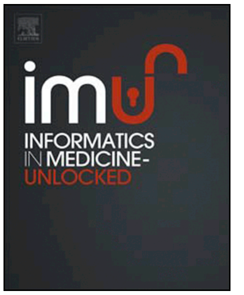
Journal Pre-proof



Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday

Addisalem Workie Demsash

PII: S2352-9148(23)00144-2

DOI: https://doi.org/10.1016/j.imu.2023.101298

Reference: IMU 101298

To appear in: *Informatics in Medicine Unlocked*

Received Date: 17 March 2023

Revised Date: 12 June 2023

Accepted Date: 17 June 2023

Please cite this article as: Workie Demsash A, Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday, *Informatics in Medicine Unlocked* (2023), doi: https://doi.org/10.1016/j.imu.2023.101298.

Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday

Addisalem Workie Demsash[a*]

[a]Mettu University, College of health science, department of health informatics, Ethiopia

***Correspondence author:** addisalemworkie599@gmil.com

Abstract

Introduction: Child morbidity and mortality in resource-limited settings is a major public health problem. The previous studies were mainly concerned with determining the prevalence of child deaths and identifying associated factors. Extracting knowledge and discovering insights from hidden patterns in child data through supervised machine learning algorithms is limited. Therefore, this study aimed to predict the under-five death of children using a best performance-supervised machine learning algorithm.

Methods: A total of 1813 samples were used from the 2019 Ethiopian Demographic and Health Survey dataset. 70% and 30% of total instances were used for training the model and measuring the performance of each algorithm with 10-fold cross-validation techniques respectively. Five supervised machine learning algorithms were considered for model building and comparison. All the included algorithms were evaluated using confusion matrix elements. Information gain value was used to select important attributes to predict child deaths. The **If/ then** logical association was used to generate rules based on relationships among attributes using Weka version 3.8.6 software.

Results: J48 is the second-best performance algorithm next to the random forest to predict child death, with 77.8% and 93.9% accuracy, respectively. Late initiation of breastfeeding, mothers with no formal education, short birth intervals, poor wealth status of the mother, and unexposed to media were the top five important attributes to predict child deaths. A total of six associated rules were generated that could determine the magnitude of child deaths. Of these, if children were rural residents, had a short birth interval, and if born as multiples (twins), then the probability of child death was 83.6%.

Conclusions: Five machine learning algorithms were included to predict child deaths and generate rules. Hence, the random forest algorithm was the best algorithm to predict child deaths. However, the study was limited since important attributes were not included in the data source, and irrelevant values were found. So, researchers are encouraged to use machine learning algorithms for future studies including important attributes that could predict child death. The current findings would be useful for stakeholders' preparedness, and taking proactive childcare interventions. Encouraging women in education, media access, and economic development programs are essential interventions for child death reduction.

Keywords: Child death, Prediction, Machine learning

Introduction

Under-five mortality is the most important indicator to measure the health status of children, and it is a key marker for the development of countries [1]. The under-five mortality rate is the probability of children dying before their fifth birthday [2]. Globally, nearly 44% of all under-five deaths occurred before their first month of birth [3], and an estimated 4.1 million child deaths occurred in 2017 [4]. According to the Centers for Disease Control and Prevention, child mortality in the United States in 2020 was predicted to be 5.4 deaths per 1,000 live births [5].

The risk of under-five mortality is highest in low-income countries. The under-five mortality rate in low-income countries was predicted to be 69 deaths per 1,000 live births in 2017, which is almost 14 times the rate in high-income countries [6, 7]. In Bangladesh, 522 under-five children died per 1,000 live births [8]. In 2001, under-five mortality in Nepal was projected to be 91 deaths per 1,000 live births [9]. Though under-five mortality shows a reduction from 166 to 67 per 1,000 live births over a period of 16 years [10], Ethiopia appears to have the fifth-highest number of new-born deaths in the world [11]. Under-five mortality is projected to cause 472,000 children to die annually in Ethiopia before their fifth birthday, which places Ethiopia sixth in the world according to the number of under-five deaths [7, 12]. According to WHO 2017, more than half of under-five deaths are due to infectious diseases that are easily preventable and treatable through simple and affordable interventions [13]. Under-five mortality is also caused by undernutrition, which further leads to stunting and wasting [14].

According to previous traditional logistic regression analysis, under-five mortality is associated with mothers' educational status and age, wealth status, the children's age (18), child size and birth order, poor sanitation, and unsafe drinking water, the wealth index in the community [15], distance to a health facility, and multiple born children [16]. Additionally, giving birth at a health facility, timely initiation of breastfeeding, mothers' preparedness for birth, media exposure, and professionals' knowledge [17] are important factors associated with under-five mortality. Even though the traditional logistic regression is suitable to determine the strength of the association between independent predictors with outcome variables, the odds ratio and relative risk do not meaningfully describe the individual predictors' ability to classify subjects, and it does not discover new insight [18]. Moreover, the complex and voluminous amount of data is less likely to be manageable in a traditional logistic regression model. Accordingly, machine learning algorithms have been used more effectively. The traditional logistic regression model is defined based on small input variables and sample sizes that would lead to incorrect relationships and reduce representativeness [19]. Hence, this makes it difficult for policymakers and stakeholders to take accurate interventions to solve the problems.

Nowadays, different machine-learning algorithms are used in public health research to predict and classify public health and biomedical data to discover new insights for a better understanding of relationships and patterns in input data [20]. From various machine learning algorithms, supervised machine learning is effective in forecasting disease prevalence, health service utilization, and maternal and child mortalities by labeling training, and testing data sets. Supervised machine learning algorithms are critical for the automated detection and discovery of meaningful patterns in data [21]. Additionally, supervised algorithms are used to find a non-linear relationship between the outcome variable and independent variables.

Previously, different studies have been done based on supervised machine learning algorithms such as decision trees, random forests, logistic regression, J48, and adaboost algorithms to predict under-five mortality [22]. Machine learning models were used for effective prediction of the undernutrition status of under-five children [23], machine learning techniques were used to predict the risk of neonatal mortality and morbidity [24], random forest and decision tree models were used to predict under-five mortality [20, 25], malnutrition among children and nutritional effects for humans were predicted using machine learning algorithms [25, 26]. This was an excellent step to discover unknown relationships, gain insight, and learn from the work. The previous machine learning-based studies are critical for establishing the baselines of the current and future research studies. However, some studies only used fewer machine learning algorithms for comparisons [7]. These would reduce the chance of comparability of the algorithms, and the best performance model might not be included.

As evidenced by the literature reviewed, studies about the prediction of under-five mortality based on machine learning modeling techniques are insufficient. Currently, voluminous, heterogeneous patient data are generated, and it is important to analyze and present these health-related data in machine learning algorithms. Policymakers and stakeholders need accurate predictions on various aspects of health parameters for effective actions. Researchers are needed to test and compare various prediction and classification algorithms to provide highly accurate results.

Moreover, this study serves as input for health program formulators and practitioners to make correct decisions. Specifically, stakeholders would use the findings of the study for setting interventions to reduce child deaths in resource-limited settings. The study would also be important for a child's mother to pay attention to the most important predictors of child deaths. This study employed supervised machine learning algorithms to train data and develop a predictive model on under-five child deaths. Hence, the study would have credibility for enhancing public health practices, and help as a framework for future similar research. Therefore, this study aimed to predict child death before celebrating their fifth birthday by using various supervised machine learning algorithms.

Research questions

1. Which supervised machine learning algorithm is best to determine child death before celebrating their fifth birthday?
2. Which important variables could predict child deaths before celebrating their fifth birthday?
3. Which important variables would jointly determine (generate association rules) child deaths before celebrating their fifth birthday?

Methods and materials

Study design and setting

The cross-sectional study design was conducted across the region of Ethiopia. Ethiopia is located in the Horn of Africa and bordered by Eritrea to the north, Djibouti, and Somalia to the east, Sudan and South Sudan to the west, and Kenya to the south. Ethiopia has nine regional states with two administrative cities. These are subdivided into different administrative units (68 zones, 817 woredas, and 16253 kebeles).

