0.02641
economy_Mali	0.03702	0.03370	0.04047
economy_Mauritania	0.02999	0.02279	0.03747
economy_Mozambique	0.03698	0.03370	0.04038
economy_Namibia	0.04039	0.05319	0.02709
economy_Niger	0.03622	0.03064	0.04201
economy_Rwanda	0.03967	0.04642	0.03267
economy_Senegal	0.03555	0.03295	0.03824
economy_South Africa	0.03849	0.03642	0.04064
economy_South Sudan	0.03218	0.03254	0.03181
economy_Togo	0.03559	0.02750	0.04399
economy_Uganda	0.03828	0.04245	0.03395
economy_Zambia	0.03917	0.04493	0.03318
economy_Zimbabwe	0.04005	0.04815	0.03164

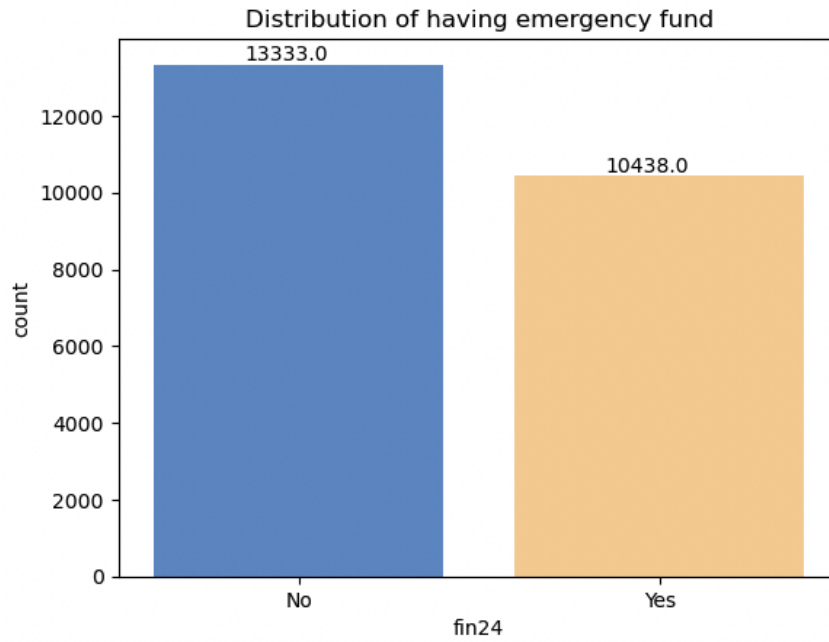*Outcome Variable of Interest: Access to Emergency Fund (Fin24)*

Figure 1: Distribution of having emergency fund

The variable "Fin24" in the dataset asked participants the question: Now, imagine that you have an emergency and you need to pay [1/20 of GNI per capita in local currency]. Is it possible or not possible that you could come up with [1/20 of GNI per capita in local currency] within the NEXT MONTH? In order to make it more straightforward for interpretations, we restrict the response into a binary variable for Yes and No. The overall distribution of access to emergency funds showed 17, 599 individuals had access to emergency funds while 14342 did not. This indicates that over half of individuals represented in the data do not have access to emergency funds.

To help conceptually understand how the outcome variables might vary among different levels of the variables, we chose 50% as the benchmark proportion for checking if the possibility of coming up with emergency funds for each variable group of interest is different from a random 50/50 chance.

