Genus

ORIGINAL ARTICLE Open Access

Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey

Fikrewold H. Bitew[1*], Samuel H. Nyarko[1,2], Lloyd Potter[1,2] and Corey S. Sparks[1]

* Correspondence: fikre.wold@gmail.com
[1]Department of Demography, College for Health, Community & Policy, University of Texas at San Antonio, 501 W. Cesar Chavez Blvd, San Antonio, TX 78207, USA
Full list of author information is available at the end of the article

Abstract

There is a dearth of literature on the use of machine learning models to predict important under-five mortality risks in Ethiopia. In this study, we showed spatial variations of under-five mortality and used machine learning models to predict its important sociodemographic determinants in Ethiopia. The study data were drawn from the 2016 Ethiopian Demographic and Health Survey. We used three machine learning models such as random forests, logistic regression, and K-nearest neighbors as well as one traditional logistic regression model to predict under-five mortality determinants. For each machine learning model, measures of model accuracy and receiver operating characteristic curves were used to evaluate the predictive power of each model. The descriptive results show that there are considerable regional variations in under-five mortality rates in Ethiopia. The under-five mortality prediction ability was found to be between 46.3 and 67.2% for the models considered, with the random forest model (67.2%) showing the best performance. The best predictive model shows that household size, time to the source of water, breastfeeding status, number of births in the preceding 5 years, sex of a child, birth intervals, antenatal care, birth order, type of water source, and mother's body mass index play an important role in under-five mortality levels in Ethiopia. The random forest machine learning model produces a better predictive power for estimating under-five mortality risk factors and may help to improve policy decision-making in this regard. Childhood survival chances can be improved considerably by using these important factors to inform relevant policies.

Keywords: Machine learning, Under-five mortality, Determinants, Ethiopia

Introduction

Globally, an estimated 5.4 million children under the age of 5 are said to have died in 2017 alone (UNICEF, WHO, World Bank Group, and United Nations, 2018). Meanwhile, the global under-five mortality rate is said to have declined by 58%, from 93 deaths per 1000 live births in 1990 to 39 in 2017 (UNICEF, WHO, World Bank Group, and United Nations, 2018). Yet still, the under-five mortality rate in low-income countries was 69 deaths per 1000 live births in 2017—almost 14 times the rate in high-income countries (5 deaths per 1000 live births) (UNICEF, WHO, World Bank Group, and United Nations, 2018). It has been observed that more than half of these deaths are due to infectious diseases (such as pneumonia and diarrhea) that are preventable and treatable through simple, affordable interventions (World Health Organization, 2017).

Despite the considerable improvements over the past decades, sub-Saharan Africa remains the region with the highest level of under-five mortality in the world, with about half of the global under-five mortality burden (UNICEF, WHO, World Bank Group, and United Nations, 2018). Ethiopia appears to have the fifth-highest number of newborn deaths in the world, following India, Pakistan, Nigeria, and the Democratic Republic of Congo (UNICEF, 2017). It is estimated that about 472,000 children die in Ethiopia each year before their fifth birthday, which places Ethiopia sixth among the countries in the world in terms of absolute numbers of under-five deaths (Federal Ministry of Health, 2005). In Ethiopia, the under-five mortality rate has declined by two-thirds from the 1990 figure of 204 per 1000 live births to 58 per 1000 live births in 2016, and thus, achieving the target for Millennium Development Goal 4 (MDG 4) (You, Hug, Ejdemyr, Idele, et al., 2015). Despite this achievement, the under-five mortality rate in Ethiopia remains higher than those of many low and middle-income countries (LMIC).

Previous studies have provided much evidence on the socioeconomic and demographic factors that are associated with under-five mortality in Ethiopia (Ayele & Zewotir, 2016; Ayele, Zewotir, & Mwambi, 2017; Bereka, Habtewold, & Nebi, 2017), using traditional regression models. In this study, we predict the important determinants of under-five mortality in Ethiopia using non-traditional regression models drawing on nationally representative data. Specifically, we employed machine learning techniques to predict under-five mortality risks in the study sample. The main aim of this study is to show the spatial distribution of under-five mortality and the potential of machine learning algorithms in predicting important sociodemographic factors underlying the spatial variations in under-five mortality. As such, we initially develop a spatial visualization of the under-five mortality rate by region in Ethiopia. This is to visually highlight the spatial disparities in under-five mortality in the country while predicting the most important factors underlying these disparities. This study informs and strengthens appropriate extant policies or intervention strategies aimed at reducing under-five mortality in the country. It also underscores the potential role of the machine learning approach in demographic research.

Methods

Data source

This study is based on data from the 2016 Ethiopian Demographic and Health Survey (EDHS), the most recent in the demographic and health survey series that is conducted every five years. The EDHS is a nationally representative household survey that collects data on a wide range of population, health, and nutrition indicators to improve maternal and child health in Ethiopia (Central Statistical Agency (CSA) [Ethiopia],, and ICF International, 2016). The survey used a multi-stage stratified sampling technique based on the 2007 National Population and Housing Census of Ethiopia to select respondents from a total of 624 clusters (187 urban and 437 rural) (Central Statistical Agency (CSA) [Ethiopia],, and ICF International, 2016). The unit of analysis is under-five children with a total sample size of 10,641 selected from 645 clusters across Ethiopia. This is based on children's data obtained from retrospective information from mothers about their children that died before age 5 within the 5 years preceding the survey (2011 to 2016).