Data source

For this study, the 2019 Ethiopian Mini Demographic and Health Survey (EMDHS) dataset was used from the Demographic and Health Survey (DHS) website (https://dhsprogram.com). The 2019 EMDHS data represents Ethiopia's second DHS. The Ethiopian Federal Ministry of Health requested the Ethiopian Public Health Institute (EPHI) to implement the survey. The survey was conducted with the financial and technical support of the World Bank, UNICEF, and the United States Agency for International Development. The survey was conducted by EPHI in collaboration with the Central Statistical Agency from March 21 to June 28, 2019. The 2019 EMDHS generates data for measuring the progress of the health sector goals set under the Growth and Transformation Plan, which is closely aligned with the Sustainable Development Goals [27].

Sampling procedures and sample size of the study

Two-stage stratified cluster sampling was used. Each region was stratified into urban and rural areas. In the selected enumeration areas, a household listing operation was done, and the results were used as a sampling frame for household selection in the second stage. Finally, a fixed number of households per cluster were selected. Samples of enumeration areas were selected independently in each stratum through implicit stratification and equal proportional allocation. Finally, a total of 8885 samples of eligible women were included in the 2019 EMDHS data set. However, the data set did not include important attributes that would predict child deaths, had missing values, and some were recorded as not applicable. After removing missing values and irrelevant data, the total sample size of this study was 1813.

Study population, inclusion, and exclusion criteria

All eligible women aged 15–49 years old who were either permanent residents of the selected households or visitors who were present in the household the night before the survey were the respondents on behalf of their children. Therefore, all sampled under-five children were the study population[27].

Study variables

Dependent variable

The dependent variable of the study was **the death of children before celebrating their fifth birthday**.

Independent variables

Socio-demographic characteristics of households, such as wealth status, educational status of mothers, sex of children, preceding birth interval and birth order, age and sex of households' heads, initiation of breastfeeding, mothers' age, media exposure, the place of residency, and region, were extracted as potential attributes to predict child deaths before their fifth birthday.

Operationalization and measurement of included variables

Under-five child mortality: According to WHO, under-five child mortality is the death of children under the age of five (death before celebrating their fifth birthday) per 1,000 live births [2]. Hence, children who died before their fifth birthday were labelled as **yes**, else **no**.

Birth interval: The period between two successive live birth is a birth interval. For this study, a birth interval of <33 months between two consecutive live births is a **short birth interval**, whereas a birth interval of 33 and above is an **optimum birth interval [28, 29]**.

Early initiation of breastfeeding: Provision of the mother's breast milk to the infants within 1 hour after birth indicates **early breastfeeding (Yes)**. If the infants are provided mothers' breast milk after 1 hour of birth indicate late initiation of breastfeeding (No) [30].

Media exposure: If the mothers had access to either radio or television or both, then the mothers were had media exposure; otherwise unexposed to media [31].

Family's wealth status: The family's wealth index was generated from the wealth index of the households. In the 2019 EMDHS dataset, the wealth index has five quintiles, such as the lowest quintile (poorest), the second quintile (poorer), the third quintile (middle), the four quintiles (rich), and the fifth quintile (richest). For this study, the first and second wealth index categories as **'poor'**, the middle wealth index category was taken as **'middle'**, and the fourth and fifth wealth index categories were categorized as **'rich'** [32].

Data management and statistical analysis

Data cleaning and labelling were performed using STATA version 15 software to prepare the data for analysis. Variables were recoded to meet the desired classification. Respondents from small regions like Harari contribute a small sample, and respondents from large regions like Amhara and Oromia, contribute much more. In such a case, the sample might not be representative across the country, and so there is a need for mathematical adjustment to make the sample representative. Hence, to ensure the representativeness of the findings at the national level, sampling weights were done before the data analysis [33, 34]. The STATA version 15 software was used for data management and logistic regression analysis. Weka version 3.8.6 software was used for data pre-processing, important attribute selection that could predict child death, and generating rules associated with the death of children before their fifth birthday.

Data pre-processing

Data pre-processing is mainly concerned to manage noise, outliers, and inconsistency in the data set. Since the study was based on 2019 EMDHS data, irrelevant data are found in the records. Therefore, all these unnecessary data values were removed from the data set for this study. At this stage, all strings and categorical variables were also transformed into nominal data types. This approach of changing data type is critical to enhancing the accuracy of the result.

Feature selection

In this study, there were two stages of variable selection for model building. In the first stage, a traditional logistic regression analysis was employed for a feature or independent variable selection. A variable with a p-value of less than 0.2 with backward stepwise logistic regression analysis was considered as a candidate for further important attribute selection. A variance inflation factor (VIF=1/ (1-$R_{i,}^{2}$, R is the unadjusted coefficient of regressing in the i^{th} independent variable)) was used to test the possible existence of a correlation between independent variables. If the VIF value is **<1**, between **1** and **5**, and **>5**, indicates that there is no correlation, moderate correlation, and high correlation between independent variables, respectively [35]. Hence, the value of the VIF for all independent predictors was **2.75**. This revealed that there was no significant correlation between variables.

In the second stage, a best-performance algorithm model with information gain values was used to find important features or attributes that have a major contribution to predicting child death before celebrating their fifth birthday. The highest information gain value of an independent predictor indicates the most important attribute it is to predict the target variable and that it is highly correlated with the target variable (death of children before celebrating their fifth birthday). Then the next important features/predictors were selected based on their order of highest information gain value for model building.

Model building

Data split and model selection

In this step, 7:3 rule was considered for training and testing the model. From a total of 1813 observations, 70% and 30% were assumed for training and measuring the performance of the model, respectively. To ensure the accurate and equal classification of the available data as training and testing dataset, K-fold cross-validation techniques were used, and it is important when there is a small samples [36]. The K-fold cross-validation technique divides the available records into equal samples, and K-1 folds are used for training the predictive model, and the remaining folds are used for testing K-times repeatedly. The average number of K-times cross-validation was used as a performance measure.

Previous similar studies have used different supervised machine-learning algorithms to predict under-five child mortality [20, 25, 26]. Then various appropriate supervised machine learning algorithms such as Naïve Bayes, logistic regression, J48, random forest, and adaboost algorithms were used for predicting under-five child mortality.

Naïve Bayes: Naïve Bayes algorithm is a supervised machine learning algorithm, which is based on the Bayes theorem and used for the classification and prediction of problems. In the Naïve Bayes algorithm, attributes are conditionally independent for the target class [20]. Naïve Bayes has a computational efficiency in that number of attributes and classification time is linear with several attributes, and not affected by training time. Naive Bayes algorithms had an incremental learning behaviour, could directly predict patterns with low variance, and their performance is measured by confusion matrix elements [37].

Logistic regression: Logistic regression is a type of regression model that is important to model the categorical dichotomous outcome variable or feature. Logistic regression is a statistical model used to classify and predict different parameters in health [38]. It might be a binary (Binary logistic) and (multiple) model used to predict binary (multiple) outcome variables. Logistic regression has different assumptions, of which the target variable is dichotomous, and independent variables that affect the target variable are independent of each other [39].

J48 classifier algorithm: A J48 classifier algorithm is one of the best machine learning algorithms that examine categorical data based on a top-down recursive divide and conquer strategy [40]. J48 classifier is a simple C4.5 decision tree for classification to create a binary tree. The algorithm is crucial for classifying the problems, and the J48 algorithm is important to ignore the missing values and be able to predict the item of missing value based on what is known about the records of another attribute. The process is to divide the available data into ranges based on the attribute values for that item that are found in the training data, and then classification is done, and rules are generated from the attributes [41].

Random forest: A random forest is a supervised machine learning algorithm used to classify and predict health problems and health service utilization [42]. Random forest is the fastest to train and work with subsets of features. Random forest is important to detect complex relationships, including nonlinear and high-order interactions, and yields the smallest prediction errors [43]. **Adaboost:** Adaboost is an ensemble meta-learning method that enhances the efficiency of the binary classification tree. Adaboost uses an iterative approach to learn from the mistakes of weak classifiers and turn them into strong ones [44, 45]. Ada Boosting is crucial to boost the performance of decision trees based on binary classification problems [46]. The overall knowledge flow of model building for data processing, analysing, and visualizing are presented in **Figure 1**.