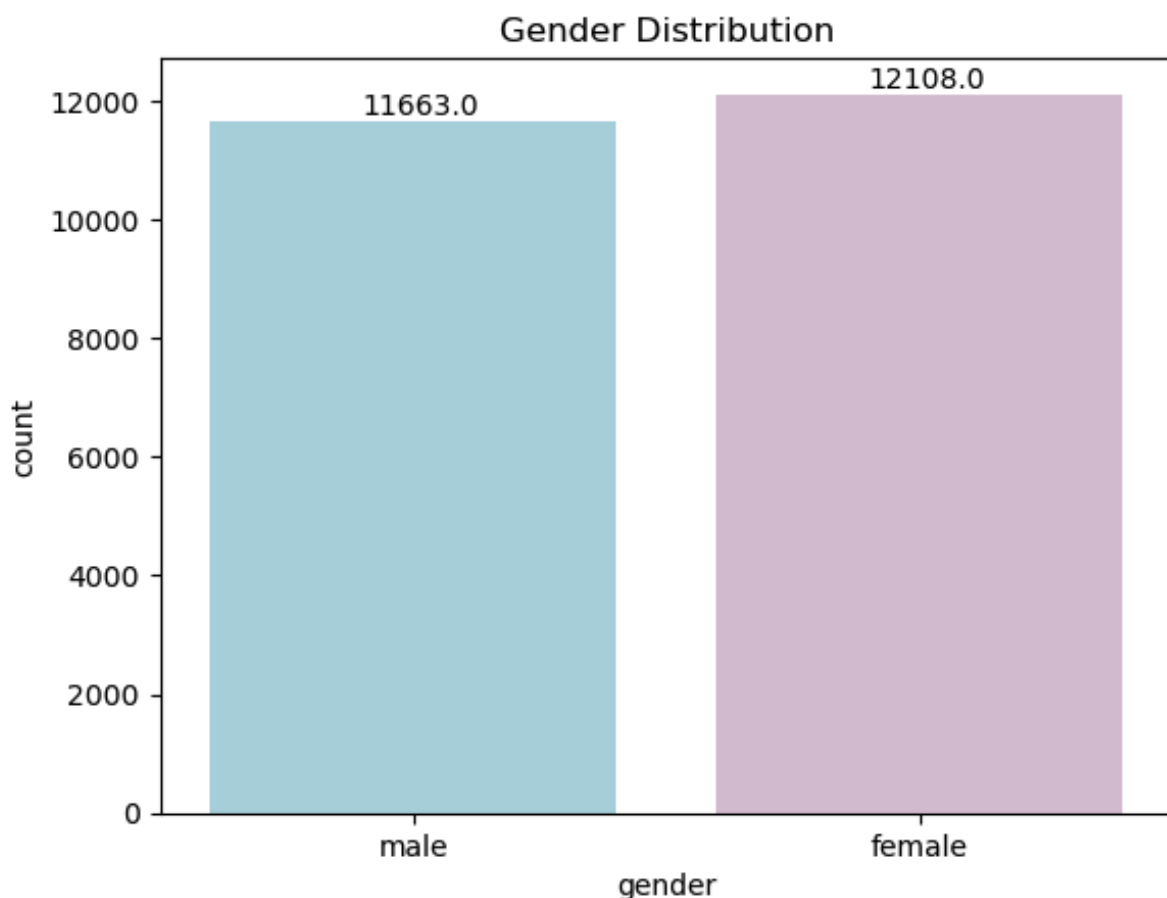*Protected Attribute of Focus: Gender (female)*

Figure 2: Distribution on gender

The variable "female" distinguishes gender. There are 12108 females and 11663 males in the dataset. It appears to have a fair distribution between different genders for the dataset as a whole.

While it looks quite balanced in the previous figure, we observe a discrepancy between the two genders when compared against our outcome variable of interest. Females have a 15% lower chance of having access to emergency funds.

In order to investigate further into how discrepancy of gender plays into other predictor variables, we created visuals for contrasting the two genders.

Education (educ)

