Evaluation of machine learning methods for predicting the risk of child mortality in South Africa

Chodziwadziwa Kabudula, Reesha Kara, Henry Wandera, Fidelia A. A. Dake, Justin Dansou, Dereje Danbe Debeko, Chipo Mufudza

Abstract

There has been extensive research on child mortality in sub-Saharan Africa. But, most of the methods applied thus far have relied on conventional regression analyses with limited prediction capability. Emerging methods in computational social science, particularly machine learning, present opportunities to identify critical features in different domains to facilitate accurate prediction of the risk of child mortality in sub-Saharan Africa. We evaluated different methods of machine learning techniques to develop the best model for predicting child mortality using training and test data from the National Income Dynamics Survey and District Health Barometer in South Africa. Logistic, Random Forest and XGBoost all show accuracy, sensitivity and specificity of about 60%. Further analysis will be explored using data from different countries with different features.

Introduction

Globally, mortality among children under five remains a major development goal and a key target of the sustainable development goals (1). Although there have been significant gains in reducing child mortality globally, the rate of decline differs across the major regions of the world. Sub-Saharan Africa is reported to have the highest burden of child mortality in the world (2). The average under-five mortality rate in sub-Saharan Africa in 2017 was 76 deaths per 1,000 live births compared to the global rate of 39 per 1,000 live births (2). Again the rate of under-five mortality in sub-Saharan Africa is much higher when compared to regions such as Australia and New Zealand where the rate in 2017 was 4 per 1,000 live births. These rates translate into much lower chances of survival for children born in sub-Saharan Africa. While only 1 in 263 children in Australia and New Zealand may die before their fifth birthday, in sub-Saharan Africa, every 1 in 13 children may die before dying before his or her fifth birthday. Furthermore, the risk of mortality differs by age, particularly in the neonatal and post neonatal periods and the neonatal mortality rate is highest in sub-Saharan Africa, and Central and Southern Asia at 27 and 26 neonatal deaths per 1000 live births respectively (2).

The underlying risk factors for under five mortality have been extensively examined using several different methods, mostly conventional regression (3), decomposition analyses (4) and survival analyses (5,6). These methods however, fail to explain to a full extent the factors that account for the variation in child mortality (7). While several factors including demographic, socioeconomic, environmental and community have been identified as risk factors for child mortality in sub-Saharan Africa, there is still a gap in knowledge particularly risk of mortality for an unborn child. New methods in artificial intelligence, including machine learning present opportunities for enhanced prediction of the risk of mortality as well as identifying at risk groups for targeting of specific interventions.

New methods in artificial intelligence, including machine learning present opportunities for enhanced prediction of the risk of mortality as well as identifying at risk groups for targeting of specific interventions. Furthermore, recent development of big data technologies such as machine learning has made possible the ability to provide more accurate estimates of statistical analyses (8). Supervised machine learning employs classification algorithms to explain the outcome variable in terms of the independent variables (9). While conventional methods investigate associations based on hypothesis testing, machine learning elucidates patterns in the set of predictor variables that identify the dependent variable (9). Additionally, “machine learning algorithms automatically scan and analyze all predictor variables in a way that prevents overlooking important predictor variables even if it was unexpected”(9:1).

Applying machine learning techniques to researching child mortality in sub-Saharan Africa will be useful for identifying populations of children who are at risk of dying and predicting the underlying risks for each identified group of children. This study uses machine learning techniques to identify community, household and parental factors that are the strongest predictors for mortality in children under the age of five years and develop prediction and classification models for neonatal, infant and childhood mortality in sub-Saharan Africa.

Methodology

Data

This study combines data from two secondary sources from South Africa; the National Income Dynamics Survey (NIDS) and the District Health Barometer. The NIDS is a face-to-face longitudinal household survey of individuals living in South Africa. The survey collects demographic, socioeconomic and health data on children and adult members of surveyed households. This study uses data from the 2017 round of the survey. Mother and child pairs were constructed to link children under five to their mothers. Additionally, household characteristics including basic demographic characteristics such the composition, level of education and employment status of household members were extracted. The second data was sourced from the District Health Barometer which is a comprehensive statistical and analytical resource that provides an overview of health performance at the primary healthcare level, including hospitals in 52 district in South Africa. The District Health Barometer indicators for 2017/2018 were used to construct community level factors in health, education, child health, immunization coverage and environmental health among others.

