Predictive models and under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey

Fikrewold Bitew (✉ fikre.wold@gmail.com)

Ethiopian Public Health Association

Samuel H. Nyarko

University of Texas at San Antonio

Lloyd Potter

University of Texas at San Antonio

Corey S. Sparks

University of Texas at San Antonio

Research article

Keywords: Predictive model, determinants, under-five mortality, Ethiopia

Posted Date: April 8th, 2020

DOI: <https://doi.org/10.21203/rs.2.13113/v3>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: There is a dearth of literature on predictive models estimating under-five mortality risk in Ethiopia. In this study, we develop a spatial map and predictive models to predict the sociodemographic determinants of under-five mortality in Ethiopia.

Methods: The study data were drawn from the 2016 Ethiopian Demographic and Health Survey. We used three predictive models to predict under-five mortality within this sample. The three techniques are random forests, logistic regression, and k-nearest neighbors. For each model, measures of model accuracy and Receiver Operating Characteristic curves are used to evaluate the predictive power of each model.

Results: There are considerable regional variations in under-five mortality rates in Ethiopia. The under-five mortality prediction ability was found to be moderate to low for the models considered, with the random forest model showing the best performance. Maternal age at birth, sex of a child, previous birth interval, water source, health facility delivery services, antenatal and post-natal care checkups, breastfeeding behavior and household size have been found to be significantly associated with under-five mortality in Ethiopia.

Conclusions: The random forest machine learning algorithm produces a higher predictive power for under-five mortality risk factors for the study sample. There is a need to improve the quality and access to health care services to enhance childhood survival chances in the country.

Background

Globally, an estimated 5.4 million children under the age of 5 are said to have died in 2017 alone [1]. Meanwhile, the global under-five mortality rate is said to have declined by 58 percent, from 93 deaths per 1,000 live births in 1990 to 39 in 2017 [1]. 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) [1]. 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 [2].

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 [1]. Ethiopia has been found to have the fifth-highest number of newborn deaths in the world, following India, Pakistan, Nigeria, and the Democratic Republic of Congo [3]. 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 an absolute number of under-five deaths [4]. In Ethiopia, the under-5 mortality rate has declined by two thirds from the 1990 figure of 204/1,000 live births to 58/1,000 live births in 2016, and thus, achieving the target for Millennium Development Goal 4 (MDG 4) [5]. Despite this

achievement, the under-five mortality rate in Ethiopia is still 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 [6-8], using traditional regression models. In this study, we ascertain the 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 5 mortality in this sample. The main aim is to determine the spatial distribution of under-five mortality and the best predictive model that shows the potential of machine learning techniques in estimating the most important sociodemographic factors affecting under-five mortality distribution. 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-5 mortality in the country.

Methods

Data source

This study draws 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 with the aim of improving maternal and child health in Ethiopia [9]. 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) [9]. The unit of analysis comprised a total of 10,641 children under age 5 of mothers selected from 645 clusters across the country. This was based on the children's data obtained from a retrospective information from mothers about their children that died before age five within the five years preceding the survey (2011 to 2016).

Study variables

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 services 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 included 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 hour of birth. The household factors used were the

source of drinking water (improved/unimproved), time to 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 services factors included antenatal visits (0, 1-4, 5+ visits), place and mode of delivery services (Facility with Cesarean Section (CS) services, facility without CS, home), and postnatal visits within two months after delivery (Yes/No). The selection of these predictor variables was based on information from existing literature on the subject [6-8].

Analytic strategy

The R programming language (version 3.6.0) and the caret package [10] 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 the widely accepted machine learning algorithms – logistic regression, a random forest model (RFM), K-nearest neighbors (KNN), – to predict under-five mortality in Ethiopia. These three models were selected for the following reasons. First, logistic regression is typically used to analyze binary data and commonly used as an inferential tool in population health research, but it also can be used as a binary classification model. Second, the KNN model is chosen based on its ability to detect linear and nonlinear boundaries between groups. The KNN method relies on finding the best value of k so that the k closest observations are used to predict the value of a given observation. “Closeness” of observations is usually measured using a distance metric such as the Euclidean distance between observations. Third, from a predictive modeling perspective, the random forest model is commonly used in machine learning situations because they are highly flexible and provide better predictive performance. Random forests repeatedly sample 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. In this regard, machine learning techniques draw on a learning process that extracts useful information from the data generation process of previous observations [11]. It is touted as a prominent application of artificial intelligence technology for ensuring good health and social care for an entire population through preventive strategies, and protection from diseases [12].