The variable "educ" distinguishes education level among participants: 1 being the com-
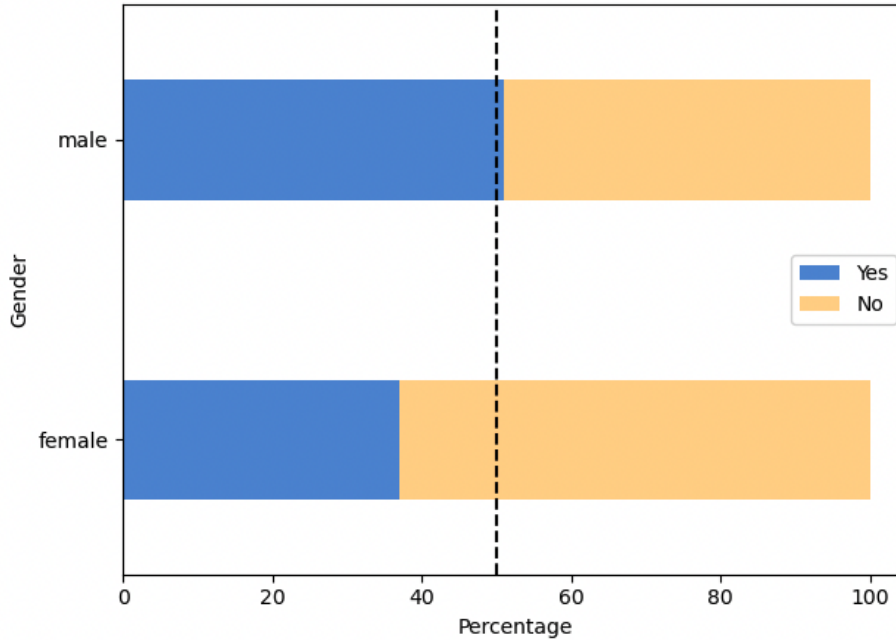
Figure 3: Distribution of having emergency fund by gender

pleted primary education or less, 2 being completed secondary education, and 3 being completed tertiary education. The majority in the dataset is participants with primary or less education (12523 participants). On the other hand, there are only 910 participants who have tertiary education. Overall, the discrepancy between education levels appears quite drastic, and might result in a varying prediction for access to emergency funds. If we break down education by gender, we see that more females than males have primary or less education while the opposite for secondary or tertiary education.

When we consider gender and education with respect to the outcome variable of interest, we observe that tertiary education levels have the highest likelihood of access to emergency funds while primary education is the least likely.

This discrepancy between education levels is reasonable since participants with tertiary education might have been more financially independent and have higher financial resilience, thus be more likely to have access to emergency funds. Nonetheless, we continue
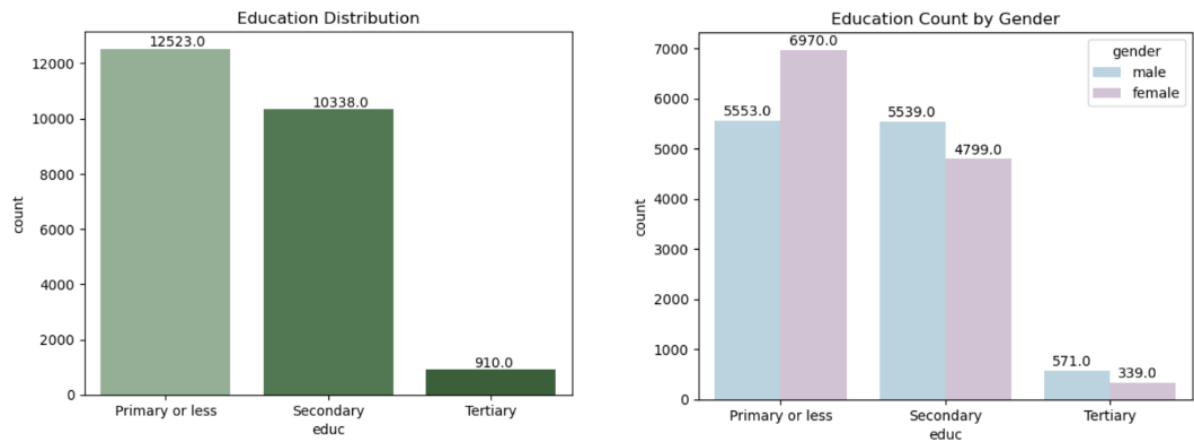
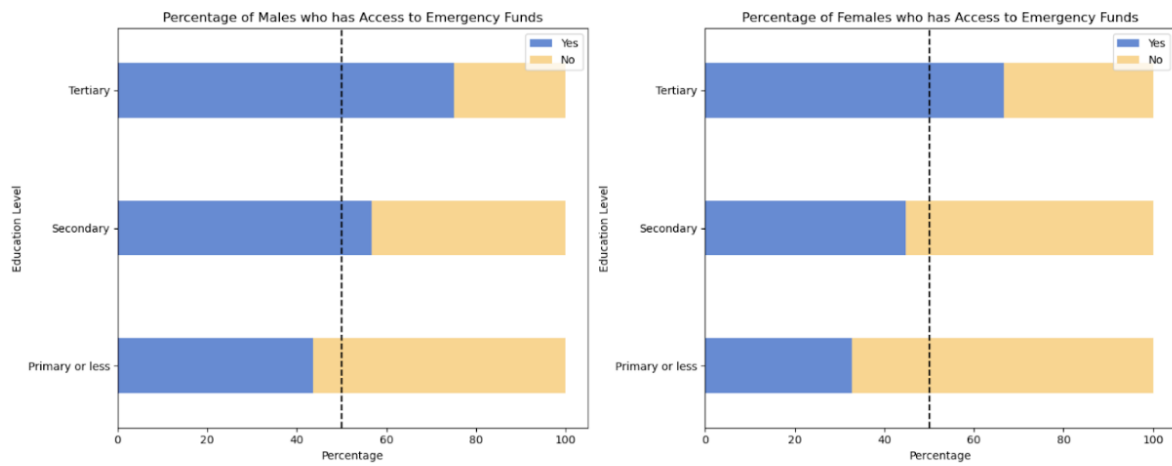Figure 4: Distribution by education level