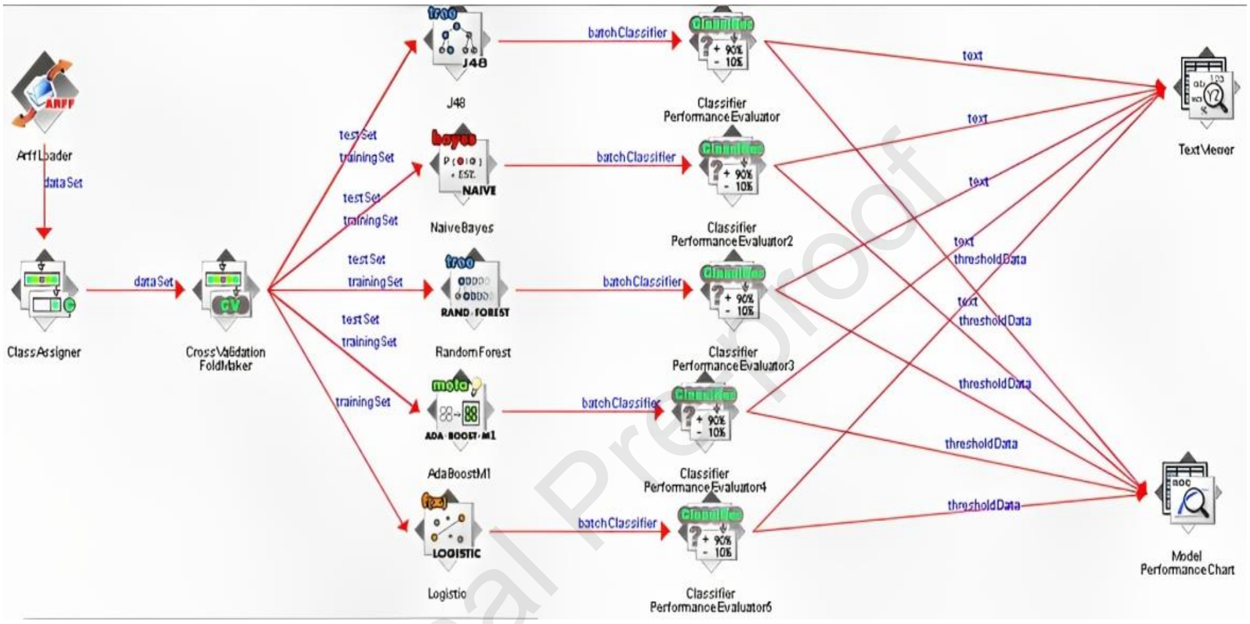


Figure 1: Knowledge and expert flow of the supervised machine learning algorithms

Imbalance data handling

Data imbalance mainly occurs in real-world applications such as medical diagnosis, pattern recognition, speech, and fraud detection. The number of observations in the classification dataset might have majority and minority classes [47]. For instance, the target variable of this study is a binary outcome: **Yes** (child died before their fifth birthday), **No** (child did not die before their fifth birthday). Hence, the classification might provide a high value for (the majority) for the Yes class, and a low value (the minority) for the No class. In such a case, the classification might give inaccurate and biased predictive results in machine learning. Therefore, handling imbalanced data is critical for accurate results. There is various technique to handle imbalanced data such as over-sampling, synthetic minority over-sampling technique (SMOTE), and under-sampling [48]. Under-sampling is about the random reduction of the majority (abundant class) to balance the data set. For this study, under-sampling was not considered to handle imbalanced data since it might lose important attributes and records in the dataset. Oversampling techniques are important to keep attributes and are used to fill in missing values. Whereas, SMOTE is an effective oversampling approach to handle imbalanced datasets. It creates new synthetic samples for the minority class by interpolating linearity between the minority class [48, 49]. SMOTE is a critical method to overcome overfitting in machine learning algorithms[50]. Hence, over-sampling and SMOTE were considered to ensure that the majority and minority classes have balanced data. This is very critical to reduce prediction errors, increase the use of data for both training and validation, and eliminate data overfitting. So, the overall model performance increased, and accurate results were generated. Consequently, 8.7% of additional synthetic records were generated to balance the minority class. Overall, the imbalanced data and balanced data were depicted in **Figure 2**.

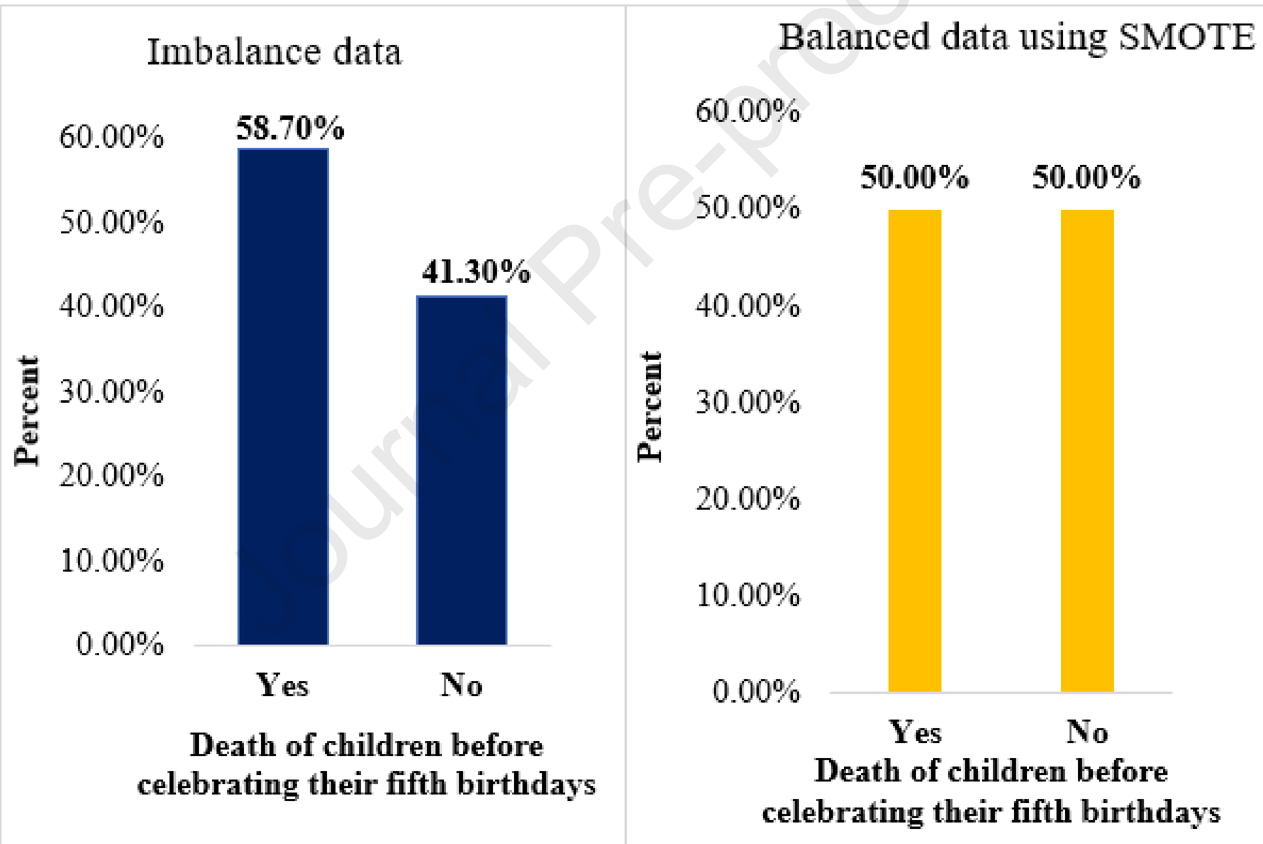


Figure 2: Distribution of death of children before celebrating their fifth birthday before and after data balancing, using the 2019 EDHS dataset.