Study variables and measurements

In this study, the outcome variable—under-five mortality—was measured as a binary outcome. Thus, under-five mortality was measured as being alive (coded as 0) or dead (coded as 1) for all the models.

The predictors (features) used in this study include individual, household, community, and health service factors. The individual-level factors consisted of maternal and child characteristics. Maternal factors include mother's age at birth (< 20, > 20), education (no education, primary, secondary/higher), contraceptive use (yes/no), and mother's body mass index (BMI) (underweight/overweight and normal). Child factors include whether the child was wanted (child wanted then, wanted later, not at all), sex of the child, birth order (1–2, 3/later), births in last 5 years, and previous birth interval (< 2, 2–4, > 4 years), as well as whether the child was breastfed within 1 h of birth. The household factors used are the source of drinking water (improved/unimproved), time to the water source, toilet facility (improved/unimproved), and household wealth index (low, middle, high), and household size. The community factors comprised residence type (urban/rural) and geographical region (Tigray, Afar, Amhara, Oromia, Somali, Benishangul-Gumuz, Southern Nations Nationalities and People Region (SNNPR), Gambella, Harari, Dire Dawa, and Addis Ababa). The health service factors include antenatal visits (0, 1–4, 5+ visits), place/mode of delivery services (facility with caesarean section (CS) services, facility without CS, home), and postnatal visits within 2 months after delivery (yes/no). The selection of these predictor variables was based on information from existing literature on the subject (Aheto, 2019; Bereka et al., 2017; Yaya, Bishwajit, Okonofua, & Uthman, 2018).

Analytic strategy

The R programming language (version 3.6.0) and the caret package (Kuhn, 2020) was used to perform the data processing and analysis. We first developed a spatial map for crude under-five mortality rates by regions in Ethiopia to document the regional disparities in under-five mortality in the country. In this regard, we estimated the rates under-five mortality by region and then merged them with an Ethiopian regional shapefile before mapping it.

We also used three widely used machine learning (ML) algorithms—logistic regression, a Random Forest (RF), K-nearest neighbors (KNN) models—to predict under-five mortality determinants in Ethiopia and compared the results of the best algorithm to the results of the traditional logistic regression model. These three models were selected for various reasons. The logistic regression is typically used to analyze binary data and is commonly used as an inferential tool in population health research, but can be also used as a binary classification model. The KNN model is chosen based on its ability to detect linear and nonlinear boundaries between groups. The K is a value that represents the number of nearest neighbors which is the core deciding factor in this classifier. It relies on finding the best value of k so that the k closest observations are used to predict the value of a given observation. Thus, when $k = 1$ then the new data object is simply assigned to the class of its nearest neighbor. The "nearness" of observations is widely measured using Euclidean distance between observations even though there are various numerical measures (Ali, Neagu, & Trundle, 2019; Larose, 2015). The main concept behind KNN depends on calculating the distances between the tested, and the trained data samples to identify its nearest neighbors. The RF model is commonly used in machine learning situations because it is highly flexible and provides better predictive performance. The RF model repeatedly samples the variables in the training data set several times, each time using a random set of predictor variables to produce decision trees. After many of these trees are formed, the forest is examined to see which variable consistently produce a better prediction. These groups of relatively uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. This is because the trees protect each other from their errors (as long as they do not all constantly err in the same direction).

ML was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks. It allows computers to learn from complex data sources, to potentially find previously unseen insights without being explicitly programmed where to look (Elisa, 2018). It can also be used to automate tasks by building analytical models using algorithms that iteratively learn from data. In demographic parlance, ML appears to address some of the major challenges in demographic research by helping to draw insights using available datasets collected for different purposes at different points in time, which in most cases may be challenging to incorporate in the traditional techniques. It may be also used to predict future occurrences of the principal components of population change (fertility, mortality, and migration) and associated factors. As such, ML techniques can be both used to predict previously identified proximate correlates and new "significant" demographic variables, and also shed light on how important previously used variables are in terms of prediction.

In this regard, we randomly sampled and trained 80% of the total sample, which was eventually used for 10-fold cross-validation to tune the model parameters. The remaining 20% random sample was used as test data to predict the measures of model performance. Because the outcome is unbalanced (there is a low fraction of under-five mortality in the data), the data were down-sampled so the proportions of data in the training set are equivalent to the cases who were alive after 5 years, and those who had died before 5 years. Model accuracy metrics such as sensitivity, specificity, positive predictive value, and negative predictive values were calculated to show how well the models perform in terms of predicting the dead and alive cases. Sensitivity ("positivity in health") refers to the proportion of subjects who have dead cases (reference standard positive) and give positive test results. Specificity ("negativity in health") is the proportion of subjects that are alive (reference standard negative) and give negative test results. Positive predictive value is the proportion of positive results that are true positives (i.e., truly dead) whereas negative predictive value is the proportion of negative results that are true negatives (i.e., truly alive). Predictive values vary depending on the prevalence of the target condition in the population being studied, even if the sensitivity and specificity remain the same (Price & Christenson, 2007).

Metrics such as the area under curve (AUC) and receiver operating characteristic (ROC) curve were also used to evaluate model performance in distinguishing between the dead and alive cases. The ROC curves compare sensitivity versus specificity across a range of values to determine the ability to predict a dichotomous outcome. The AUC is a measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve (Florkowski, 2008). Thus, the higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes (Florkowski, 2008).