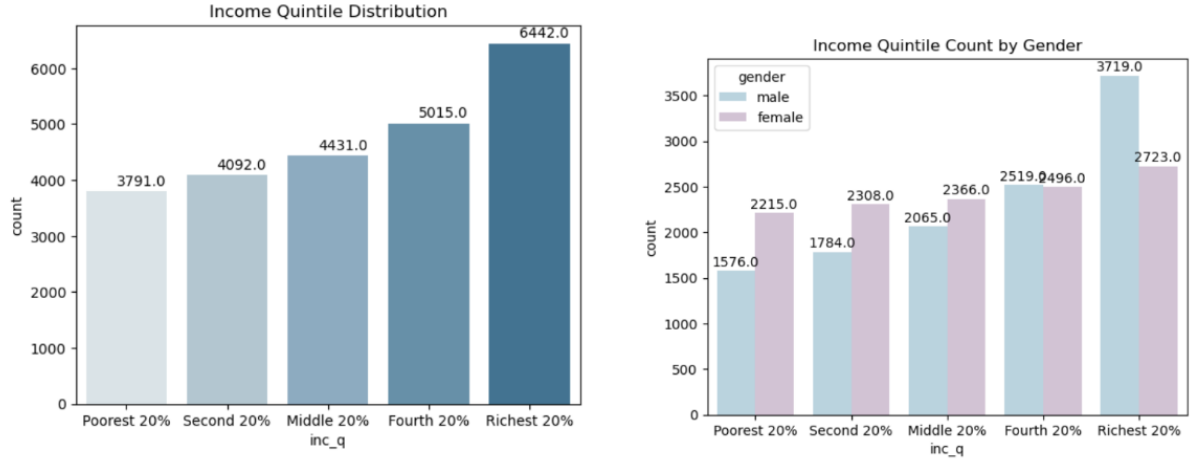Figure 5: Distribution of access to emergency funds by education level and gender

Figure 6: Income Quantile Distribution

to observe that females have a lower access to emergency funds, among the same education level. It shows that gender imbalance is still evident when we position it in relationship to other predictors.

Income Quantile (Inc_q)

The variable "inc_q" distinguishes the income quintile within an economy. It is separated into 5 quantiles with 1 being the poorest and 5 being the richest. Income quintile 5, or the richest 20%, is the largest quantile. There are more male than females in the richest 20% quintile, while females compose the majority of the poorest 20%, second poorest 20%, and middle 20%, which are the poorest quantiles.

