



March 2022



## Machine learning for applied gender analysis: a hypothesis generation research technique

In the past decade, there has been a growing interest in using machine learning (ML) approaches for public health and social science research, applied to topics such as outcome risk factor identification, disease prediction, and health care resource allocation. [1]. The challenges to in-person data collection triggered by the COVID-19 pandemic have further catalyzed interest in ML applications with big data (such as social media, remote sensing, and other data from online sources) in particular [2]. ML models can be particularly useful in addressing some of the limitations associated with traditional statistical models, such as multi-collinearity, restrictions of small prevalence outcomes or unbalanced datasets, and other challenges associated with large numbers of independent variables.

To date, ML approaches have received limited application on issues specific to women's health or gender. In the current brief, we discuss our ML related work to understand gender issues. We describe an exploratory research technique, Iterative Thematic Analysis(ITA), which integrates ML models with qualitative coding methods. We have used ITA to identify correlates of multiple gender related outcomes, such as child marriage, non-marital sexual violence, marital sexual violence, and help-seeking behavior for marital violence [3-6]. These outcomes are key areas of research, yet less understood in terms of known risk factors. Additionally, most have very low prevalence, making the use of traditional regression models challenging to identify correlated variables.

Our work focuses on India, and uses a large dataset- the Demographic Health Survey, also known as the National Family Health Survey (NFHS-4) in India. NFHS-4, a nationally representative survey was conducted in 2015-16, and interviewed nearly 700,000 women between ages 15 to 49 on a wide range of topics related to their health and overall well-being. A subset of these women (nearly 80,000), were also asked about their experiences of violence. The NFHS-4 dataset includes approximately 5000 variables. We adopt an exploratory approach, aiming to identify potential correlates of a given outcome from all collected. ITA is a tool for hypothesis generation, which uses a large pool of both theoretically related and unrelated variables within a population-level dataset to identify potential correlates for an outcome. These correlates can then be further tested and studied by more localized research.

We developed the ITA technique through a process of collaboration and deliberation that included experts in gender research, survey data analysis and computer science.

Our work uses two types of machine learning models

Both are forms of regularized regression models. Regularization imposes a penalty on the size of regression coefficients by trying to shrink them towards zero.

**L1- regularized regression model/ lasso model**

Useful for variable selection and shrinkage. Can force some coefficient estimates to be exactly equal to zero – ideal for removing irrelevant variables

**L2- regularized regression model/ ridge model**

Shrinks coefficient values towards zero, but never to an absolute zero. Suitable when working with variables that have some known potential relationships.

Figure 1: Two types of machine learning models used in the Iterative Thematic Analysis (ITA) method

**ITERATIVE THEMATIC ANALYSIS (ITA) – A SYSTEMATIC APPROACH TO IDENTIFYING CORRELATED THEMES:** ITA uses two different types of ML models : lasso and ridge models (Figure 1). We run these two models, lasso followed by ridge, using the strategy of ITA. Figure 2 depicts the steps in ITA. We first run a lasso model with the complete dataset, i.e., over 5000 variables as the independent variables and our outcome of interest as the dependent variable. We use lasso since it is useful in identifying irrelevant variables by shrinking their coefficients to zero. The variables with non-zero coefficients from lasso are then included in a ridge regression model.

1. Mhasawade, V., Zhao, Y., & Chunara, R. (2021). Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8), 659-666.
2. McDougal, L., Raj, A., Yore, J., Boyd, S., Penumetcha, N., Pryor, E. C., ... & Gaddis, I. (2021). Strengthening gender measures and data in the COVID-19 era: an urgent need for change. Available form: [https://data2x.org/wp-content/uploads/2021/03/COVID-19\\_Gender\\_Data\\_and\\_Measures\\_Evidence\\_Review\\_FINAL.pdf](https://data2x.org/wp-content/uploads/2021/03/COVID-19_Gender_Data_and_Measures_Evidence_Review_FINAL.pdf). [Last accessed 2021 Sep 05].
3. Raj, A., Dehingia, N., Singh, A., McDougal, L., & McAuley, J. (2020). Application of machine learning to understand child marriage in India. *SSM-Population Uealth*, 12, 100687.
4. Raj, A., Dehingia, N., Singh, A., McAuley, J., & McDougal, L. (2021). Machine learning analysis of non-marital sexual violence in India. *EClinicalMedicine*, 39, 101046.
5. McDougal L, Dehingia N, Bhan N, Singh A, McAuley J, Raj A. (2021). Opening closed doors: Using machine learning to explore factors associated with marital sexual violence in India. *BMJ Open*.11:e053603.
6. Dehingia N, Dey AK, McDougal L, McAuley J, Singh A, Raj A. (2022) Help seeking behavior by women experiencing intimate partner violence in India: a machine learning approach to identifying risk factors. *PLoS One*. PLoS ONE 17(2): e0262538.



