MDPI

Article

Analysis of Access to Emergency Funds in Sub-Saharan Countries– A Human Rights-Based Approach

Rose Porta[1,‡,*], Alejandra Munoz Garcia[1,‡], Margaret Bassney[1,‡], Aushanae Haller[1,‡]

[1] Department of Statistical and Data Sciences, Smith College, Northampton MA;

* Correspondence: rporta@smith.edu

† These authors contributed equally to this work.

Version November 12, 2022 submitted to Water

Simple Summary: A Simple summary goes here.

Abstract: Having access to emergency funds is a valuable resource that many people end up needing at least once in their lives. Those who have access to emergency funding and other financial services have the capacity to remain afloat when unexpected predicaments arise, while those who are without this privilege have no choice but to endure crises and simply hope for the best. The purpose of our project is to analyze the access adults have to emergency funds and financial services in Sub-Saharan countries using a 2017 dataset from the Global Findex Database. Additionally, an important goal of our project is to employ a variety of different approaches in an attempt to minimize bias and maximize fairness, particularly when examining the performance for males and females. We also aim to determine how adults in the Sub-Saharan African region access financial services as well as establish the amount of bias we have within our models using exploratory data analysis, a baseline model, and a variety of fairness metrics. We hope to implement our findings in a Jupyter notebook where this information can be made accessible to a broader undergraduate audience.

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

1. Version

This Rmd-skeleton uses the mdpi Latex template published 2019/02. However, the official template gets more frequently updated than the 'rticles' package. Therefore, please make sure prior to paper submission, that you're using the most recent .cls, .tex and .bst files (available here).

2. Introduction

Science is often viewed as a way to offer trustworthy research backed solutions and answers. A lot of that research involves statistical methods performed on data however, what happens when the data and statistical methods are not as objective and trustworthy as is so often assumed? The conclusions drawn from the data are biased and unfair, most often towards minorities and protected classes of people. To contribute to a human rights based approach to data analysis, we evaluate fairness metrics on a machine learning algorithm to measure bias. We use a Global Findex data set which contains financial information about 35 Sub Saharan countries. Specifically, we create models to predict access to emergency funds, then analyze the fairness of those models. We focus on group and individual fairness metrics for the protected attribute sex. In addition we investigate the data set itself to understand where potential biases might have been implanted.

Data sets and algorithms have real world impacts on real people. The inherent bias in data sets can carry over into machine learning algorithms that are used to profile and categorize people [1,2]. Since data set's are not collected in a vacuum and often represent the discriminatory environments in which they are collected [3], we must find ways to make data sets and statistical methods more equitable. In this study we explore fairness methods that can be used to evaluate machine learning models. The "impossibility theorem" is the idea that not all fairness metrics can be satisfied at the same time[4]. Although fairness is complex and there are multiple approaches to make a model fair [5,6], it's important to continue to question how data and algorithms can be biased and how to mitigate that bias.

While there have been previous studies implementing fairness techniques in different contexts[7,8], we implement them in an exploratory context meant to teach how and when to use these techniques thus giving us more freedom to branch beyond a specific question while supporting previous work about the importance of these fairness metrics[3,9]. We analyse the data, data collection methods, prediction models, and the fairness metrics to assess how biased our data is and understand how we can de-bias when possible.

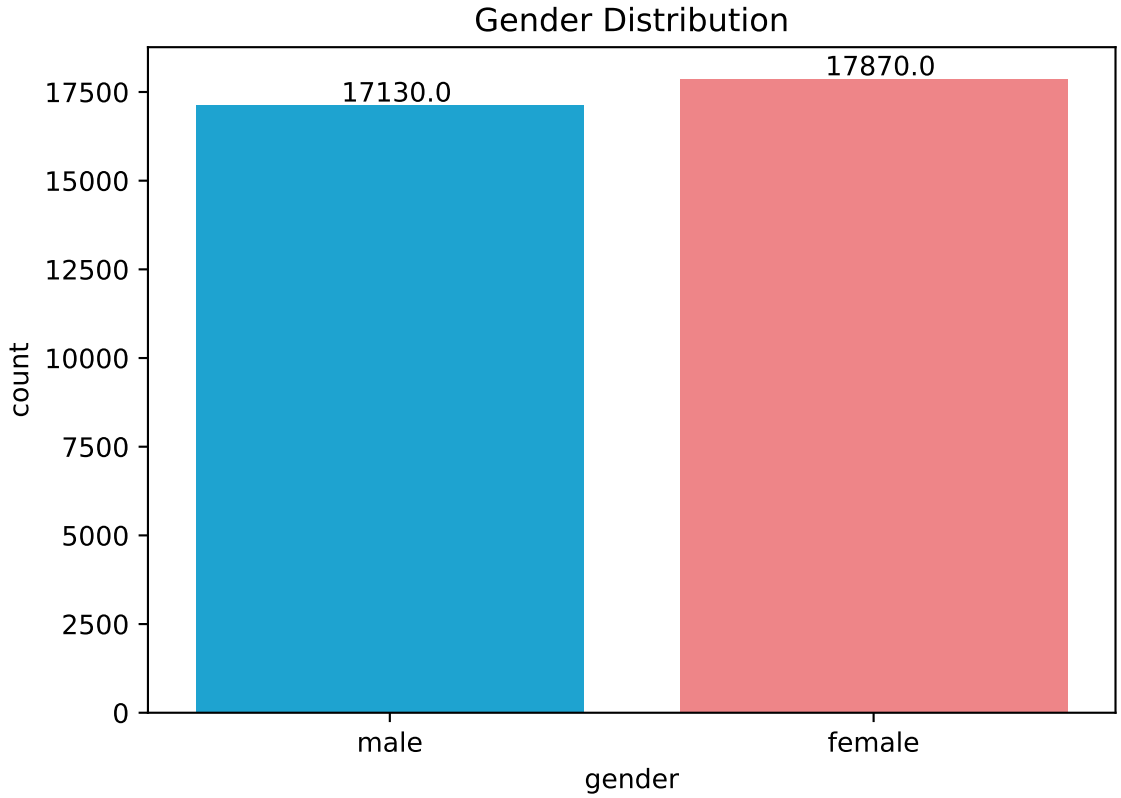
3. Data

Our data is derived from The World Bank in The Global Findex Database, comprising the most comprehensive data sets on how adults save, borrow, make payments, and manage risk in more than 140 economies around the world. The data set was created to record various measures of financial equity and inclusion, with the intention that such information could reveal opportunities to expand access to financial services and to promote greater use of digital financial services for individuals who do not have a bank account. Conducted by Gallup, Inc for the annual Gallup World Poll, the participants responded to the questionnaire either on the phone or in-person. There were several variables of interest in this dataset when creating models to predict access to emergency funds, including demographic and financial information. For this analysis, we are using only a subset of the data including countries in the Sub-Saharan region (35 countries total). Our data set includes 35000 observations and 105 variables in total.