Variables

The outcome variable for this study is child morality which as coded as 1 if a child under the age of 5 died within period of the survey and 0 if a child under the age of 5 did not die. The independent variables that were treated as features included characteristics of the mother (e.g. age, level of education and type of employment), household characteristics (e.g. location of the household whether rural or urban) and community characteristics including doctor to patient ration, population density and immunization coverage.

Methods of analysis

The study employed Machine learning (ML) approaches to find patterns and predict the risk of mortality. ML uses algorithms to discover patterns in data using variable and model selection methods (10). Both supervised and unsupervised approaches were employed in analysing the data. The unsupervised approach will be used to explore different patterns and groupings available in a heterogeneous environment including, community, household, and parental factors within South Africa. Specifically, clustering techniques will be used to group individuals with similar features to identify the different risk groups for child mortality[1]. Additionally, supervised ML will be used to predict child mortality. This can be accomplished via a range of

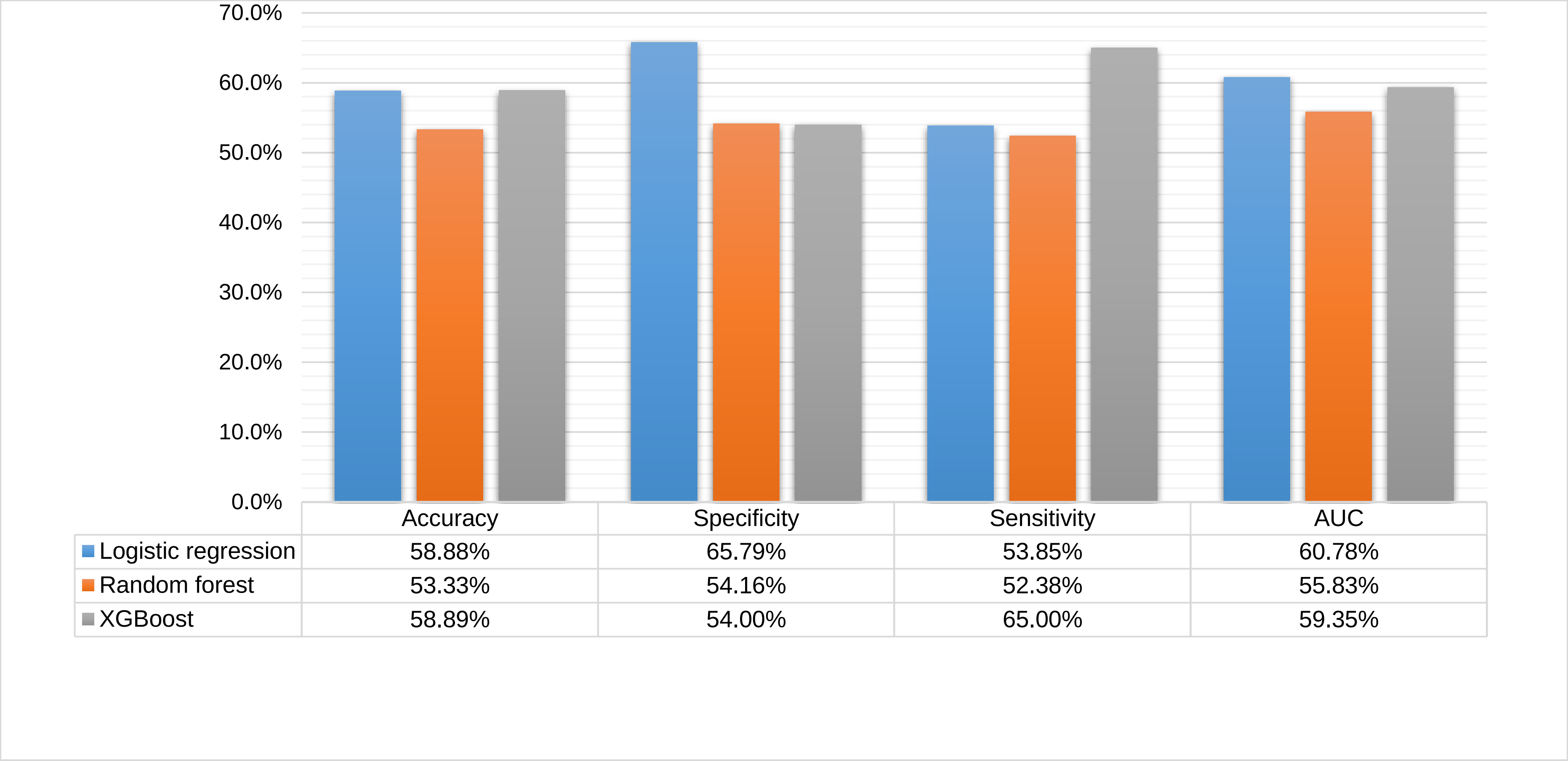
[1] There are different clustering algorithms in ML learning including heuristic and non-heuristic methods. The choice of the clustering algorithm depends on the function and purpose of the research, (Rodriguez et al., 2019). These can be rule based, function based or nonparametric including k-means density based clustering, Gaussian Mixtures, agglomerative, etc.