The results of all the models were weighted using person weights provided by the data. For the traditional logistic regression model, we infer the importance and significance of predictors using odds ratios and confidence intervals derived from the model estimation, while for the ML models, the Mean Decrease in Gini was calculated for each variable, which is a measure of variable importance for these models. The top 10 categories of variables based on their Mean Decrease in Gini were automatically generated and then presented in diagrams for each ML model.

Results

Descriptive results of the background characteristics

Table 1 shows the results of under-five mortality by the sample characteristics. Of the 10, 641 under-five children in the sample, there appears to be a significant difference in mortality prevalence between both sexes with female children experiencing higher (6.7%) than males (4.2%). There were also considerable differences by birth intervals with under-five mortality being more prevalent among children with less than 2 years of birth intervals (9.3%) than children with 2–4 and over 4 years of birth intervals (4.45% and 4.53%, respectively). Under-five mortality was also significantly prevalent among children using unimproved water sources (5.8%) than those who used improved water sources (2.9%). Significant differences were also observed regarding antenatal visits and postnatal care, with under-five mortality being considerably prevalent among children whose mothers did not receive antenatal (5.6%) and postnatal care (4.2%). Children who were breastfed within more than 1 h of birth had a significantly higher prevalence of death (9.8%) than those breastfed within 1 h of birth (4.5%) while there was also evidence of a significant difference in under-five mortality regarding the number of people in the household. The rest of the characteristics did not show any significant difference in mortality prevalence among their categories.

Spatial distribution of under-five mortality

Figure 1 shows the spatial distribution of crude under-five mortality rates by regions in Ethiopia. The under-five mortality rate in the map is presented as the number of under-five deaths per 1000 live births. The Afar region recorded the highest under-five mortality rate of 125 per 1000 live births, followed by Benshangul–Gumuz, and Somali, which recorded 98 and 94 per 1000 live births, respectively. The lowest under-five mortality rate is recorded in Addis Ababa, with a rate of 39 per 1000 live births.

Predicting under-five mortality

Below, we report the results of the three machine learning models (logistic regression, random Forests, and the K-nearest neighbor models) (Table 2). The under-five

Table 1 Descriptive statistics of child mortality outcome by study characteristics, EDHS 2016 (*N* = 10,641)

Characteristics	Child alive before age 5 Percent/mean	Child dead before age 5 Percent/mean	Chi-square test of equality
Child dead/alive	94.9	5.1	
Child sex			***p* = .0001**
Male	95.8	4.2	
Female	93.3	6.7	
Birth order			*p* = 0.71
1st or 2nd	94.7	5.3	
3rd or higher	94.4	5.6	
Birth interval			***p* = .0001**
< 2 years	90.7	9.3	
2–4 years	95.5	4.45	
> 4 years	95.5	4.53	
Mothers age at first birth			*p* = 0.47
< 20 years	94.3	5.7	
> 20 years	94.9	5.1	
Age of the mother			*p* = 0.94
15–19	94.9	5.0	
20–34	94.4	5.6	
35–49	94.6	5.4	
Residence			*p* = 0.28
Rural	94.4	5.6	
Urban	95.7	4.3	
Education			*p* = 0.34
No education	94.1	5.9	
Primary	95.1	4.8	
Secondary and Higher	95.5	4.5	
Wealth index			*p* = 0.63
Low	94.7	5.3	
Middle	94.7	5.3	
High	94.1	5.9	
Water source			*p* = 0.51
Unimproved	94.3	5.73	
Improved	94.8	5.23	
Time to water source	167.4	164.6	*p* = 0.89*
Toilet facility			***p* = 0.005**
Unimproved	94.2	5.8	
Improved	97.1	2.9	
Place and mode of delivery services			*p* = 0.07
Fac with CS delivery	96.1	3.9	
Fac without CS delivery	94.1	5.9	
Home	94.2	5.8	
Contraceptive use			*p* = 0.23
Yes	95.1	4.9	

Table 1 Descriptive statistics of child mortality outcome by study characteristics, EDHS 2016 (*N* = 10,641) *(Continued)*

Characteristics	Child alive before age 5 Percent/mean	Child dead before age 5 Percent/mean	Chi-square test of equality
No	94.2	5.8	
Child wanted			*p* = 0.74
Then	94.6	5.4	
Later	94.5	5.5	
Not at all	93.6	6.4	
Antenatal visits			***p* = 0.002**
No visit	94.4	5.6	
1–4 visits	96.7	3.3	
5+ visits	97.6	2.4	
Postnatal care visits			***p* = 0.009**
No	95.8	4.2	
Yes	98.3	1.7	
Region			*p* = 0.43
Oromia	94.2	5.8	
Addis Ababa	95.9	4.1	
Afar	91.5	8.5	
Amhara	94.9	5.1	
Ben-Gumuz	94.2	5.8	
Dire Dawa	93.7	6.3	
Gambella	93.3	6.7	
Harari	94.5	5.5	
SNNP	93.3	6.7	
Somali	96.7	3.3	
Tigray	93.8	6.2	
Mother's BMI			*p* = 0.41
Underweight	93.7	6.3	
Normal	94.6	5.4	
Overweight	95.4	4.6	
Breastfed			***p* = .0001**
Within an hour of birth	95.5	4.5	
Greater than an hour of birth	90.2	9.8	
Household size	6.1	5.4	***p* = .0001***