Model evaluation

The performance of all the included algorithms has been evaluated using the confusion matrix. The accuracy of actual and predicted classes has been visualized by the confusion matrix model [51]. The predicted and actual classifications of under-five child mortality were compared using confusion matrix elements, such as true positive, false positive, true negative, and false-negative. The receiver operators' curve (ROC) was also used for model evaluation based on sensitivity, and specificity relationships. Since ROC is based on probability, the area under the ROC curve (AUC) is crucial to representing the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Hence, the higher the AUC, the better the model is at predicting true classes as true and false classes as false. Usually, the AUC value is good if it is greater than 80%, fair if it is between 70% and 80%, poor if it is between 60% and 70%, and failed if it is less than 60% [52]. The formula for the confusion matrix's element is presented in **box 1**.

Box 1: Formula for the element of the confusion matrix

$$\textbf{Accuracy} = (\text{True positive} + \text{True negative})/(\text{True Ppositive} + \text{True negative} + \text{False positive} + \text{False negative})$$

$$\textbf{Sensitivity} = \text{True positive}/(\text{True positive} + \text{False negative})$$

Not that Sensitivity=Recall=True Positive Rate

$$\textbf{Specificity} = \text{True negative}/(\text{True negative} + \text{False positive})$$

$$\textbf{False positive rate} = \text{False positive}/(\text{False positive} + \text{True negative})$$

$$\mathbf{F_{measure}} = 2 * \text{True positive}/(2 * \text{True positive} + \text{False positive} + \text{False negative})$$

$$\textbf{Precision} = \textbf{Postive predictive value} = \text{True positive}/(2(\text{True positve }) + \text{False positive}))$$

True positive: The model correctly predicts a positive class of response outcome.

False positive: The model incorrectly predicts a positive class in the response outcome.

True negative: The model correctly predicts a negative class in the response outcome.

False-negative: The model incorrectly predicts a negative class in the response outcome.

Sensitivity: Sensitivity is the test to measure correctly positive predicted events out of a total number of positive events, and it shows the value of how many positives are predicted out of total positive classes.

Specificity: Specificity is the proportion of real negative cases that got predicted as negative. This indicates that there will be another proportion of real negative cases, which would get predicted as positive and could be termed as false positives.

Precision: Precision is a positive predictive value, and it is the correct events divided by the total number of positive events that the classifier predicts.

F_measure: F measure is the inverse relationship between accuracy and recall. The higher value of the F-measure score predicts a better model.

Prediction and association rule mining

Once the model is built and its performance assessed, the death of children before their fifth birthday is predicted based on the independent predictors. Important variables selected based on a best-performance model were used to predict child mortality before their fifth birthday. Although important variables are used to predict child deaths before their fifth birthday, the predictive model does not show which nominal variables are jointly associated with child deaths before their fifth birthday.

Therefore, association rule mining analysis (the **If** (antecedent)/ **then** (consequent) statements) is used to discover relationships between seemingly independent relational attributes. Association rule mining analysis is important for non-numerical and categorical types of data attributes. It is important to observe frequently occurring patterns and identify the dependencies between attributes by supporting how frequently the if/then relationship appears in the observations and confidence in the number of times the relationships are true. The **if/ then** association rule mining analysis is critically important to select important features that jointly determine under-five child mortality and is the easiest way to interpret [53].

The **If then** association rule is the pair of X and Y (X, Y) attributes expressed as X->Y, where X is an antecedent and Y a consequent that is as X happens Y would also happen [54]. These rules are critically important for the prevention and control of health problems and crucial for health policymakers' proactive decision-making purposes. For the association rule mining analysis, the apriori algorithm method was used to identify strong and frequently related attributes.

Various studies had widely used **if/then** rules in healthcare research, such as predicting childhood care and child mortality [55], predicting parasite infection [56], the pattern of new cases and stroke [57, 58], and maternal healthcare service utilization and service discontinuation [59] to identify important features. The relationship between X and Y attributes is expressed in the following way [58].

If the left attribute >**1** | X and Y are positively associated to determine under-five child mortality. if the left attribute <**1** | X and Y negatively associated to determine under-five child mortality.

If the left attribute=**1** | No relation between X and Y to determine under-five child mortality.

The detail of data preparation, model building, important variable selection, and analysis workflow is presented in **Figure 3**.

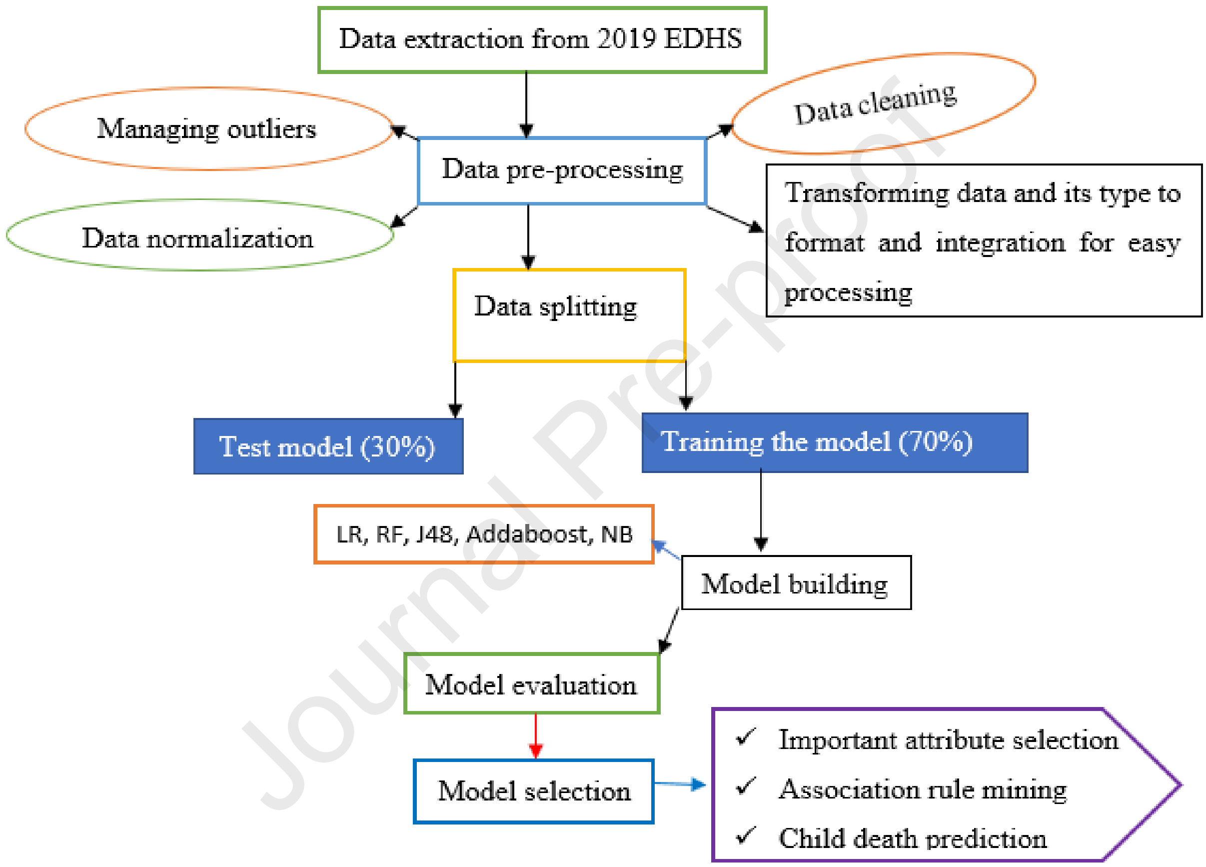


Figure 3: Workflow for data pre-processing, and child death prediction processing.