3.1. Demographics

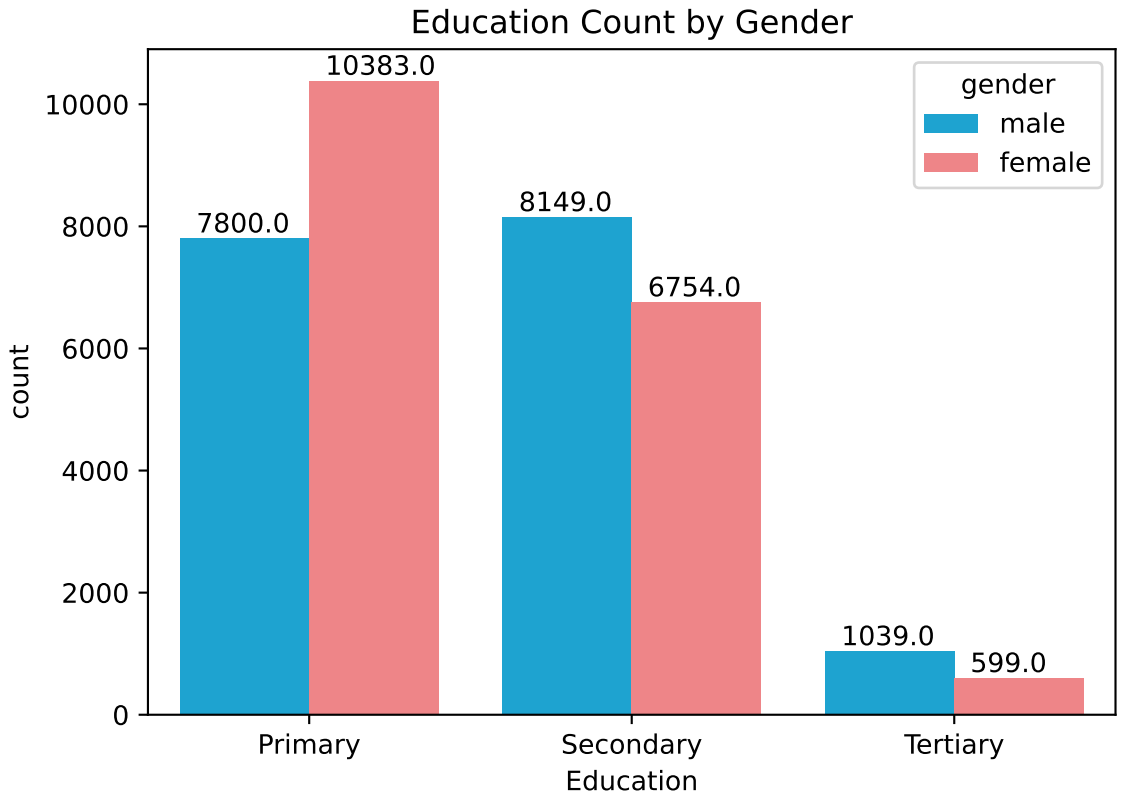
3.1.1. *gender*

The variable *gender* distinguishes gender. There are 16,716 males in this dataset and 17,388 females. This is a fairly equal distribution that we can see in the graph below.



3.1.2. *Education*

The *Education* variable corresponds to the highest level of education attained with 'Primary', 'Secondary' and 'Tertiary' being the three options. Here is the distribution of education by gender:



The bar plot above shows us that there are more women with primary education, but more men with secondary or tertiary education. Overall, we can see that there are more men with higher education than women. About 1,000 more men have received a secondary education and there is about double the amount of men with tertiary education compared to women showing a clear disparity.

3.1.3. *economy*

The final demographic variable of interest is the *economy* variable that separates respondents by which country they live in. There are 35 different countries from Sub-Saharan Africa with exactly 1000 respondents from each.

```
df2['economy'].unique()
```