NB all estimates include sample design and person weights, per DHS instructions. **t* test was used instead of chi-square. Significant variables are in bold

mortality prediction accuracy was found to be low for all models, between 46.3 and 67.2% accuracy on the test dataset, with the RF model having the highest overall accuracy. The RF model had high sensitivity, meaning that it was more accurate in identifying dead cases, but had low specificity, meaning that it was poor in identifying the alive cases. However, the model correctly identified 70% of the real dead cases (28/(28+12)) and 67% of real alive cases (698/(698+343)), which means that the RF model is relatively better at predicting both real positive (dead) and negative (alive) cases. The

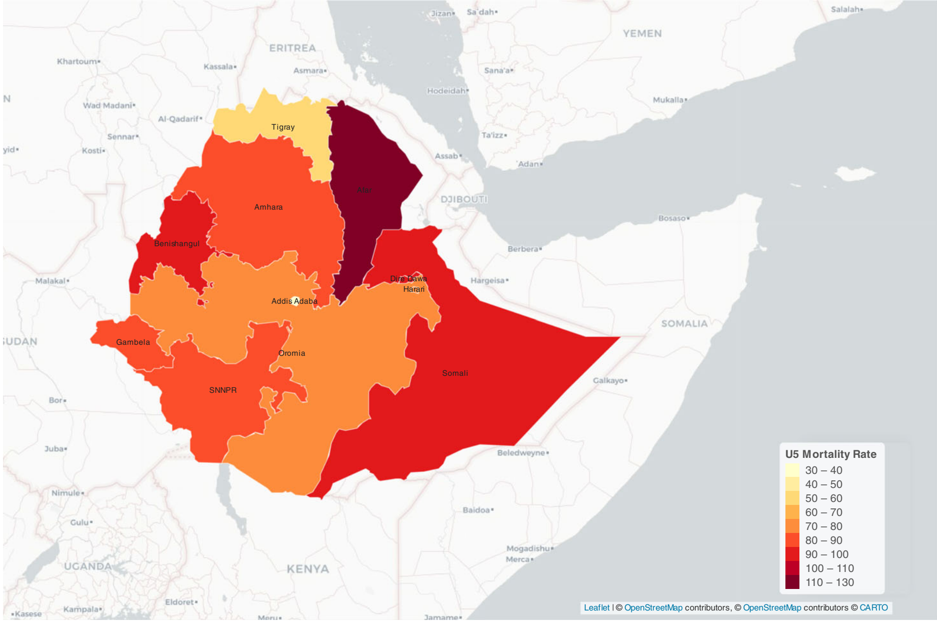

Fig. 1 Spatial distribution of crude under-five mortality rates by regions in Ethiopia.Source: Created by the authors from EDHS estimates

logistic and KNN models both show lower overall accuracy (59.9 and 46.3%, respectively), and lower sensitivity, specificity, and positive as well as negative predictive values.

A visualization of the receiver operating characteristics (ROC) curve is shown in Fig. 2. Among the three machine learning models employed in this study, the curve of the RF model shows the highest AUC value, indicating it is the best at classifying dead and alive cases, among the models.

Figures 3, 4, and 5 show the variable importance measures, measured by the scaled mean decrease in the Gini coefficient for each variable, as calculated during the k-fold cross-validation process. This is an effective measure of how important a variable is for predicting under-five mortality across all the cross-validation estimates. The three

Table 2 Model accuracy metrics for all models as evaluated on the test data

Confusion matrix		**Random Forest**		**Logistic regression**		**KNN model**	
		Predicted		**Predicted**		**Predicted**	
		Alive	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**
Observed	**Alive**	698	343	632	418	475	566
	Dead	12	28	16	24	14	26
			%		%		%
Accuracy			67.2		59.9		46.3
Sensitivity			98.3		97.5		97.1
Specificity			7.5		5.34		4.4
Positive predictive value			70.0		59.9		45.6
Negative predictive value			67.0		60.0		65.0
AUC			72.0		66.1		55.5

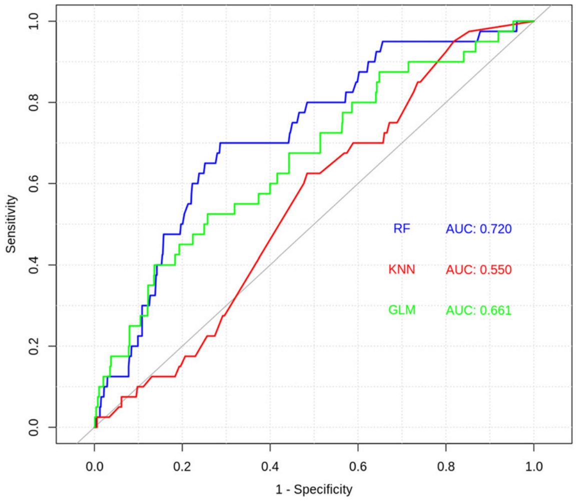

Fig. 2 ROC Curves for the three models

figures show very similar results, with household size (nhh) and breastfeeding behavior (bfeed) being among the top 3 variables in all three models. Other important factors that appeared in the top five variables are the time to water source (time_water), number of births (births5_ys), birth interval (b_interval), and child sex (male).

Unlike the ML model results presented above, the traditional logistic regression model is the only one that allows direct interpretation of the model coefficients (Table 3). Table 3 shows the estimated odds ratios and confidence intervals for the model parameters. Factors associated with under-five mortality were sex, birth order, birth interval, water source, place of delivery, antenatal visit, postnatal care, breastfeeding, and household size. Increased risks of under-five mortality were found among males, higher birth order children, and children born in a facility without C-section services. On the contrary, reduced risks were found among children with longer birth intervals, improved water sources, children who received antenatal and postnatal care as well as those from larger households.