Results

Sociodemographic characteristics of the study participants

Following data pre-processing, a total of 1813 samples of children were included for analysis. From a total sample, almost three-fourths (74.02%) of children's mothers had no formal education. The average age of mothers was 16.67. Four out of ten (39.66%) and one-fourth (26.58%) of the children's mothers were from Oromia, and South Nation Nationality and People's Region (SNNPR), respectively. The majority (76.3%) and seven hundred sixty

(41.92%) of children were rural residents and from poor mothers, respectively. Six hundred forty-seven (35.68%) were orthodox religious flowers. Seven out of ten (69.1%) respondents had no media exposure, and 56% of children were male.

Table 1: Socio-demographic characteristics of children and respective respondents using the 2019 EDHS dataset.

Variable	Category	Frequency (n)	Percent (%)
Mothers' educational status	No formal education	1342	74.02
	Primary	445	24.44
	Secondary	25	1.37
	Higher	3	0.17
Region	Tigray	63	3.47
	Afar	18	0.97
	Amhara	352	19.42
	Oromia	719	39.66
	Somali	123	6.77
	Benishangul	29	1.60
	SNNPR	482	26.58
	Gambela	9	.50
	Harari	4	.27
	Addis Ababa	5	.26
	Dire Dawa	9	.50
Mothers' age (year)	15-19	24	1.32
	20-24	125	6.89
	25-29	361	19.92
	30-34	414	22.84
	35-39	476	26.25
	40-44	413	22.78
Family's wealth index	Poor	760	41.92
	Middle	456	25.15
	Rich	597	32.93
Mother/caregiver religion	Orthodox	647	35.68
	Catholic	13	.72

	Protestant	628	34.64
	Muslim	504	27.80
	Traditional, and other	21	1.16
Place of residency	Urban	429	23.7
	Rural	1384	76.3
Sex of children	Male	1015	56.0
	Female	798	44.0
Media exposure	No	1253	69.1
	Yes	560	30.9

Performance of the included models in predicting child deaths before their fifth birthday

Five machine learning algorithms were included to predict under-five mortality in Ethiopia. The Naïve Bayes, J48, random forest, adaboost, and logistic regression classification algorithms were involved in predicting child mortality. All models built based on the included five machine learning algorithms were evaluated using performance measures from the confusion matrix. The accuracy, AUC, F-measure, and precision were used to evaluate the models' performances. If the AUC value is close to the left-top corner, show the best performance model. Accordingly, the random forest is the best performance model to predict child die before their fifth birthday, with an AUC value of 93.9%, and its sensitivity, specificity, precision, and f-measure indicate 90.7%, 84.4%, 84%, and 87.2%, respectively. From a total of 1813 instances, 1702 instances (93.9% accuracy) were correctly classified, and the remaining 111 instances (6.1%) were incorrectly classified. J48 was the second-best performance model for the prediction of children's death before their fifth birthday, with an AUC value of 77.8%. The sensitivity, specificity, precision, and f-measure value for the j48 algorithms was 89.0%, 78.7%, 75.5%, and 76.2%, respectively. The overall details of the algorithms' performance measures for under-five child mortality were presented in **Table 2** and **Figure 4**.

Table 2: Model performance of all the include supervised machine learning algorithms.

Parameters (%)	**Classification algorithms**				
	J48	Naïve Bayes	Random forest	Adaboost	Logistic regression
Sensitivity	89.0	83.5	90.7	82.4	84.9
Specificity	78.7	75.3	84.4	74.3	76.6

Precision	75.5	62.6	84.0	61.8	63.6
F-measure	76.2	72.9	87.2	71.3	73.0
AUC	77.8	61.3	93.9	60.9	62.3

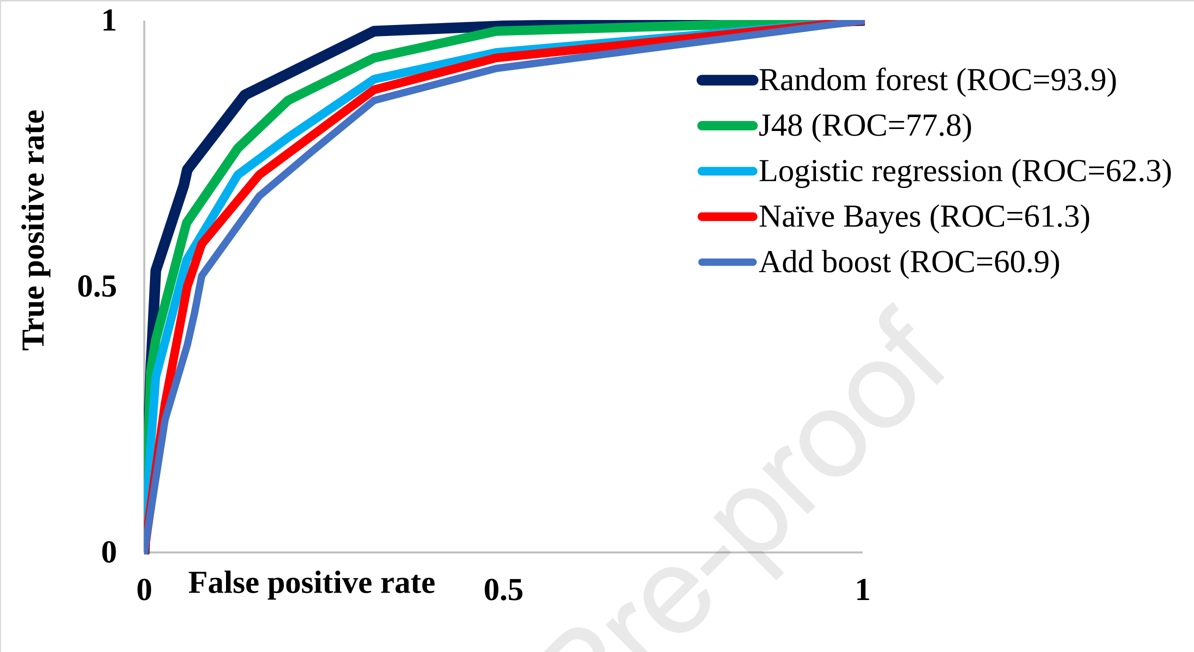


Figure 4: ROC Curve for AUC of each algorithm model

Importance attributes to predict the death of children before their fifth birthday

Important predictors of under-five child mortality were measured based on gain information coefficients with a 10-fold cross-validation process. For important predictor selection, the best performance model (random forest) was considered. Accordingly, late initiation of breastfeeding, no formal education of mothers, short birth intervals, poor wealth status, and unexposed to media were the top five important attributes of the death of children before their fifth birthday. Other important attributes of child death before celebrating their fifth birthday were presented in **Table 3** and **Figure 5.**

Table 3: Information gain value for each important predictor variable.

Predictor variables	**Type**	**Measurement**	**Information gain value**
Early breastfeeding, (No)	Nominal	Scale	0.083
Educational status, (No)	Nominal	Scale	0.069
Birth interval, (Short)	Nominal	Scale	0.067
Wealth status, (Poor)	Nominal	Scale	0067

Media exposure, (Unexposed)	Nominal	Scale	0.058
Place of residency, (Rural)	Nominal	Scale	0.057
Sex of children, (Female)	Nominal	Scale	0.019
Age of mother, (15-19 years)	Nominal	Scale	0.018
Birth order, (>5)	Nominal	Scale	0.014