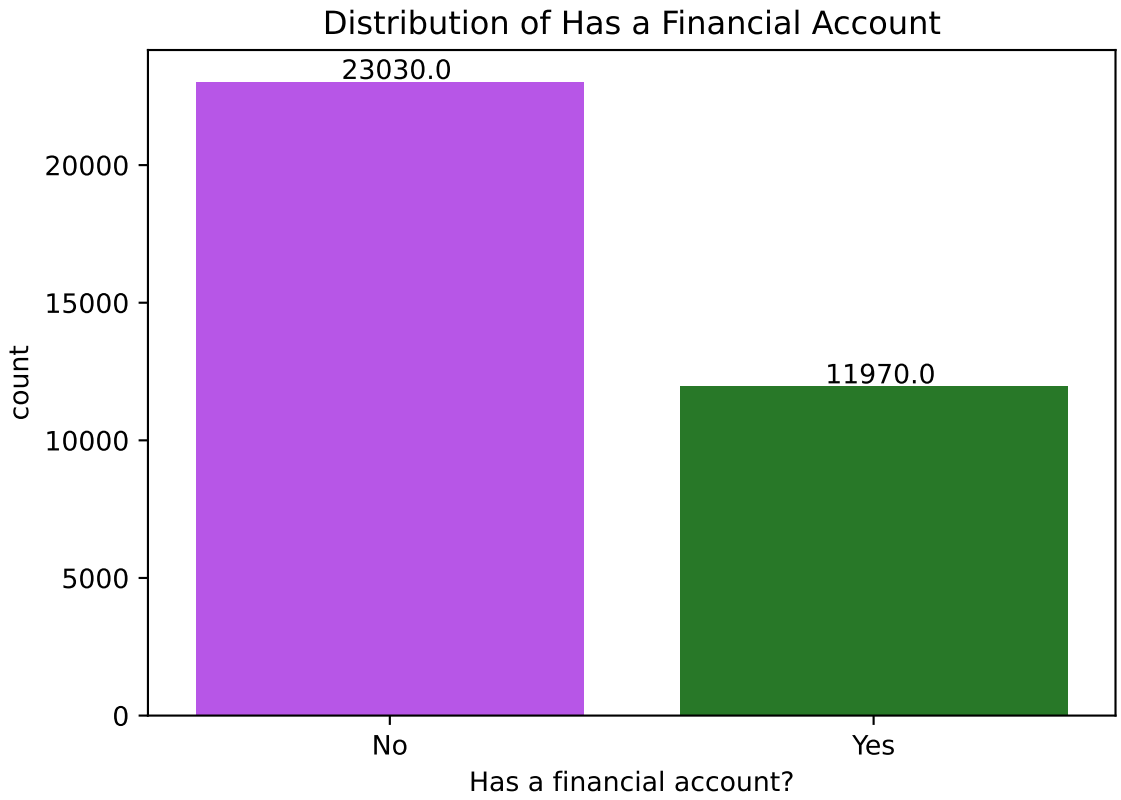
```
## array(['Benin', 'Botswana', 'Burkina Faso', 'Cameroon',
##        'Central African Republic', 'Chad', 'Congo, Dem. Rep.',
##        'Congo, Rep.', "Cote d'Ivoire", 'Ethiopia', 'Gabon', 'Ghana',
##        'Guinea', 'Kenya', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi',
##        'Mali', 'Mauritania', 'Mauritius', 'Mozambique', 'Namibia',
##        'Niger', 'Nigeria', 'Rwanda', 'Senegal', 'Sierra Leone',
##        'South Africa', 'South Sudan', 'Tanzania', 'Togo', 'Uganda',
##        'Zambia', 'Zimbabwe'], dtype=object)
```

3.2. Financial

From the financial related variables, we were most interested in a few specific financial variables that we thought would have an impact on access to emergency funds.

3.2.1. *account_fin*

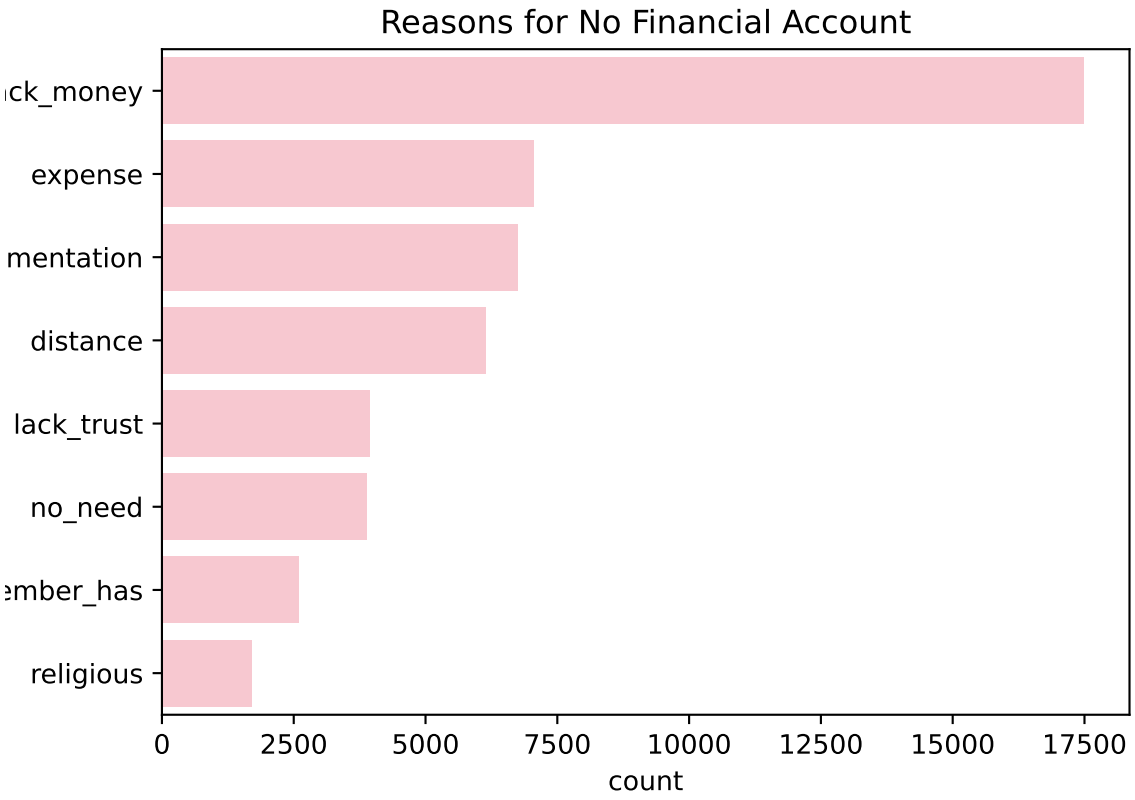
The first variable being *account_fin* which distinguishes those who have a financial account from those who don't:



We can see that about two thirds of individuals do not have an account. This is likely connected to the lack of access to emergency funds displayed above given that if an individual does not have a financial account, we would expect they are less likely to have a source of emergency funds, as emergency funds are generally stored in an account.