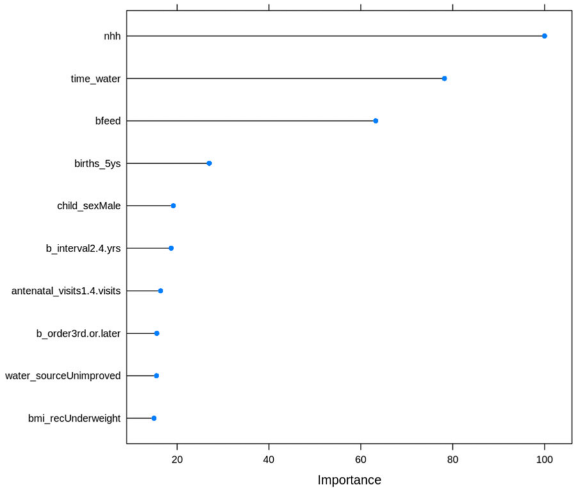

Fig. 3 Variable importance measures for the random Forest model

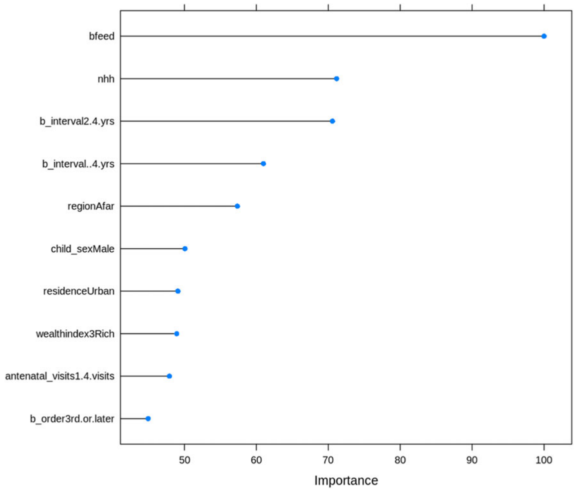

Fig. 4 Variable importance measures for logistic regression model

Discussion

This study briefly described spatial variations in under-five mortality and predicted under-five mortality risks in Ethiopia using machine learning techniques. The spatial map provides evidence of considerable regional disparities in under-five mortality rates in Ethiopia similar to what has been observed in Ghana (Aheto, 2019). Tigray and some regions in the central part of the country show the lowest under-five mortality rates whereas regions in the eastern and western parts of the country have the highest under-five mortality rates. Providing evidence on the spatial variations of under-five mortality in the country may provide the need to better understand the underlying risk factors. Regarding the predictive analysis, the prediction accuracies and AUC statistics are found to be highest for the RF model. The RF model shows a higher predictive power compared to the other ML models included in this study. In this regard, the RF model shows that household size, time to the water source, breastfeeding behavior, births in the preceding 5 years, sex of a child, birth intervals, birth order, antenatal

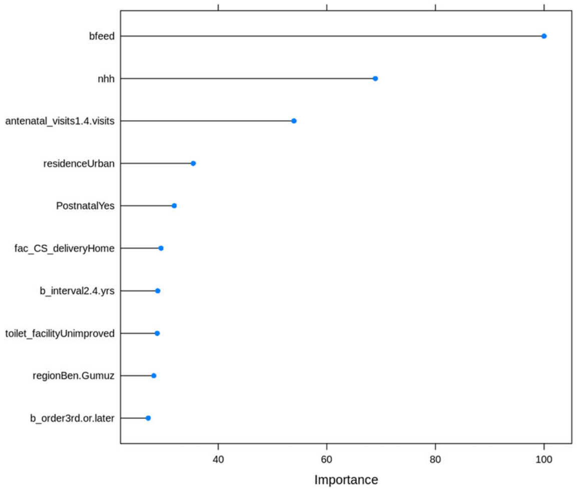

Fig. 5 Variable importance measures for K-nearest neighbor model

Table 3 Logistic regression analysis of under-five mortality in Ethiopia

Variables	Odds ratio	Lower 95 % CI	Upper 95% CI	*p* value
(Intercept)	0.033	0.006	0.193	**0.0001**
Mothers age first birth (Ref: < 20)				
> 20	0.600	0.353	1.018	0.059
Sex (Ref: female)				
Male	2.018	1.398	2.913	**0.0001**
Birth order (Ref: 1st/2nd)				
3rd or higher	2.129	1.131	4.008	**0.020**
Birth interval (Ref: < 2)				
2–4 years	0.527	0.309	0.898	**0.019**
> 4 years	0.385	0.190	0.779	**0.008**
Time to water source	1.000	0.999	1.000	0.244
Water source (Ref: unimproved)				
Improved	0.585	0.348	0.985	**0.044**
Toilet facility (Ref: improved)				
Unimproved	1.713	0.744	3.943	0.206
Births in last 5 years	1.163	0.744	1.816	0.508
Residence (Ref: rural)				
Urban	0.527	0.181	1.541	0.243
Mother's education (Ref: no education)				
Primary	0.928	0.513	1.680	0.805
Secondary/higher	1.856	0.480	7.178	0.370
Wealth index (Ref: low)				
Middle	1.342	0.698	2.581	0.378
High	1.694	0.937	3.064	0.082
Contraceptive use (Ref: using)				
Not using	1.174	0.735	1.876	0.502
Region				
Addis Ababa	1.124	0.485	2.605	0.786
Afar	0.573	0.228	1.435	0.235
Amhara	0.885	0.354	2.211	0.794
Ben-Gumuz	1.494	0.587	3.803	0.400
Dire Dawa	1.021	0.408	2.554	0.965
Gambella	0.623	0.243	1.597	0.325
Harari	1.175	0.495	2.790	0.715
SNNP	1.221	0.376	3.960	0.740
Somali	1.504	0.287	7.881	0.629
Tigray	1.733	0.519	5.787	0.372
Mother's BMI (Ref: normal)				
Overweight	0.527	0.170	1.640	0.269
Underweight	1.402	0.868	2.264	0.168
Place of delivery (Ref: fac with CS delivery)				
Facility without CS delivery	2.850	1.182	6.869	**0.020**
Home	1.185	0.617	2.275	0.610