We observe a similar gradient in access to emergency funds among the different income quintiles. This discrepancy is expected since participants with the level of income are directly related to financial stability and resilience. However, the gap between females and male is still concerning: male that are in the top two income quintile will have at least 50% chance of access to emergency funds, whereas only females that are the top 1 income quintile will have a comparable likelihood.
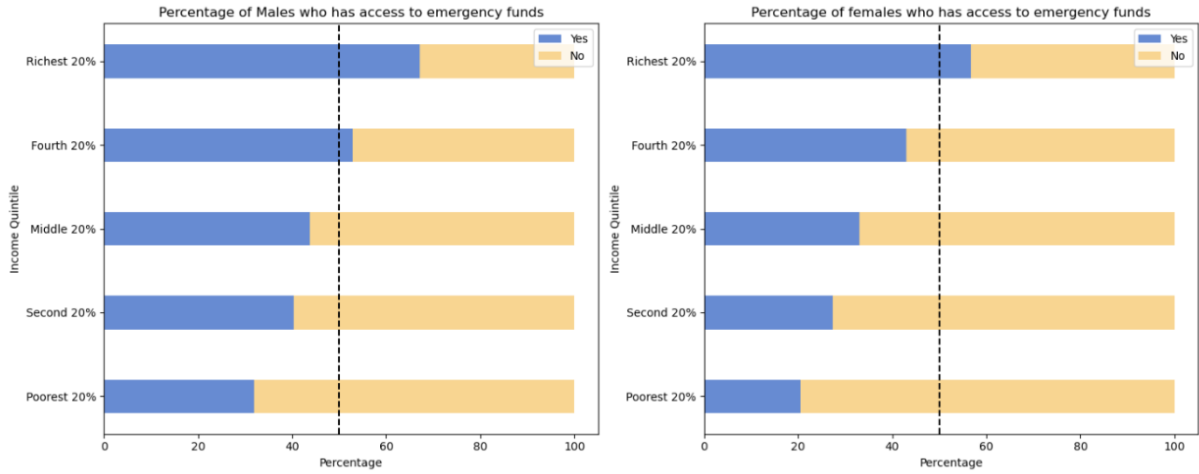
Employment Status (emp_in)

11

Figure 7: Distribution of access to emergency funds by income quantiile and gender

Employment Status asks whether or not the participant is in the workforce. There are 16648 individuals who were in the workforce while 7123 were not, and more females are out of the workforce than males, thus suggesting that females are less likely to achieve economic independence than males.

The disparity between females and male is quite alarming when we break it down by employment status. Regardless of the employment status, females do not even have a 50% chance of having access to emergency funds.

Country (economy)

The "economy" variable represents the country that the participants live in. There are 27 different countries from Sub-Saharan Africa with around 800-900 respondents each. The countries included are Benin, Botswana, Burkina Faso, Cameroon, Central African Republic, Chad, Congo Dem. Rep. Congo Rep., Côte d'Ivoire, Ethiopia, Gabon, Ghana, Guinea, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, South Africa, South Sudan, Tanzania, Togo, Uganda, Zambia, and Zimbabwe.

While looking at the outcome variable of interest broken down by country and gender,
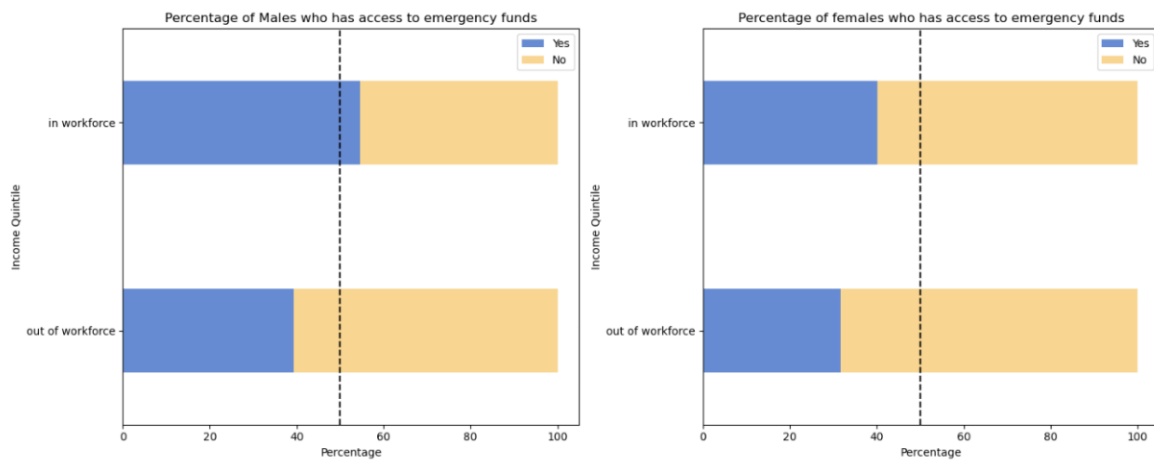
12