3.2.2. *reason*

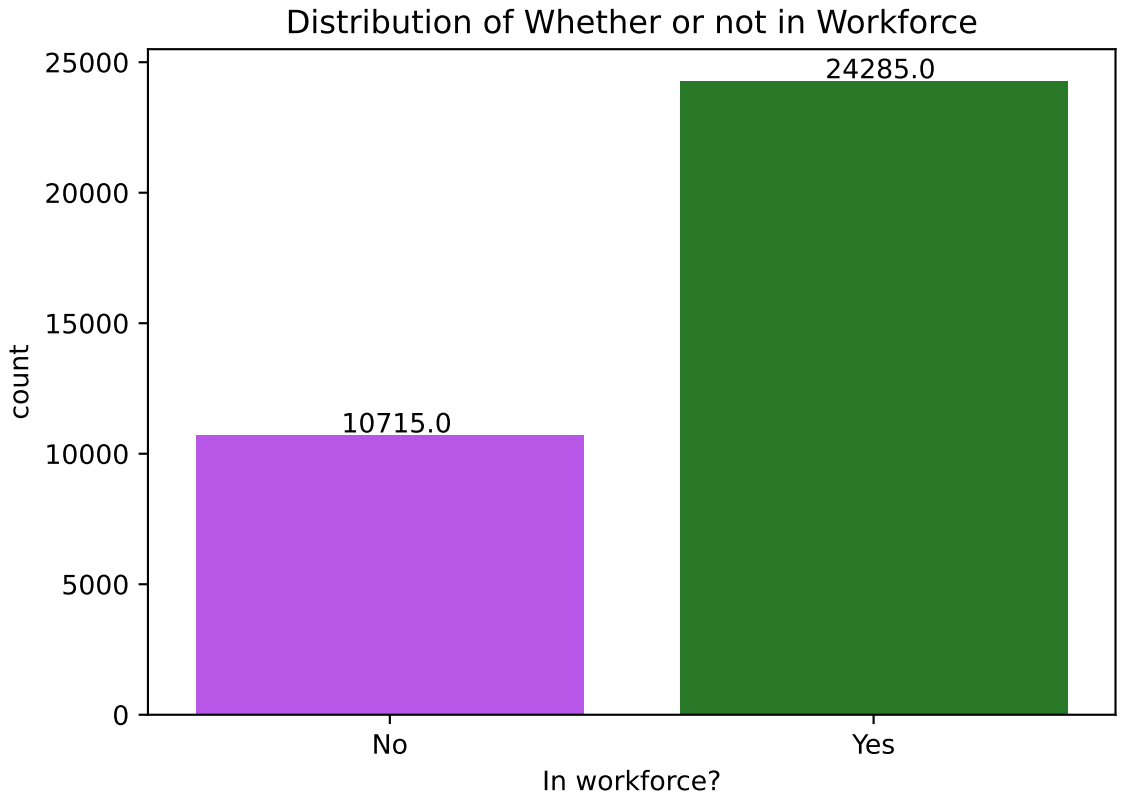
Those who do not have a financial account were asked why in the *reason* variable, that provides a list of possible reasons for not having a financial account:



97

98 3.2.3. *emp_in*

99 Employment status was another financial variable of interest represented by *emp_in*, which asks
100 whether or not the participant is in the workforce. It appears that about three-fourths of individuals
101 are in the workforce:



3.2.4. *inc_q*

And lastly, we evaluated *inc_q*, which represents income quantile. Income quantile is separated into 5 quantiles with 1 being the poorest and 5 being the richest. The mean for all of the countries in the dataset is 3.241. This means that all the countries average out to be about middle class.

```
## 3.241085714285714
```

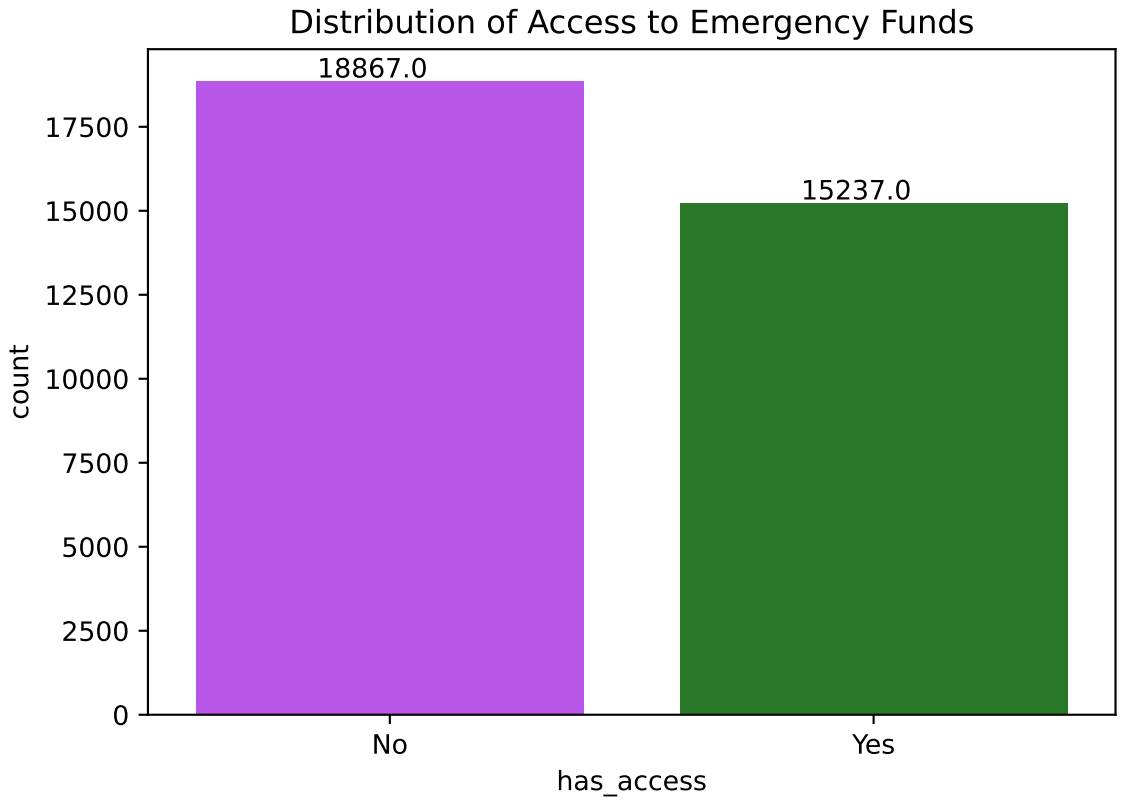
The majority of the data set has individuals within the richest quantile, Quantile 5.