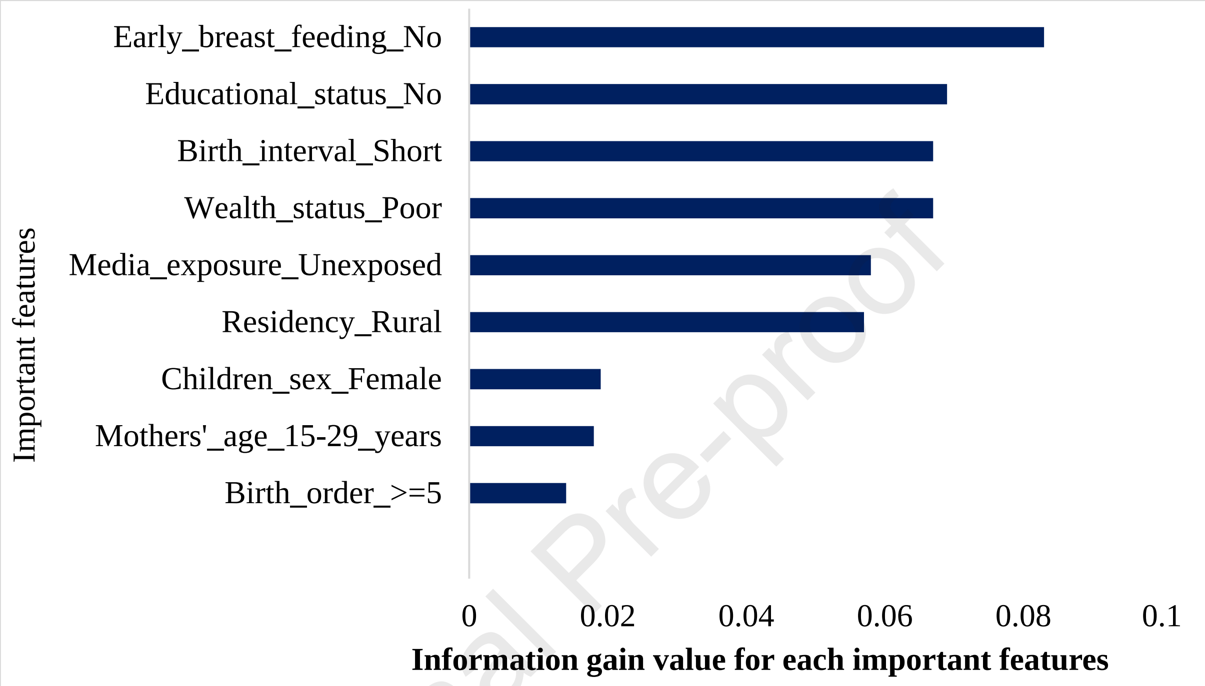


Figure 5: Important variable selection using information gain value

Association rule building

Rule 1: If residency=0 (Rural), birth interval =1 (short), and child twin=2 (Multiple), **then** the probability of child death before celebrating their fifth birthday would be 83.6% (left=1.67).

Rule 2: If birth interval =1 (Short), early breastfeeding=0 (No), and mothers' educational status=1 (No education), **then** the probability of child death before celebrating their fifth day birthday would be 80.3% (left=1.59).

Rule 3: If the child twin=1 (Single), early breastfeeding =0 (No), and the mothers' wealth status=0 (Poor), **then** the probability of child death before celebrating their fifth birthday would be 77.8% (left=1.51).

Rule 4: If wealth status=2 (Middle), media exposure =0 (No), residency= 0 (Rural), and mothers' age=1 (15-19 years of age), **then** the probability of child death before celebrating their fifth birthday would be 74.1% (left=1.43).

Rule 5: If early breast feeding=1 (Yes), wealth status=0 (Poor), and birth interval=1 (Short), **then** the probability of child death before celebrating their fifth birthday would be 72.5% (left=1.38).

Rule 6: If residency=1 (Urban), media exposure =0 (No), and wealth status=0 (Poor), **then** the probability of child death before celebrating their fifth birthday would be 59.7% (left=1.23).

Discussion

For this study, the 2019 EDHS dataset was used, with a total of 1813 instances. 70% of the total instances were randomly sampled and trained in each algorithm, and 30% of total instances were also randomly sampled and used to test the algorithms' performance. The aim was to compare different supervised machine learning algorithms using the confusion matrix element and determine important attributes that are relatively accurate in predicting the death of children before celebrating their fifth birthday. Accordingly, five supervised machine learning algorithms, such as Naïve Bayes, logistic regression, J48, random forest, and adaboost algorithms, were included to achieve the objectives of the study.

The model performance of each of the five supervised machine learning algorithms was compared by their classification accuracy and AUC score values. According to the analysis report, the sensitivity and AUC value of the random forest model was 90.7% and 93.9% with 10-fold cross-validation, respectively. Hence, the random forest algorithm model was the most accurate model to predict child death before celebrating their fifth birthday other remaining algorithms. This finding was supported by various similar studies that report random forest is the best model to predict under-five child mortality [7, 23], contraceptive discontinuation [59], and important features for stunting and malnutrition among under-five children [60-62]. The J48 algorithm was the second-best model for predicting the death of children before celebrating their fifth birthday, with 89.0% and 77.8% of sensitivity and AUC, respectively. This finding was supported by a report that states the J48 algorithm is the best model for predicting under-five child mortality [22].

In this study, the second objective was to select important attributes that could predict the death of children before celebrating their fifth birthday. From the attributes selected to predict child death before celebrating their fifth birthday, late initiation of breastfeeding, mothers having no formal education, short birth intervals for children, poor wealth status of the mother, and being unexposed to media were the top five important attributes to predict child death before celebrating their fifth birthday. Other attributes, such as being a rural resident, being female, having a mother's age between 15-19 years old, and having a birth order of less than five were also important attributes for a child's death before celebrating their fifth birthday.

Late initiation of breastfeeding was the top-ranked attribute for predicting the death of children before celebrating their fifth birthday. This finding was supported by similar studies that found that children who were not breastfed were more likely to die before their fifth birthday [7, 22, 63]. This might be because children might not get important nutrients and complementary foods from breast milk promptly [64, 65]. Moreover, late initiation of breastfeeding might contribute to child stunting and infection-related neonatal mortality among all live births [66, 67], and children might get nutrients from breast milk that have a significant benefit for ensuring child health and survival. Mothers' lack of formal education was the second important attribute to predict child death before celebrating the fifth birthday. This finding is supported by similar studies done in Ethiopia [22], and Nigeria [68]. This might be because uneducated women might not have awareness and knowledge about childcare and feeding, and low-educated mothers might not have the skills to access health information or not purposefully seek it [69]. Consequently, children's mothers might delay accessing appropriate health services at the right time [70], and so children might be more likely to die before their fifth birthday.

A short birth interval was the third most important attribute to predict child death before celebrating their fifth birthday. This finding was in line with studies done in Ethiopia [22], and similar resource-limited settings [71, 72]. This might be because short birth intervals might lead to high fertility and population growth, which undermine child development stages, and so children are at high risk for dying before their fifth birthday [73]. Being poor, and media unexposed were another fourth and fifth important attributes to predict child deaths before celebrating their fifth birthday. This might be because poor women cannot afford to feed and care for their children, and women who have not had media exposure might not access information that could be critical for child care and survival.

Generating rules for child mortality was another objective of this study. Consequently, six rules associated with child mortality were generated. According to Rule 1, the probability of a child dying before reaching the age of five is 83.6% if the child is a rural resident, has a short birth interval, and is born as a multiple (twin). This might be because a rural area health facility might not be available, short birth intervals might cause many children to be present and unable to get the appropriate and recommended nutrients, and mothers might not properly manage and care for multiple children born at the time. So, a combination of these factors might be the cause of a child's death before celebrating their fifth birthday. The probability of child death was 80.3% if there was a short birth interval among children, children did not initiate breastfeeding early, and mothers were not formally educated. The **if/ then rules** are critical to discovering the hidden relationship between attributes, extracting knowledge from a set of data, and being accurate in representing knowledge and information about child mortality. This is vital to support public health proactive action, decision-making purposes, and the storage of knowledge.

Strengths and limitations of the study

In this study, five supervised machine learning algorithms were used to classify and predict child deaths before their fifth birthday. Association rules were also generated that more than one variable would predict a child’s death jointly. This study used nationally representative data, and the findings might have a representative nature.

However, the supervised machine learning algorithms do not have coefficients like odds and incident rate ratios. Therefore, the strength and direction of associations are unknown. Moreover, since the data source was the 2019 EMDHS data set, important attributes such as birth weight and birth size of children were not included. Plus, the current study was more emphasized mothers and child related attributes, that fathers’ related attributes such as fathers’ education, income level were missed, and though the author tries to address issues of endogeneity, we readers should kindly be informed that the issues of endogeneity may alter the interpretation of the result. Hence, the author recommends future researchers conduct similar studies by addressing the limitation mentioned in this study.