Table 3 Logistic regression analysis of under-five mortality in Ethiopia *(Continued)*

Variables	Odds ratio	Lower 95 % CI	Upper 95% CI	*p* value
Antenatal visits (Ref: no visit)				
1–4 visits	0.616	0.381	0.995	**0.048**
5+ visits	0.437	0.208	0.917	**0.029**
Postnatal care (Ref: no)				
Yes	0.264	0.080	0.872	**0.029**
Child wanted (Ref: wanted then)				
Wanted later	0.768	0.369	1.599	0.482
Not at all	1.407	0.749	2.642	0.289
Breastfeeding (Ref: > an hour of birth)				
Within 1 h of birth	0.242	0.147	0.398	**0.0001**
vHousehold size	0.498	0.345	0.719	**0.0001**

NB significant variables are in bold

visits, type of water source, and mother's BMI are the top 10 important predictors of under-five mortality in Ethiopia. The important role played by some of these factors in under-five mortality levels is widely documented in the literature (Abir, Agho, Page, Milton, & Dibley, 2015; Dendup, Zhao, & Dema, 2018; Howell, Holla, & Waidmann, 2016; Yaya et al., 2018).

In comparison, the findings of the best performing ML model appear to be virtually consistent with the traditional logistic regression analysis which also shows that a child's sex, birth interval, birth order, water source, place of delivery, antenatal visits, postnatal care, household size, and breastfeeding behavior play a significant role in under-five mortality levels in Ethiopia. Only the number of births in the preceding 5 years and the mother's BMI appear to play an important role in the ML models but play an insignificant role in the traditional logistic regression analysis. This is an indication that ML models may produce some "new variables" or previously unseen insights by the traditional regression models which may play a crucial role in policy decision making. From the traditional logistic regression findings, male children have shown a significantly higher risk of dying before age 5 compared with female children. This is consistent with the finding of a cross-sectional study conducted in Bangladesh (Abir et al., 2015). It has been shown that male children have an increased risk of dying in the first month of life because of high vulnerability to infectious disease. This may be because female neonates are more likely to develop early fetal lung maturity in the first week of life, which may result in a lower incidence of respiratory diseases in female compared with male neonates (Khoury, Marks, McCarthy, & Zaro, 1985). Also, higher birth order of children appears to be associated with a significantly higher risk of under-five mortality. Analogously, the unfavorable effect of higher birth order on childhood survival chances has been well documented in Africa (Howell et al., 2016) as well as some parts of Asia (Dendup et al., 2018; Hong & Hor, 2013) and may provide a better understanding of the spatial variations in the country.

Furthermore, the risk of under-five mortality has increased significantly among children with less than 2 years of birth interval than children with more than 2 years of birth interval. Affirmatively, there is much evidence that longer birth intervals improve the survival chance of succeeding children (Kozuki & Walker, 2013; Yaya et al., 2018). A short preceding birth interval can be said to influence under-five mortality through three main mechanisms: first, closely spaced births may cause depletion of the mother. The second mechanism is through competition for scarce household resources among children, while the third is the transmission of infectious diseases between the closely spaced children (Majumder, May, & Pant, 1997). While the first mechanism is biological, the last two are said to be behavioral effects of a short preceding birth interval (Koenig, Phillips, Campbell, & Dsouza, 1990).

Additionally, this study finds that the use of unimproved drinking water is associated with an increased risk of under-five mortality. Lack of access to clean water has been considered as one of the important factors that contribute to more than 80% of child deaths in the world (UNICEF, 2018). There is also considerable evidence from studies in developing countries that show that household sanitation and a clean water supply promote child health and survival (Ezeh, Agho, Dibley, Hall, & Page, 2014; Mugo, Agho, Zwi, Damundu, & Dibley, 2018). In Ethiopia, the proportion of the population using improved drinking water sources is only 57%, and those who use improved sanitation are less than 5% (World Health Organization, 2017). This may have serious implications for variations in under-five mortality in the country. This study further provides evidence that children whose mothers do not use any contraceptives have a significantly higher risk of under-five mortality than their counterparts whose mothers use modern contraceptives.