3.3. Emergency Funds

To explore access to emergency funds in our dataset, we were interested 3 variables we thought could be related:

3.3.1. *has_access*

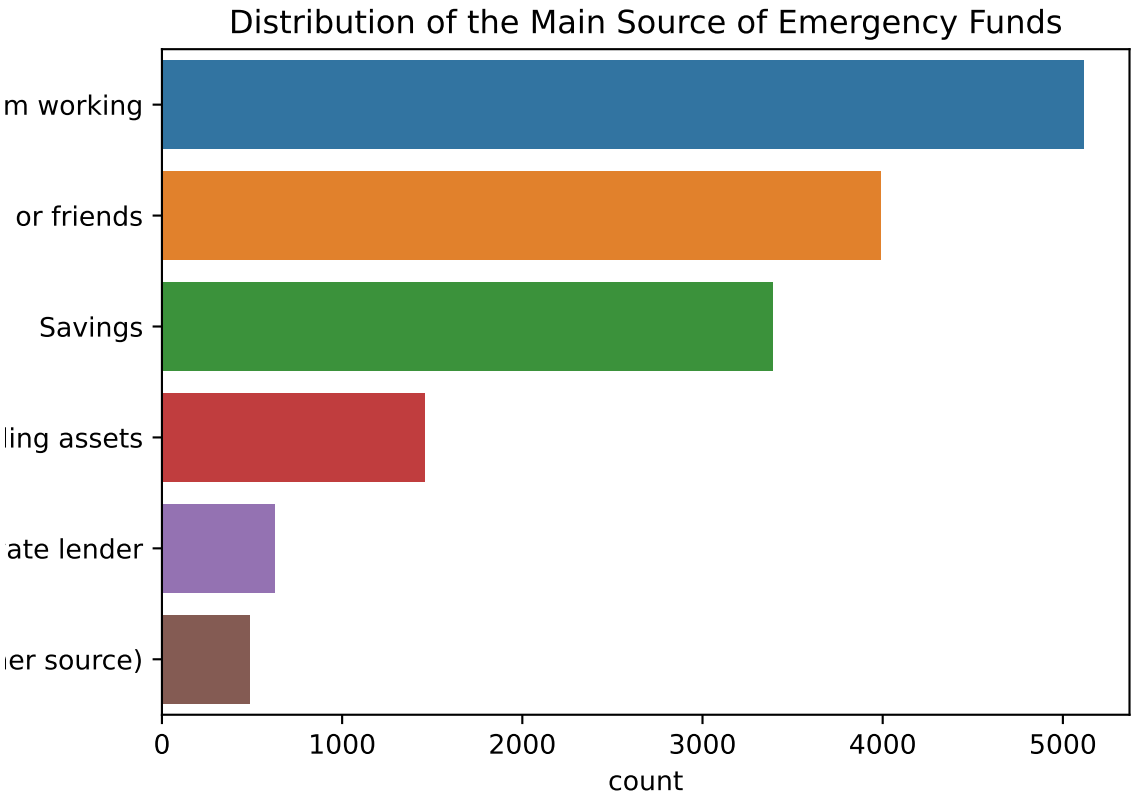
The variable *has_access* directly asks participants if they have access to emergency funds, with "emergency funds" defined as 1/20th of the GNI (gross national income) per capita for the country. GNI per capita is the country's total income in a year/ the country's population size. For context, in the United States, "emergency funds" would be defined as about $3,000.



The barchart above displays the overall distribution of access to emergency funds. We can see that over half of individuals represented in the data do not have access.

3.3.2. *main_source_funds*

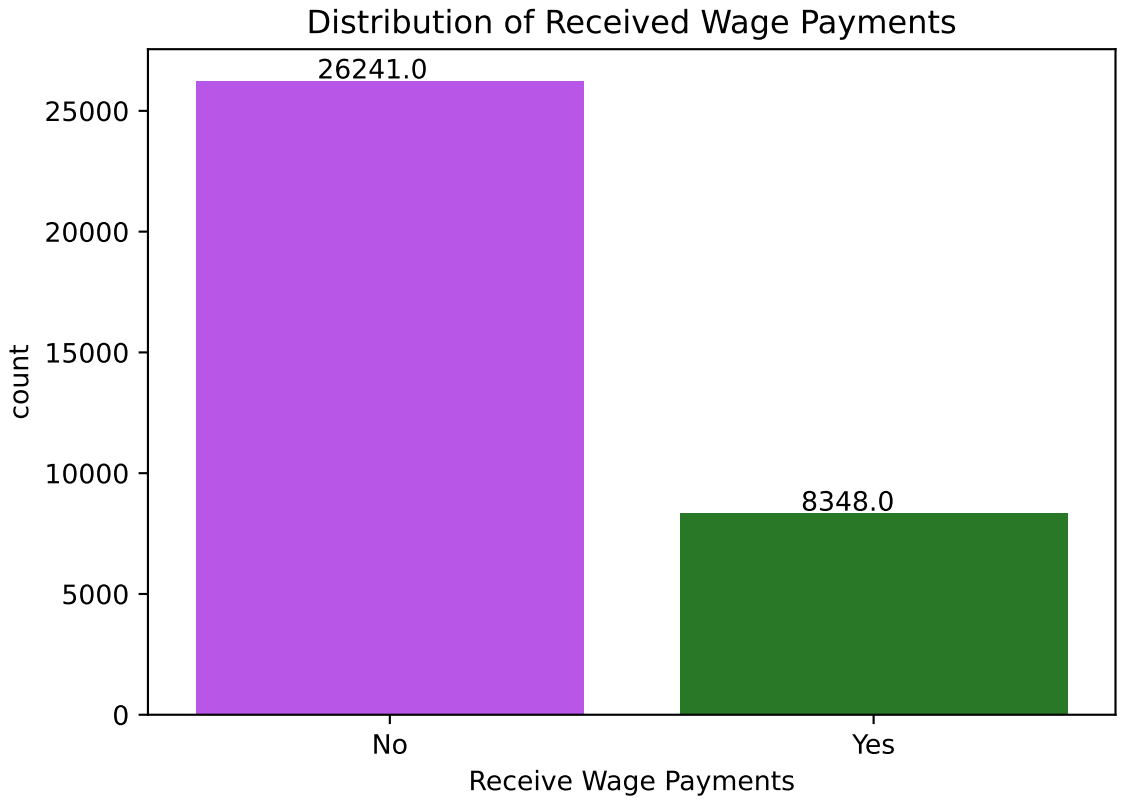
We proceeded to explore the source of emergency funds using the *main_source_funds* variable, which provides a list of options for where participants receive their main source of emergency funds:



The barchart above displays the overall distribution of the main source of emergency funds. Most of the individuals with access to emergency funds receive their funding from work, their family and friends, or their savings.