Conclusions and recommendations

This study aimed to identify the best-supervised machine learning algorithms to classify and select important attributes to predict the death of children before their fifth birthday. In line with the objectives, six supervised machine learning algorithms were considered that accurately predict the death of children before celebrating their fifth birthday. Different confusion matrix element was used to compare the candidate-supervised machine learning algorithms. Based on the result, the random forest algorithm was the best performance model to predict the death of children before celebrating their fifth birthday. Attributes such as late initiation of breastfeeding, mothers with no formal education, the short birth interval, poor wealth status of the mother, and being unexposed to media were the top important attributes to predict child deaths.

Generating associated rules for child death was another objective of the study. Accordingly, six rules were generated that were associated with the deaths of children before celebrating their fifth birthday. The findings of this study would have practical implications by supporting policymakers and stakeholders in developing childcare intervention mechanisms and preparing themselves to care for children as early as possible. Stakeholders are recommended to encourage mothers to initiate breastfeeding at the appropriate time. Improving mothers' wealth status, closing the gap in media access, and creating awareness among mothers would be critical interventions to enhance the survival of children in Ethiopia. The generated rules would also have theoretical implications by extracting and representing knowledge. Moreover, researchers would use this study as a baseline and framework for further research studies, including important attributes that would predict child mortality in low-income countries.

Abbreviations

AUC: Area Under ROC curve**, DHS:** Demographic and Health Survey, **EDHS:** Ethiopian Demographic and Health Survey**, EPHI:** Ethiopian Public Health Institute, **ROC:** Recursive Operator Characteristics, **SMOTE:** synthetic minority over-sampling technique, **SNNPR:** South Nation Nationality and People's region.

Declarations

Ethical approval and consent to participate

Ethical clearance was not necessary for this study since it was based on publicly available data sources. Informed consent from the study participants was also not applicable to this study.

Consent for publication

Not applicable.

Availability of data and materials

The dataset used for analysis is available on the DHS program website. All the data generated and analyze are included in the study.

Patient and public participation

Not applicable.

Funding

No funding was received.

Author's contributions

AWD had substantial contributions to the study design, conception, data management, and analysis, result in the interpretation, and discussion of the findings. The author read and approved the final manuscript for submission.

Reference

1. Khazaei, S., et al., *Variations of infant and under-five child mortality rates around the world, the role of human development index (HDI).* Int J Pediatr, 2016. **4**(5): p. 1671-1677.
2. *Under-five mortality rate, World health organization*. https://www.who.int/data/nutrition/nlis/info/under-five-mortality-rate.
3. Unicef, *statistical snapshot. Child mortality: Accessed from https://data.unicef.org/resources/2013-statistical-snapshot-child-mortality/.* New York, 2013.
4. Organization, W.H., *Meeting report: WHO technical consultation: nutrition-related health products and the World Health Organization model list of essential medicines–practical considerations and feasibility: Geneva, Switzerland, 20–21 September 2018*. 2019, World Health Organization.
5. CDC, *Reproductive health, infant mortality mortality: Accessed from https://www.cdc.gov/reproductivehealth/maternalinfanthealth/infantmortality.htm*. 2022.
6. Unicef, *Levels & trends in child mortality: report 2012: estimates/developed by the UN Inter-Agency Group for Child Mortality Estimation.* 2012.
7. Bitew, F.H., et al., *Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey.* Genus, 2020. **76**: p. 1-16.
8. Abir, T., et al., *Risk factors for under-5 mortality: evidence from Bangladesh Demographic and Health Survey, 2004–2011.* BMJ Open, 2015. **5**(8): p. e006722.
9. Ghimire, P.R., et al., *Under-five mortality and associated factors: evidence from the Nepal demographic and health survey (2001–2016).* International Journal of environmental research and public health, 2019. **16**(7): p. 1241.
10. *Under-five mortality in Ethiopia*. 2016. https://dhsprogram.com/publications/publication-fr328-dhs-final-reports.cfm.
11. Sahile, A., D. Bekele, and H. Ayele, *Determining factors of neonatal mortality in Ethiopia: An investigation from the 2019 Ethiopia Mini Demographic and Health Survey.* Plos one, 2022. **17**(12): p. e0267999.
12. MoH, F., *National strategy for child survival in Ethiopia.* Addis Ababa, Ethiopia, 2005: p. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwilipjng7P9AhWC8LsIHdqMCxIQFnoECAoQAQ&url=https%3A%2F%2Fextranet.who.int%2Fnutrition%2Fgina%2Fsites%2Fdefault%2Ffilesstore%2FETH%25202005%2520National%2520Strategy%2520for%2520Child%2520Survival.pdf&usg=AOvVaw1UdLjqk5Tv0rW8sVijqyp5.

13. Organization, W.H., *World health statistics 2016: Monitoring health for the SDGs sustainable development goals*. 2016: World Health Organization.
14. Bitew, F.H., C.S. Sparks, and S.H. Nyarko, *Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia.* Public Health Nutrition, 2022. **25**(2): p. 269-280.
15. Chilot, D., et al., *Geographical variation of common childhood illness and its associated factors among under-five children in Ethiopia: a spatial and multilevel analysis.* Scientific Reports, 2023. **13**(1): p. 868.
16. Kadobera, D., et al., *The effect of distance to the formal health facility on childhood mortality in rural Tanzania, 2005–2007.* Global health action, 2012. **5**(1): p. 19099.
17. Tegene, T., et al., *Newborn care practice and associated factors among mothers who gave birth within one year in Mandura District, Northwest Ethiopia.* Clinics in Mother and Child Health, 2015. **12**(1).
18. Pepe, M.S., et al., *Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker.* American Journal of Epidemiology, 2004. **159**(9): p. 882-890.
19. Ij, H., *Statistics versus machine learning.* Nat Methods, 2018. **15**(4): p. 233.
20. Saroj, R.K., et al., *Machine Learning Algorithms for understanding the determinants of under-five Mortality.* BioData mining, 2022. **15**(1): p. 1-22.
21. Osisanwo, F., et al., *Supervised machine learning algorithms: classification and comparison.* International Journal of Computer Trends and Technology (IJCTT), 2017. **48**(3): p. 128-138.
22. Tesfaye, B., et al., *Determinants and development of a web-based child mortality prediction model in resource-limited settings: a data mining approach.* Computer methods and programs in biomedicine, 2017. **140**: p. 45-51.
23. Fenta, H.M., T. Zewotir, and E.K. Muluneh, *A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones.* BMC Medical Informatics and Decision Making, 2021. **21**(1): p. 1-12.
24. Jaskari, J., et al., *Machine learning methods for neonatal mortality and morbidity classification.* Ieee Access, 2020. **8**: p. 123347-123358.
25. Thangamani, D. and P. Sudha, *Identification of malnutrition with use of supervised data mining techniques–decision trees and artificial neural networks.* Int J Eng Comput Sci, 2014. **3**(09).
26. Kuttiyapillai, D. and R. Ramachandran, *Improved text analysis approach for predicting effects of nutrient on human health using machine learning techniques.* IOSR J Comput Eng, 2014. **16**(3): p. 86-91.
27. Ethiopian Public Health Institute (EPHI) [Ethiopia] and ICF. Rockville, M., USA: EPHI and ICF., *Ethiopia Mini Demographic and Health Survey 2019: Final Report*. 2021:Accessed from. https://dhsprogram.com/publications/publication-FR363-DHS-Final-Reports.cfm.
28. Wakeyo, M.M., et al., *Short birth interval and its associated factors among multiparous women in Mieso agro-pastoralist district, Eastern Ethiopia: A community-based cross-sectional study.* Front Glob Womens Health, 2022. **3**: p. 801394.
29. Kassie, S.Y., et al., *Spatial distribution of short birth interval and associated factors among reproductive age women in Ethiopia: a spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey.* BMC Pregnancy and Childbirth, 2023. **23**(1): p. 1-14.
30. Lyellu, H.Y., et al., *Prevalence and factors associated with early initiation of breastfeeding among women in Moshi municipal, northern Tanzania.* BMC Pregnancy Childbirth, 2020. **20**(1): p. 285.
31. Demsash, A.W., et al., *Spatial distribution of vitamin A rich foods intake and associated factors among children aged 6–23 months in Ethiopia: a spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey.* BMC Nutrition, 2022. **8**(1): p. 77.
32. Demsash, A.W., et al., *Spatial and multilevel analysis of sanitation service access and related factors among households in Ethiopia: using 2019 Ethiopian national dataset.* PLOS Global Public Health, 2023. **3**(4): p. e0001752.