We randomly selected and trained an 80% sample of the original data, 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 children in the data who die), the data were down-sampled so the proportions of data in the training set are equivalent for the cases who were alive after 5 years, and those who had died before 5 years. The performance of these algorithms was evaluated using various metrics including the Area

Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve, which are useful in deciding which model provides the best discriminatory power between the dead and alive cases. The positive and negative predictive accuracy of each model was also calculated to show how well the model performs in terms of predicting the dead and alive cases, respectively. The results from all of the above models were weighted using person weights provided by the DHS. For the logistic regression model, we infer the importance and significance of predictors using traditional t-statistics and odds ratios derived from the model estimation, while for the random forest and KNN methods, these are not available. For these models, the Mean Decrease in Gini was calculated, which is a measure of variable importance for these models.

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 2-4 and over 4 years of birth intervals (4.455 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 one hour of birth had a significantly higher prevalence of death (9.8%) than those breastfed within one hour 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 a 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 results from the three machine learning models (logistic regression, Random Forests, and the k-nearest neighbor models) to predict the under-five mortality outcome (Table 2). The under-five mortality prediction accuracy was found to be low for all models, at between 46.3 to 67.2% accuracy on the test set, with the random forest model having the highest overall accuracy. The random forest model

had high sensitivity, meaning that it was accurate at distinguishing the alive cases from the dead cases, but low specificity, meaning that it was not good in discerning the dead cases. More metrics show that the model is relatively good at predicting both positive (alive) and negative (dead) cases. The model was able to correctly identify 70% of dead cases ($28/(28+12)$), which suggests it is relatively good at predicting the dead cases. The logistic and KNN models both show lower overall accuracy (59.9 and 46.3%, respectively), and lower sensitivity, specificity, and positive and negative predictive values. The results for the receiver operating characteristics (ROC) curve are shown in Figure 1. Among the three machine learning models employed in this study, the curve for the random forest model shows the highest AUC value, indicating it is best at separating the two classes, among the models considered.

The logistic regression model is the only one of the three that allows for direct interpretation of the model coefficients. Table 3 shows the estimated odds ratios and confidence intervals for the model parameters. Factors associated with increased risk of under 5 mortality were male sex, higher birth order and being born in a facility without C-section services. Protective factors were longer birth intervals, improved water source, having received antenatal and postnatal care as well as larger household size.

Figures 3 through 5 show the variable importance measures, as measured by the scaled mean decrease in the Gini coefficient for each variable, as measured 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 figures show very similar results, with household size (nhh) and breastfeeding behavior (bfeed) among the top three variables in all three models. Other important factors that fall out in the top 5 variables are the time to water source (time_water), number of births (births5_ys), birth interval (b_interval) and child sex (male).

Discussion

The study develops a series of predictive models and a regional map for under-five mortality 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 found in Ghana [13]. 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 underlying risk factors may help to better understand the spatial variations of under-five mortality in the country. Regarding the predictive model, the prediction accuracies and AUC statistics are found to be highest for the random forest model. It shows the higher predictive power of the random forest model compared to the other models considered here. In this regard, the Random Forest Model shows that household size, time to the water source, breastfeeding behavior, number of recent births, child sex and length of birth intervals are the strongest predictors of child mortality.

The logistic regression models show that a child's sex, preceding birth interval, water source, place and mode of delivery, antenatal care checkup, postnatal care, household size, and breastfeeding behavior are significantly associated with under-five mortality in Ethiopia. In this study, children of teenage mothers

show a higher risk of under-five mortality than children of older mothers. Consequently, male children have shown a significantly higher risk of dying before age five compared with female children. This is consistent with the finding of a cross-sectional study conducted in Bangladesh [14]. 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 is 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 [15]. In this study, 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 [16] as well as some parts of Asia [17, 18) and may probably be due to fierce competition for scarce household resources. Also, the risk of under-five mortality has increased significantly among children with less than 2 years preceding birth interval than children with more than 2 years or birth interval. Affirmatively, there is much evidence that longer birth intervals improve the survival chance of succeeding children [19, 20]. 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 sibling competition while the third is the transmission of infectious diseases between the closely spaced children [21]. While the first mechanism is biological, the last two are said to be behavioral effects of a short preceding birth interval [22].

Furthermore, this study finds that the use of an unimproved source of 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 percent of child deaths in the world [23]. 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 [24, 25]. In Ethiopia, the proportion of the population using improved drinking-water sources is only 57%, and those who use improved sanitation are less than five percent [2]. This may have serious implications for the under-five mortality levels in the country. This study further provides evidence that children whose mothers do not use any contraceptives have a significantly higher risk for 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 risk. Health facilities with CS services are very scarce in Ethiopia; even where they are available, transportation challenges encourage women to deliver at home delivery when facility-based delivery is available at a minimal cost [26]. Moreover, this study provides evidence of 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 [27, 28]. The implication is that children whose mothers do not receive antenatal and postnatal care services may experience more 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 [29, 30] and has to play a key part in childhood survival interventions. Quite surprisingly, larger household size appears to be associated with reduced under-five mortality risk in this study, contrary to what has been documented in the literature [18]. However, this may well be underscored by some household-level contextual factors in the country such as availability of considerable social support from 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. Also, using a cross-sectional survey data such as the DHS only provides a snapshot of the scenario unlike using a longitudinal approach. There are also possible biases in the memorization or non-disclosure of deaths by mothers which may underestimate the number of deaths. Nevertheless, the machine learning techniques used provided a strong case for predicting the underlying risk factors of under-five mortality in the study sample.