3.3.3. *Recieve Wage Payments*

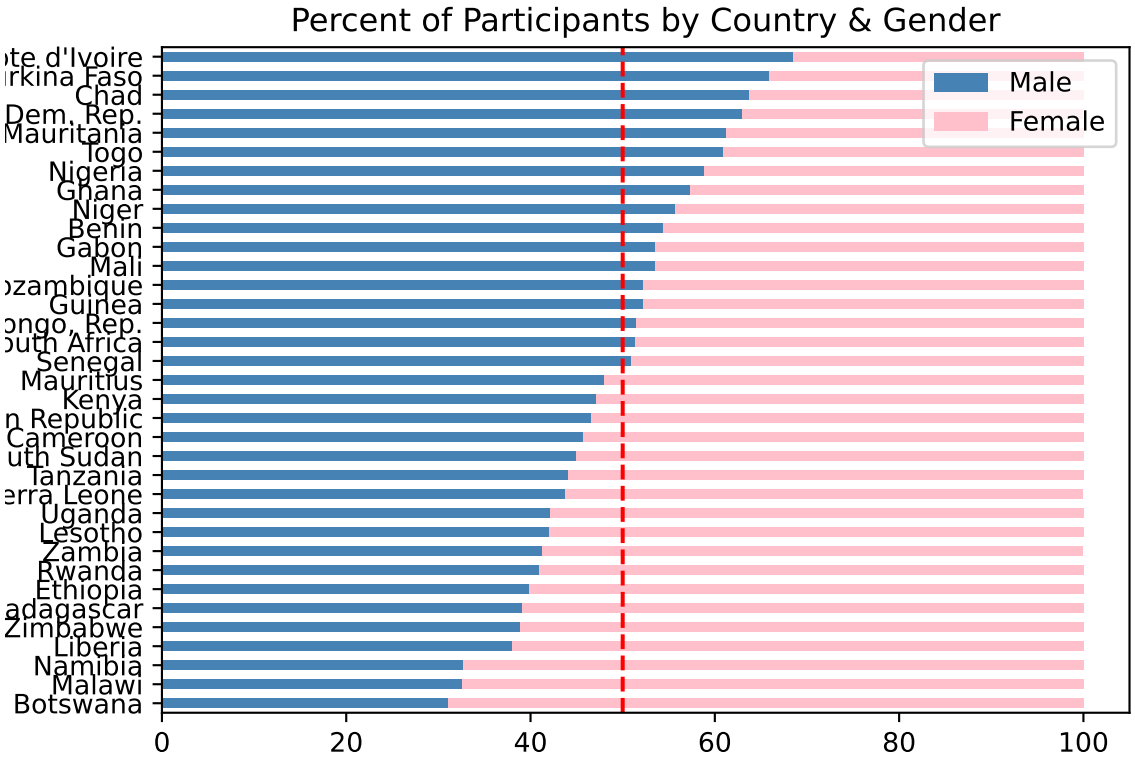
Diving further into the "Money from Working" category, we can see that only 8196 individuals receive wage payments from the *Receive Wage Payments* variable. This analysis suggests that receiving wage payments may be a key factor in determining access to emergency funds.