33. Levy, P.S. and S. Lemeshow, *Sampling of populations: methods and applications*. 2013: John Wiley & Sons.
34. Demsash, A.W., and A.D. Walle, *Women's health service access and associated factors in Ethiopia: application of geographical information system and multilevel analysis.* BMJ Health & Care Informatics, 2023. **30**(1): p. e100720.
35. *Variance Inflation Factor (VIF)*. 2023. https://www.investopedia.com/terms/v/variance-inflation-factor.asp.
36. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
37. Webb, G.I., E. Keogh, and R. Miikkulainen, *Naïve Bayes.* Encyclopedia of machine learning, 2010. **15**: p. 713-714.
38. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Applied logistic regression*. Vol. 398. 2013: John Wiley & Sons.
39. Uddin, S., et al., *Comparing different supervised machine learning algorithms for disease prediction.* BMC medical informatics and decision making, 2019. **19**(1): p. 1-16.
40. Kaur, G. and A. Chhabra, *Improved J48 classification algorithm for the prediction of diabetes.* International journal of computer applications, 2014. **98**(22).
41. Sharma, A.K. and S. Sahni, *A comparative study of classification algorithms for spam email data analysis.* International Journal on Computer Science and Engineering, 2011. **3**(5): p. 1890-1895.
42. Saroj, R.K., et al., *Machine Learning Algorithms for understanding the determinants of under-five Mortality.* BioData Min, 2022. **15**(1): p. 20.
43. Fu, G., X. Dai, and Y. Liang, *Functional random forests for curve response.* Sci Rep, 2021. **11**(1): p. 24159.
44. Tkachev, V., et al., *Flexible Data Trimming Improves Performance of Global Machine Learning Methods in Omics-Based Personalized Oncology.* Int J Mol Sci, 2020. **21**(3).
45. Yu, Y., et al., *Machine Learning Methods for Predicting Long-Term Mortality in Patients After Cardiac Surgery.* Front Cardiovasc Med, 2022. **9**: p. 831390.
46. *What is AdaBoost Algorithm Model?: Accessed from https://data-flair.training/blogs/adaboost-algorithm/.*
47. Alghamdi, M., et al., *Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project.* PloS One, 2017. **12**(7): p. e0179805.
48. *Handling Imbalanced Datasets in Machine Learning*. 2020. https://www.section.io/engineering-education/imbalanced-data-in-ml/.
49. Zenu, S., et al., *Determinants of first-line antiretroviral treatment failure among adult patients on treatment in Mettu Karl Specialized Hospital, South West Ethiopia; a case-control study.* Plos one, 2021. **16**(10): p. e0258930.
50. Elhassan, T. and M. Aljurf, *Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method.* Global J Technol Optim S, 2016. **1**: p. 2016.
51. Narkhede, S., *Understanding auc-roc curve.* Towards Data Science, 2018. **26**(1): p. 220-227.
52. El Khouli, R.H., et al., *Relationship of temporal resolution to diagnostic performance for dynamic contrast-enhanced MRI of the breast.* Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2009. **30**(5): p. 999-1004.
53. Molnar, C., *Interpretable machine learning*. 2020: Lulu. com.
54. Shi, R., et al., *Obesity is negatively associated with dental caries among children and adolescents in Huizhou: a cross-sectional study.* BMC Oral Health, 2022. **22**(1): p. 76.
55. Ivančević, V., et al., *Using association rule mining to identify risk factors for early childhood caries.* Computer methods and programs in biomedicine, 2015. **122**(2): p. 175-181.

56. Zafar, A., et al., *Machine learning-based risk factor analysis and prevalence prediction of intestinal parasitic infections using epidemiological survey data.* PLOS Neglected Tropical Diseases, 2022. **16**(6): p. e0010517.
57. Tandan, M., et al., *Discovering symptom patterns of COVID-19 patients using association rule mining.* Computers in biology and medicine, 2021. **131**: p. 104249.
58. Li, Q., et al., *Mining association rules between stroke risk factors based on the Apriori algorithm.* Technology and Health Care, 2017. **25**(S1): p. 197-205.
59. Kebede, S.D., et al., *Prediction of contraceptive discontinuation among reproductive-age women in Ethiopia using Ethiopian Demographic and Health Survey 2016 Dataset: A Machine Learning Approach.* BMC Medical Informatics and Decision Making, 2023. **23**(1): p. 1-17.
60. Fenta, H.M., et al., *Determinants of stunting among under-five years children in Ethiopia from the 2016 Ethiopia Demographic and Health Survey: Application of ordinal logistic regression model using complex sampling designs.* Clinical Epidemiology and Global Health, 2020. **8**(2): p. 404-413.
61. Talukder, A. and B. Ahammed, *Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh.* Nutrition, 2020. **78**: p. 110861.
62. Kassie, G.W. and D.L. Workie, *Determinants of under-nutrition among children under five years of age in Ethiopia.* BMC Public Health, 2020. **20**(1): p. 1-11.
63. Alemayehu, T., J. Haidar, and D. Habte, *Determinants of exclusive breastfeeding practices in Ethiopia.* Ethiopian Journal of Health Development, 2009. **23**(1).
64. Woldeamanuel, B.T., *Trends, and factors associated with early initiation of breastfeeding, exclusive breastfeeding, and duration of breastfeeding in Ethiopia: evidence from the Ethiopia demographic and health survey 2016.* International breastfeeding journal, 2020. **15**(1): p. 1-13.
65. Arimond, M. and M.T. Ruel, *Progress in developing an infant and a child feeding index: an example using the Ethiopia Demographic and Health Survey 2000*. 2002.
66. Debes, A.K., et al., *Time to initiation of breastfeeding and neonatal mortality and morbidity: a systematic review.* BMC public health, 2013. **13**: p. 1-14.
67. Edmond, K.M., et al., *Delayed breastfeeding initiation increases the risk of neonatal mortality.* Pediatrics, 2006. **117**(3): p. e380-e386.
68. Antai, D., *Regional inequalities in under-5 mortality in Nigeria: a population-based analysis of individual-and community-level determinants.* Population health metrics, 2011. **9**: p. 1-10.
69. Demsash, A.W., et al., *Spatial distribution of vitamin A rich foods intake and associated factors among children aged 6–23 months in Ethiopia: a spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey.* BMC Nutrition, 2022. **8**(1): p. 1-14.
70. Demsash, A.W., and A.D. Walle, *Women's health service access and associated factors in Ethiopia: application of geographical information system and multilevel analysis.* BMJ Health & Care Informatics, 2023. **30**(1).
71. Darroch, J.E., G. Sedgh, and H. Ball, *Contraceptive technologies: responding to women's needs.* New York: Guttmacher Institute, 2011. **201**(1): p. 1-51.
72. Kozuki, N. and N. Walker, *Exploring the association between short/long preceding birth intervals and child mortality: using reference birth interval children of the same mother as a comparison.* BMC public health, 2013. **13**(3): p. 1-10.
73. Wakeyo, M.M., et al., *Short Birth Interval and Its Associated Factors among Multiparous Women in Mieso Agro-Pastoralist District, Eastern Ethiopia: A Community-Based Cross-sectional Study.* Frontiers in Global Women's Health, 2022: p. 105.

Acknowledgment

We would like to express our deepest appreciation to the DHS program for permitting data access.

Competing interests

The author declared that there is not any competing interests.