Conclusions

This study provides evidence of considerable regional disparities in under-five mortality rates in Ethiopia, with the highest rates observed in the Afar, Benishangul – Gumuz and Somali regions. In this study, the Random Forest Model provides a modestly higher predictive power than the logistic regression and k-nearest neighbor models in predicting under-five mortality risks in Ethiopia. Under-five mortality in Ethiopia is significantly associated with maternal age at first birth, sex of a child, previous birth interval, water source, health facility delivery services, antenatal and post-natal care checkups, household size and breastfeeding immediacy. Children of teenage mothers and mothers, male children, short birth interval children, children from unimproved water source households, children delivered at facilities without CS services as well as children whose mothers do not receive antenatal and post-natal care, who are not breastfed immediately and who live in smaller households all have increased risks for under-five mortality in Ethiopia. This study highlights the potential of machine learning methods in predicting under-five mortality risk factors and points to crucial areas for policy development. Our findings reinforce the need to improve the quality and access to health care services such as antenatal, delivery, and post-natal care as well as family planning services in the country to enhance childhood survival chances. Also, based on the findings, expanding access to improved drinking water will help to substantially reduce under-five mortality in the country in the future.

Abbreviations

AUC, Area Under Curve; BMI, Body Mass Index; CS, Caesarean Section; EDHS, Ethiopian Demographic and Health Survey; LMIC, Low and Middle-income Countries; MDG, Millennium Development Goal; ROC, Receiver Operating Characteristic; SNNPR, Southern Nations Nationalities and People Region.

Declarations

Ethics approval and consent to participate

The study used secondary data from the EDHS. Ethical approval not applicable.

Consent to publish

Not applicable

Availability of data and methods

The dataset analysed in this study are available on The DHS Program website.

Competing interests

The authors declare that they have no competing interests.

Funding

No specific funding was received for this study

Authors' contributions

FB conceived and designed the study. FB and SHN performed the analysis with technical support from LP and CSS. FB 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.

Acknowledgments

Not applicable.

References

1. UNICEF, WHO, World Bank Group and United Nations. Levels and trends in child mortality report 2018. New York: UNICEF; 2018.
2. World Health Organization. World health statistics 2017: Monitoring health for the SDGs, and Sustainable Development Goals. Geneva: WHO; 2017
3. UNICEF. The State of the World's Children. 2017. <https://www.unicef.org/sowc/>. Accessed March 15, 2019.
4. Federal Ministry of Health. National Strategy for Child Survival in Ethiopia. Addis Ababa: Federal Ministry of Health; 2005.
5. You D, Hug L, Ejdemo S, Idele P, et al. Global, regional, and national levels and trends in under-5 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. 2015; 386(10010): 2275-2286.