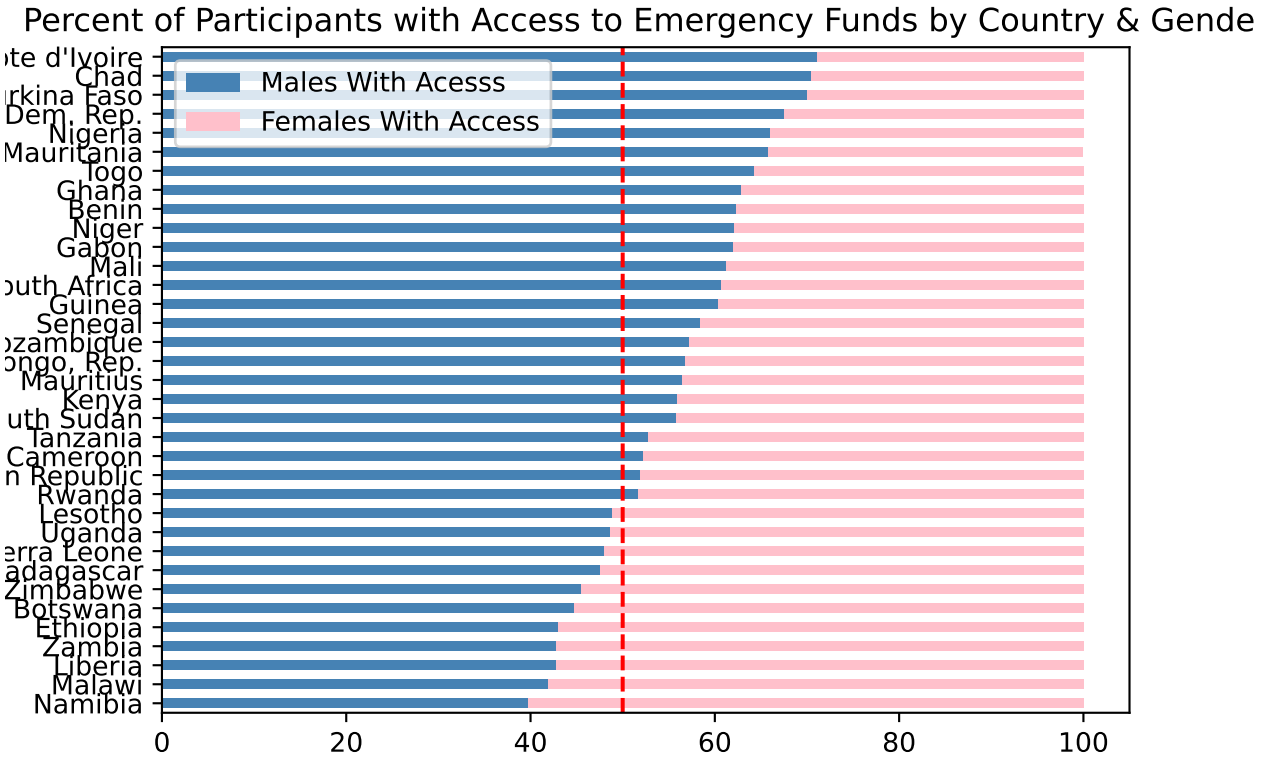


131

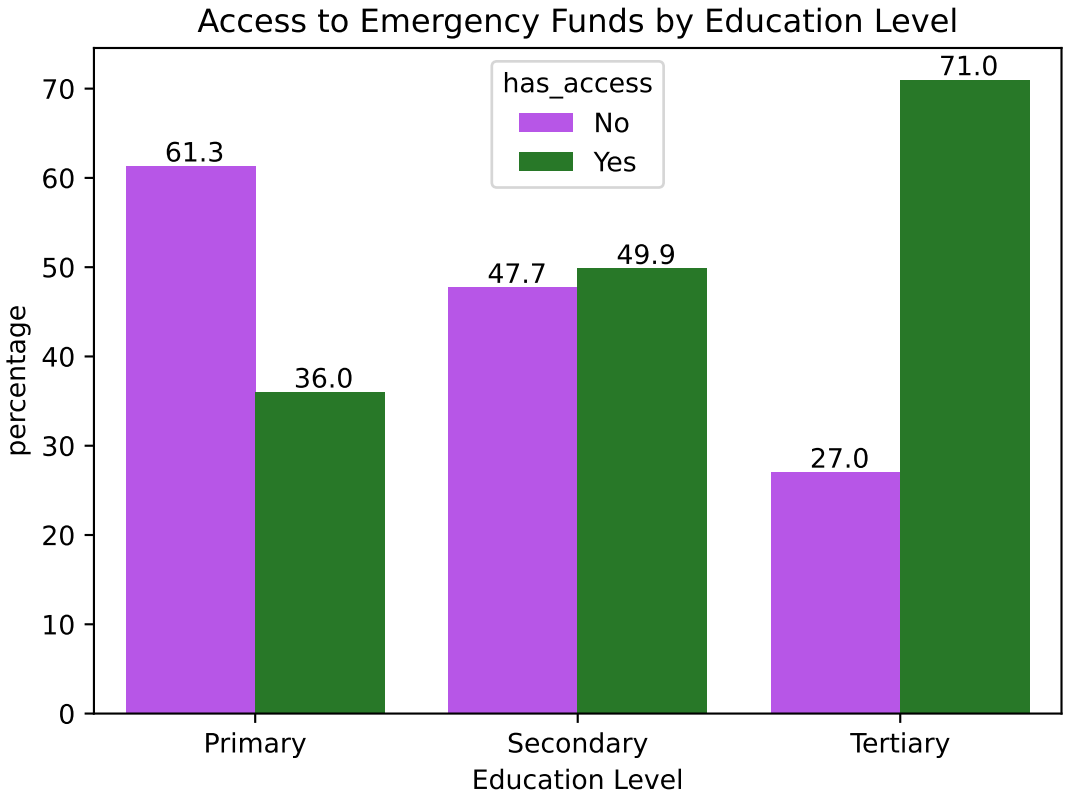
132 3.3.4. *gender*, *economy* and *Education* in relation to Emergency Funds

133 Finally, we sought to find if there were disparities in access to emergency funds by *gender*, *economy*,
134 and *Education*.


Percent of Participants by Country & Gender
Male
Female
Botswana
Malawi
Namibia
Liberia
Zimbabwe
Ethiopia
Rwanda
Zambia
Lesotho
Uganda
Tanzania
Cameroon
Guinea
Senegal
Mauritius
Kenya
Gabon
Mali
Benin
Niger
Ghana
Nigeria
Togo
Mauritania
Chad
0
20
40
60
80
100


Percent of Participants with Access to Emergency Funds by Country & Gende
Males With Acesss
Females With Access
Botswana
Malawi
Namibia
Liberia
Zimbabwe
Ethiopia
Rwanda
Zambia
Lesotho
Uganda
Tanzania
Cameroon
Guinea
Senegal
Mauritius
Kenya
Gabon
Mali
Benin
Niger
Ghana
Nigeria
Togo
Chad
0
20
40
60
80
100

In the side-by-side barplots above, we can see that although only about 50% of the countries have a higher percentage of men represented in the questionnaire (left bar plot), in 75% of the countries more men have access to emergency funds than women (right bar plot).



Additionally, in the barplot above we can see the distribution of funds based on an individual's highest education level. 63% of people with only a primary education do not have access to emergency funds compared to 37% of people who do. These numbers are more evenly distributed for those with secondary education, with about 49% of people not having access to emergency funds, while 51% of people do have access. Finally, for those with a tertiary level of education we can see that about 72% of people have access to emergency funds while only 28% of that group does not have access. Overall, we can make the assumption that people with a higher level of education are more likely to have access to emergency funds.

4. Methods

4.1. Software

We conducted our analysis in a Google Colab Notebook primarily employing the python libraries pandas and numpy for data cleaning and exploratory analysis, as well as sklearn and aif360 for implementing machine learning methods, fairness metrics, and de-biasing techniques. The notebook detailing our full analysis is available via our public GitHub repository.

4.2. Data Cleaning

Before fitting the model, we performed several pre-processing steps on the data in order to remove unnecessary or redundant information, address missing values, and ensure that the variables were coded such that they would be processed appropriately by the model. Some of these steps were

performed after fitting the initial model, and some choices were made based on the impacts of those choices on the performance of the model.