This study also finds that delivery in health facilities without CS services and at home is associated with a higher under-five mortality risk. This may be mainly related to dealing with delivery complications that may raise under-five mortality risks. Health facilities with CS services are very scarce in Ethiopia, and where they are available, transportation challenges encourage women to deliver at home even when facility-based delivery is available at a minimal cost (Shiferaw, Spigt, Godefrooij, Melkamu, & Tekie, 2013). Moreover, the study finds a positive effect of antenatal and postnatal care checkups on under-five survival chances. This is consistent with the significant association observed between antenatal and postnatal care and lower under-five mortality risk in the literature (Bitew & Nyarko, 2019; Machio, 2018). The implication is that children whose mothers do not receive antenatal and postnatal care services may experience several proximate under-five mortality risk factors, such as congenital and infectious diseases, than their counterparts. This study has also shown a considerable positive effect of early timing of breastfeeding on childhood survival chances. Breastfeeding has long been shown as an important protective factor against under-five mortality, particularly among developing countries (Azuine, Murray, Alsafi, & Singh, 2015; Nyarko, Tanle, & Kumi-Kyereme, 2014) and may play a key part in childhood survival interventions in Ethiopia. Quite surprisingly, larger household size appears to be associated with reduced under-five mortality risk in this study, contrary to what is documented in the literature (Dendup et al., 2018). However, this may well be underscored by some household-level contextual factors in the country such as availability of considerable social support from parents and siblings.

This study is not without limitations. The survey comprised only surviving women, and since neonatal and maternal mortalities may occur concurrently, this may have led to an underestimation of the under-five mortality rates. Ultimately, unlike the traditional regression models, the ML results appear to be mostly uninterpretable because they have no regression coefficients and for that matter no direction of effect. In effect, ML models generally predict or classify specific variables based on the level of importance of their role in determining the under-five mortality levels in the current study. In this case, extant empirical literature from studies using the traditional methodologies may be used to determine the direction of these important variables. There are also possible biases in the memorization or non-disclosure of deaths by mothers which may underestimate the number of deaths. Nevertheless, machine learning techniques are considered to be very useful in predicting population health and other phenomena and lead to better policy decisions (Ashrafian & Darzi, 2018; Holzinger, 2017).

Conclusions

The findings show that considerable regional disparities in under-five mortality rates persist in Ethiopia, with the highest rates being found in the Afar, Benishangul—Gumuz, and Somali regions. Also, the RF model provides a moderately better predictive power than the logistic regression and KNN ML models in predicting under-five mortality determinants in Ethiopia. Even though the RF model and the traditional logistic regression model have shown similar factors, the RF model appears to reveal some important factors that may not be identified by the traditional logistic regression model. This model may, therefore, proffer better policy directions regarding under-five childhood survival. Thus, household size, time to the water source, breastfeeding behavior, number of births in the past 5 years, sex of a child, birth intervals, antenatal visits, birth order, type of water source, and mother's BMI may play an important role in under-five survival chances in Ethiopia. This study highlights the use of machine learning algorithms to predict and better understand very important under-five mortality risk factors to improve crucial policy directions. As a corollary, ML methods may also apply to other areas of demographic research including fertility and migration studies. Our findings reinforce the need to focus on the most important predicted factors including breastfeeding, birth interval control, and antenatal care among others in developing policies aimed at enhancing childhood survival chances. Also, based on the findings, expanding access to improved drinking water will help to substantially reduce future under-five mortality levels in Ethiopia.

Abbreviations

AUC: Area under curve; BMI: Body mass index; CS: Caesarean section; EDHS: Ethiopian Demographic and Health Survey; KNN: K-Nearest Neighbors; LMIC: Low- and middle-income countries; MDG: Millennium Development Goal; ML: Machine learning; RF: Random forest; ROC: Receiver operating characteristic; SNNPR: Southern Nations Nationalities and People Region

Authors' contributions

FHB conceived and designed the study. FHB and CSS performed the analysis with technical support from SHN. FHB wrote the initial draft of the manuscript with technical support from SHN, LP, and CSS. All authors critically reviewed the manuscript for important intellectual content and then approved the final version of the manuscript for publication.

Funding

No funding was received for this study

Availability of data and materials

The datasets analyzed in this study are freely available at the DHS Program repository

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Demography, College for Health, Community & Policy, University of Texas at San Antonio, 501 W. Cesar Chavez Blvd, San Antonio, TX 78207, USA. [2]Institute for Demographic and Socioeconomic Research, The University of Texas at San Antonio, 501 W. Cesar Chavez Blvd, San Antonio, TX 78207, USA.

Received: 30 April 2020 Accepted: 16 October 2020
Published online: 04 November 2020

References

Abir, T., Agho, K. E., Page, A. N., Milton, A. H., & Dibley, M. J. (2015). Risk factors for under-five mortality: evidence from Bangladesh Demographic and Health Survey, 2004–2011. *BMJ Open*, *5*(8), e006722.

Aheto, J. M. K. (2019). Predictive model and determinants of under-five child mortality: evidence from the 2014 Ghana Demographic and Health Survey. *BMC Public Health*, *19*, 64.

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*, *1*(12), 1559.

Ashrafian, H., & Darzi, A. (2018). Transforming health policy through machine learning. *PLoS Medicine*, *15*(11), e1002692.

Ayele, D. G., & Zewotir, T. T. (2016). Childhood mortality spatial distribution in Ethiopia. *Journal of Applied Statistics*, *43*(15), 2813–2828.

Ayele, D. G., Zewotir, T. T., & Mwambi, H. (2017). Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. *Journal of Health, Population, & Nutrition*, *36*(1), 25.