classification algorithms which include: Logistic regression, Support Vector Machines, Random Forest, K Nearest Neighbour and Naives Bayes.

Preliminary Results

The results presented in Figure 1 indicates that among the various methods employed, the logistic model exhibits the highest level of accuracy, specificity and proportion of the dependent variable that is correctly predicted by the model while the XGBoost approach shows the highest level of sensitivity. This notwithstanding, all the various approaches fail to achieve the minimum accepted level of accuracy and AUC.

Figure 1: Model performance indicators of factors predictors of child mortality in South Africa using different machine learning approaches



	Accuracy	Specificity	Sensitivity	AUC
🟦 Logistic regression	58.88%	65.79%	53.85%	60.78%
🟧 Random forest	53.33%	54.16%	52.38%	55.83%
■ XGBoost	58.89%	54.00%	65.00%	59.35%

References

1. Golding N, Burstein R, Longbottom J, Browne AJ, Fullman N, Osgood-zimmerman A, et al. Mapping under-5 and neonatal mortality in Africa , 2000 – 15 : a baseline analysis for the Sustainable Development Goals. Lancet [Internet]. 2017;390(10108):2171–82. Available from: http://dx.doi.org/10.1016/S0140-6736(17)31758-0

2. United Nations Children’s Fund, World Health Organization, The World Bank Group, United Nations. Levels and Trends in Child Mortality Report 2018. 2018.

3. Kujala S, Waiswa P, Kadobera D, Akuze J, Pariyo G, Hanson C. Trends and risk factors of stillbirths and neonatal deaths in Eastern Uganda (1982 – 2011): a cross-sectional , population- based study. 2017;22(1):63–73.

4. Malderen C Van, Amouzou A, Barros AJD, Masquelier B, Oyen H Van, Speybroeck N. Socioeconomic factors contributing to under-five mortality in sub-Saharan Africa : a decomposition analysis. 2019;1–19.

5. Id SY, Bishwajit G, Okonofua F, Id OAU. Under five mortality patterns and associated maternal risk factors in sub-Saharan Africa : A multi-country analysis. 2018;1–14.

6. Ezeh OK, Agho KE, Dibley MJ, Hall JJ, Page AN. Risk factors for postneonatal , infant , child and under-5 mortality in Nigeria : a pooled cross-sectional analysis. 2015;1–9.

7. Podda M, Bacciu D, Micheli A, Bellù R, Placidi G, Gagliardi L. A machine learning approach to estimating preterm infants survival : development of the Preterm Infants Survival Assessment (PISA) predictor. 2018;(August):1–9.

8. Wang P, Li YAN, Reddy CK. 1 Machine Learning for Survival Analysis: A Survey. 2017;X(X):1–39.

9. Sakr S, Elshawi R, Ahmed AM, Qureshi WT, Brawner CA, Keteyian SJ, et al. Comparison of machine learning techniques to predict all-cause mortality using fitness data : the Henry ford exercIse testing (FIT) project. 2017;1–15.

10. Buskirk TD, Kirchner A, Eck A, Signorino CS. An Introduction to Machine Learning Methods for Survey Researchers what are machine learning methods ? Surv Pract. 2018;11(1):1–10.

11. Rodriguez MZ, Id CHC, Casanova D, Bruno OM, Amancio DR, Costa LF, et al. Clustering algorithms : A comparative approach. 2019. 1–34 p.