First, we removed the arbitrary variables *economycode* (country code) and *regionwb* (region); *economycode* is essentially a duplicate of *economy* (country name), and *regionwb* is the same for all rows (value is Sub-Saharan) since this is the variable that we initially filtered by. Next, we checked for variables with a high percentage of missing values. Several variables have many missing values because they are follow-up questions to a previous question that are only asked if the respondent gives a specific for a previous question. We chose to drop all variables with more than 30% missing values (58 variables total) because we observed by running the model multiple times that variables with NA percentages above this threshold had no impact on the model accuracy or fairness. Thus, it made sense to remove them if their presence is negligible when included in the model. Next, we checked for variables with high levels of redundancy (i.e. little variation), defining a high level of redundancy as 95% or more of the values being the same. The only variable with a high level of redundancy was *pay_online* (a binary variable indicating whether or not an individual has paid online for something), and we chose to remove it because removing it had no impact on the model accuracy and fairness.

Many of the questions in the survey were structured such that the potential responses were "yes", "no", "don't know", and "refuse". Based on our exploratory data analysis, it seems that for most questions, the numbers of "don't know" and "refuse" responses are very low. Furthermore, these responses would not give us much useful information when implementing a predictive model. Thus, we chose to replace all "don't know" and "refuse" values with *NA* values. We then replaced all remaining missing values (including missing values not removed previously and "don't know" and "refuse" values) with the column mean rounded to the nearest whole number (i.e. the most frequent value if the variable is categorical).

Next, we re-coded the country variable into a variable with five categories based on the percentage of sampled individuals in the country who have access to emergency funds (1 = < 20% have access, 2 = between 20% and 40% have access, etc.). All other variables in our data set were coded such that the model could appropriately interpret them, so we did not have to do any additional re-coding. The majority of the variables are binary with 0 = no, 1 = yes, and the rest are either categorical variables with hierarchical categories such as income quantile and education level or continuous numerical variables such as age.

Finally, we created a cartesian product to combine two variables – one indicating whether or not an individual has a financial account and the other indicating whether or not the individual has saved money in the past 12 months– in order to increase the accuracy of the model after fitting the initial model.

After this pre-processing, we have 42 predictor variables and 35,000 observations.

4.3. Model Selection

Given that we are aiming to predict a binary outcome (possible or not possible to come up with emergency funds), our model needs to be a classification model. We tried two of the most common model types used for classification– logistic regression and decision tree classifier. We ultimately chose the decision tree classifier model over the logistic regression model because the baseline accuracy was higher (61% versus 55%). Furthermore, the decision tree model makes more intuitive sense given our data since most of our predictors are binary variables, and binary predictors fit well into the tree structure. We were able to further improve the accuracy of the decision tree model to 68% by using cross-validation to specify the max depth as 6. The decision tree model uses machine learning to predict outcomes by organizing the variables into a tree that branches off at each decision point based on the value of the variable at that point. The most influential variables are at the top, and the outcome variable is at the end of each branch. We split our data into 70% training, 30% testing because this is the standard train-test split used for machine learning algorithms.

5. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.[3]

5.1. Subsection Heading Here

Subsection text here.

5.1.1. Subsubsection Heading Here

Bulleted lists look like this:

- First bullet
- Second bullet
- Third bullet

Numbered lists can be added as follows:

1. First item
2. Second item
3. Third item

The text continues here.
All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

Figure 1. This is a figure, Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Table 1. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	Title 2	Title 3
entry 1	data	data
entry 2	data	data

This is an example of an equation:

$$\mathbb{S} \tag{1}$$

Example of a theorem:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:
Example of a proof:

Proof of Theorem 1. Text of the proof. Note that the phrase 'of Theorem 1' is optional if it is clear which theorem is being referred to. □

The text continues here.

6. Discussion

Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

7. Conclusion

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

8. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Acknowledgments: All sources of funding of the study should be disclosed. Please clearly indicate grants that you have received in support of your research work. Clearly state if you received funds for covering the costs to publish in open access.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "X.X. and Y.Y. conceive and designed the experiments; X.X. performed the experiments; X.X. and Y.Y. analyzed the data; W.W. contributed reagents/materials/analysis tools; Y.Y. wrote the paper.'' Authorship must be limited to those who have contributed substantially to the work reported.

Conflicts of Interest: Declare conflicts of interest or state 'The authors declare no conflict of interest.' Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funding sponsors in the design of the study; in the collection, analyses or interpretation of data in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state 'The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, an in the decision to publish the results'.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text. For example, explanations of experimental details that would disrupt the flow of the main text, but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with 'A', e.g., Figure A1, Figure A2, etc.

References

1. Navarro, C.L.A.; Damen, J.A.; Takada, T.; Nijman, S.W.; Dhiman, P.; Ma, J.; Collins, G.S.; Bajpai, R.; Riley, R.D.; Moons, K.G.; others. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *bmj* **2021**, *375*.
2. Hellström, T.; Dignum, V.; Bensch, S. Bias in Machine Learning–What is it Good for? *arXiv preprint arXiv:2004.00686* **2020**.
3. Barocas, S.; Hardt, M.; Narayanan, A. Fairness and machine learning.
4. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* **2016**.
5. Kypraiou, S. What is Fairness? **2021**. Publisher: PubPub.
6. Green, B.; Hu, L. The myth in the methodology: Towards a recontextualization of fairness in machine learning. Proceedings of the machine learning: the debates workshop, 2018.
7. Deho, O.B.; Zhan, C.; Li, J.; Liu, J.; Liu, L.; Duy Le, T. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology* **2022**.
8. Kim, J.Y.; Cho, S.B. An Information Theoretic Approach to Reducing Algorithmic Bias for Machine Learning. *Neurocomputing* **2022**.
9. Anahideh, H.; Asudeh, A.; Thirumuruganathan, S. Fair active learning. *Expert Systems with Applications* **2022**, *199*, 116981.

Sample Availability: Samples of the compounds ..... are available from the authors.