Figure 8: Distribution of employment status

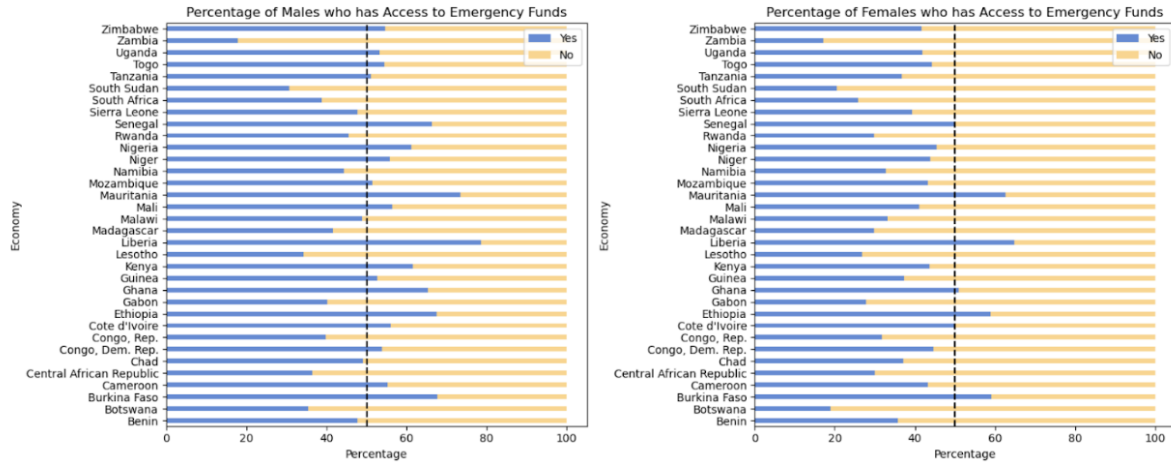Figure 9: Distribution of access to emergency funds by employment status and gender

Figure 10: Distributions of Access to Emergency Funds by country and gender

we see that Liberia has the highest proportion of both females and male having access to emergency funds, while Zambia has the lowest proportion of accessing emergency funds in both females and males. It's also important to note that the difference between males and females is largest in countries like Botswana and Kenya.

We iterate the visualizations for each main financial variable of interest in the dataset by gender, by gender and economy, and by gender and education level, to check if there's recurring discrepancy among genders. Due to the page constraints, we decide to not include all visualizations and attach our Jupyter Notebook for reference.

References

[1] Rashmi Umesh Arora. 222Links between financial inclusion and financial stability: A study of BRICS. In *Handbook of BRICS and Emerging Economies*. Oxford University Press, 12 2020.

[2] Asli Demirgüç-Kunt, Leora Klapper, Dorothe Singer, and Saniya Ansar. *The Global*

Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19. World Bank, Washington, DC, 2022.

[3] Will Fleisher. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 480–490, New York, NY, USA, 2021. Association for Computing Machinery.

[4] Sofia Kypraiou. What is Fairness? *Feminist AI*, sep 13 2021. https://feministai.pubpub.org/pub/what-is-fairness-.

[5] Kypraiou, Sofia. Synthetic data for AI equality. `https://aiequalitytoolbox.com/wp-content/uploads/2023/10/Copy-of-Synthetic-Data-for-AI-Equality-website.pdf`, 2023. Accessed on January 29, 2024.

[6] Drago Plecko and Elias Bareinboim. Causal fairness analysis, 2022.

[7] Rose Porta, Alejandra Munoz Garcia, Margaret Bassney, and Aushanae Haller. Analysis of access to emergency funds in sub-saharan countries– a human rights-based approach. Manuscript under preparation. Department of Statistical and Data Sciences, Smith College, Northampton MA, 2022.

[8] Tim Räz. Group fairness: Independence revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 129–137, New York, NY, USA, 2021. Association for Computing Machinery.