Azuine, R. E., Murray, J., Alsafi, N., & Singh, G. K. (2015). Exclusive breastfeeding and under-five mortality, 2006-2014: A cross-national analysis of 57 low- and-middle income countries. *International Journal of MCH AIDS*, *4*(1), 13–21.

Bereka, S. G., Habtewold, F. G., & Nebi, T. D. (2017). Under-five mortality of children and its determinants in Ethiopian Somali Regional State, Eastern Ethiopia. *Health Science Journal*, *11*, 3.

Bitew, F., & Nyarko, S. H. (2019). Modern contraceptive use and intention to use: implication for under-five mortality in Ethiopia. *Heliyon*, *5*, e02295.

Central Statistical Agency (CSA) [Ethiopia], & ICF International (2016). *Ethiopia Demographic and Health Survey 2016*. Addis Ababa, Ethiopia, Calverton, MD, USA: Central Statistical Agency, ICF International.

Dendup, T., Zhao, Y., & Dema, D. (2018). Factors associated with under-five mortality in Bhutan: an analysis of the Bhutan National Health Survey 2012. *BMC Public Health*, *18*, 1375.

Elisa, N. (2018). Could Machine Learning be used to address Africa's Challenges? *International Journal of Computer Applications*, *180*(18), 0975–8887.

Ezeh, O. K., Agho, K. E., Dibley, M. J., Hall, J., & Page, A. N. (2014). The impact of water and sanitation on childhood mortality in Nigeria: evidence from demographic and health surveys, 2003–2013. *International Journal of Environmental Research and Public Health*, *11*(9), 9256–9272.

Federal Ministry of Health (2005). *National Strategy for Child Survival in Ethiopia*. Addis Ababa: Federal Ministry of Health.

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves, and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, *29*(Suppl 1), S83.

Holzinger, A. (2017). Introduction to machine learning and knowledge extraction (MAKE). *Machine Learning and Knowledge Extraction*, *1*(1), 1–20.

Hong, R., & Hor, D. (2013). *Factors associated with the decline of under-five mortality in Cambodia, 2000-2010: Further analysis of the Cambodia Demographic and Health Surveys*. Calverton: ICF International.s.

Howell, E. M., Holla, N., & Waidmann, T. (2016). Being the younger child in a large African family: a study of birth order as a risk factor for poor health using the demographic and health surveys for 18 countries. *BMC Nutrition*, *2*, 61.

Khoury, M. J., Marks, J. S., McCarthy, B. J., & Zaro, S. M. (1985). Factors affecting the sex differential in neonatal mortality: the role of respiratory distress syndrome. *American Journal of Obstetrics and Gynecology*, *151*(6), 777–782.

Koenig, M. A., Phillips, J. F., Campbell, O. M., & Dsouza, S. (1990). Birth intervals and childhood mortality in rural Bangladesh. *Demography*, *27*(2), 251–265.

Kozuki, N., & Walker, N. (2013). Exploring the association between short/long preceding birth intervals and child mortality: using reference birth interval children of the same mother as comparison. *BMC Public Health*, *13*, S6.

Kuhn, M. (2020). *Caret: Classification and Regression Training. R package version*, *6*, *0–85* https://CRAN.R-project.org/package=caret.

Larose, D. T. (2015). *Data mining and predictive analytics*. New York: Wiley.

Machio, P. M. (2018). Determinants of neonatal and under-five mortality in Kenya: Do antenatal and skilled delivery care services matter? *Journal of African Development*, *20*(1), 59–67.

Majumder, A. K., May, M., & Pant, P. D. (1997). Infant and child mortality determinants in Bangladesh: Are they changing? *Journal of Biosocial Science*, *29*(4), 385–399.

Mugo, N. S., Agho, K. E., Zwi, A. B., Damundu, E. Y., & Dibley, M. J. (2018). Determinants of neonatal, infant, and under-five mortality in a war-affected country: analysis of the 2010 Household Health Survey in South Sudan. *BMJ Global Health*, *3*(1), e000510.

Nyarko, S. H., Tanle, A., & Kumi-Kyereme, A. (2014). Determinants of childhood mortality in Ghana. *International Journal of Social Science Research*, *3*, 61–77.

Price, C. P., & Christenson, R. H. (2007). *Evidence-based laboratory medicine: principles, practice, and outcomes*, (2nd ed.,). Washington DC: American Association for Clinical Chemistry Press.

Shiferaw, S., Spigt, M., Godefrooij, M., Melkamu, Y., & Tekie, M. (2013). Why do women prefer home births in Ethiopia? *BMC Pregnancy and Childbirth*, *13*, 5.

UNICEF. (2017). *The State of the World's Children*. https://www.unicef.org/sowc/. Accessed March 15, 2019.

UNICEF (2018). *Every Child Alive. The urgent need to end newborn deaths*. Genèva, Switzerland: UNICEF.

UNICEF, WHO, World Bank Group & United Nations (2018). *Levels and trends in child mortality report 2018*. New York: UNICEF.

World Health Organization (2017). *World health statistics 2017: Monitoring health for the SDGs, and Sustainable Development Goals*. Geneva: WHO.

Yaya, S., Bishwajit, G., Okonofua, F., & Uthman, O. A. (2018). Under five mortality patterns and associated maternal risk factors in sub-Saharan Africa: A multi-country analysis. *PLoS ONE*, *13*(10), e0205977.

You, D., Hug, L., Ejdemyr, S., Idele, P., et al. (2015). Global, regional, and national levels and trends in under-five mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *Lancet*, *386*(10010), 2275–2286.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.