6. Ayele DG, Zewotir TT. Childhood mortality spatial distribution in Ethiopia. *J Applied Stat.* 2016; 43(15): 2813-2828.
7. Ayele DG, Zewotir TT, Mwambi H. Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. *J Health, Pop Nutr.* 2017; 36(1): 25.
8. Bereka SG, Habtewold FG, Nebi TD. Under-five mortality of children and its determinants in Ethiopian Somali Regional State, Eastern Ethiopia. *Health Sci J.* 2017; 11: 3.
9. Central Statistical Agency (CSA) [Ethiopia], ICF International. Ethiopia Demographic and Health Survey 2016. Addis Ababa, Ethiopia, Calverton, MD, USA: Central Statistical Agency, ICF International; 2016.
10. Kuhn M. Caret: Classification and Regression Training. R package version 6.0-85. 2020. <https://CRAN.R-project.org/package=caret>.
11. Holzinger A. Introduction to machine learning and knowledge extraction (MAKE). *Mach. Learn. Knowl. Extr.* 2017; 1(1): 1-20.
12. Ashrafian H, Darzi A. Transforming health policy through machine learning. *PLoS Med.* 2018; 15(11): e1002692
13. Aheto JMK. Predictive model and determinants of under-five child mortality: evidence from the 2014 Ghana Demographic and Health Survey. *BMC Pub Health.* 2019; 19: 64.
14. Abir T, Agho KE, Page AN, Milton AH, Dibley MJ. Risk factors for under-5 mortality: evidence from Bangladesh Demographic and Health Survey, 2004–2011. *BMJ Open.* 2015; 5(8): e006722.
15. Khoury MJ, Marks JS, McCarthy BJ, Zaro SM. Factors affecting the sex differential in neonatal mortality: the role of respiratory distress syndrome. *Am J Obstetr Gynecol.* 1985; 151(6): 777-782.
16. Howell EM, Holla N, Waidmann T. 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 Nutri.* 2016; 2:61
17. Hong R, Hor D. 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; 2013.
18. Dendup T, Zhao Y, Dema D. Factors associated with under-five mortality in Bhutan: an analysis of the Bhutan National Health Survey 2012. *BMC Pub Health.* 2018; 18:1375.
19. Yaya S, Bishwajit G, Okonofua F, Uthman OA. Under five mortality patterns and associated maternal risk factors in sub-Saharan Africa: A multi-country analysis. *PLoS ONE.* 2018; 13(10): e0205977.
20. Kozuki, N, Walker N. Exploring the association between short/long preceding birth intervals and child mortality: using reference birth interval children of the same mother as comparison. *BMC Pub Health.* 2013; 13: S6.
21. Majumder AK, May M, Pant PD. Infant and child mortality determinants in Bangladesh: Are they changing? *J Biosoc Sci.* 1997; 29(4): 385-399.

22. Koenig MA, Phillips JF, Campbell OM, Dsouza S. Birth intervals and childhood mortality in rural Bangladesh. *Demogr.* 1990; 27(2): 251-265.
23. UNICEF. Every Child Alive. The urgent need to end newborn deaths. Genève, Switzerland: UNICEF; 2018.
24. Ezech OK, Agho KE, Dibley MJ, Hall J, Page AN. The impact of water and sanitation on childhood mortality in Nigeria: evidence from demographic and health surveys, 2003–2013. *Int J Res Pub Health.* 2014; 11(9): 9256-9272.
25. Mugo NS, Agho KE, Zwi AB, Damundu EY, Dibley MJ. Determinants of neonatal, infant and under-five mortality in a war-affected country: analysis of the 2010 Household Health Survey in South Sudan. *BMJ Glob Health.* 2018; 3(1): e000510.
26. Shiferaw S, Spigt M, Godefrooij M, Melkamu Y, Tekie M. Why do women prefer home births in Ethiopia? *BMC Preg. Childbirth.* 2013; 13: 5.
27. Machio, PM. Determinants of neonatal and under-five mortality in Kenya: Do antenatal and skilled delivery care services matter? *J Afri Dev.* 2018; 20(1): 59–67.
28. Bitew F, Nyarko SH. Modern contraceptive use and intention to use: implication for under-five mortality in Ethiopia. *Heliyon.* 2019; 5: e02295.
29. Nyarko SH, Tanle A, Kumi-Kyereme A. Determinants of childhood mortality in Ghana. *Int J Soc Sci Res.* 2014; 3: 61-77.
30. Azuine RE, Murray J, Alsafi N, Singh GK. Exclusive breastfeeding and under-five mortality, 2006-2014: A cross-national analysis of 57 low- and-middle income countries. *Int J MCH AIDS.* 2015; 4(1): 13-21.

Tables

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

Characteristics	Child Alive Before Age 5	Child Dead Before Age 5	Chi-Square test of equality
	Percent/ Mean	Percent/Mean	
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
1 st or 2 nd	94.7	5.3	
3 rd 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	
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*

All estimates include sample design and person weights, per DHS instructions. *t-test was used instead of Chi-Square. Significant variables are in bold.

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

Confusion matrix observed	Random Forest			Logistic Regression		KNN Model	
		Alive	Dead	Alive	Dead	Alive	Dead
Predicted	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							
			67.0		59.9		45.6
Positive predictive value							
			70.0		60.0		65.0
AUC			72.0		66.1		55.5

Table 3: Results from the logistic regression model

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: 1 st /2 nd)				
3rd or higher	2.129	1.131	4.008	0.020
Birth interval (Ref: <2)				
2-4 yrs	0.527	0.309	0.898	0.019
> 4 yrs	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
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 hour of birth	0.242	0.147	0.398	>0.0001
Household size	0.498	0.345	0.719	>0.0001

Significant variables are in bold.

Figures

Figure 1

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

Figure 2

ROC Curves for the three models

Variable Importance for Random Forest Model

Figure 3

Variable importance measures for Random Forest Model

Variable Importance for Logistic Regression Model

Figure 4

Variable importance measures for Logistic Regression Model

Figure 5

Variable importance measures for K-nearest neighbor Model