Research Article

Journal of Epidemiology and Medical Statistics

Predicting Neonatal Mortality Using Demographic and Health Survey Data: Comparing Logistic Regression and Machine Learning under Severe Class Imbalance

Erick Cheruiyot Kirui[1*] and Stephen Muthii Wanjohi[2]

Department of Mathematics and Actuarial Science, Murang'a University of Technology, Kenya

[*]Corresponding authors
Erick Cheruiyot Kirui, Department of Mathematics and Actuarial Science, Murang'a University of Technology, Kenya.

Received: July 20, 2026; **Accepted:** July 27, 2026; **Published:** August 04, 2026

Abstract

Neonatal mortality continues to be a significant public health problem, especially in low and middle-income countries, where improvements in neonatal mortality have been slower than progress in overall child survival. Logistic regression has been applied in the past to determine the factors associated with neonatal mortality from the Demographic and Health Survey (DHS) data, but machine learning techniques have been gaining popularity as alternative methods for predictive modelling. Evidence comparing these approaches based on nationally representative household survey data, however, is limited, especially when considering the case of severe class imbalance. This study evaluated alternative strategies to address class imbalance and examined the effect of outcome-endogenous predictors to compare the predictive performance of logistic regression, random forest and gradient boosting classifiers with respect to the prediction of neonatal mortality. Secondary analysis was performed with data from 18,978 live births with 422 neonatal deaths (2.22%) and 27 predictors related to the mother, household, and child. The data were split with stratified sampling into training set (75%) and testing set (25%). Three strategies of handling imbalanced data were used in this study: baseline (no adjustment), class weighting, and Synthetic Minority Over-sampling Technique (SMOTE). The area under the receiver operating characteristic curve (ROC-AUC), the area under the precision-recall curve (PR-AUC), the sensitivity, the specificity, the precision, the F1 score, the calibration curve and the permutation feature importance were used to assess model performance. Class-weighted gradient boosting had the best overall predictive performance (ROC-AUC = 0.921, PR-AUC = 0.311, sensitivity = 83.0%) and logistic regression had similar discrimination ability but was easier to interpret (ROC-AUC = 0.920). Explicitly dealing with class imbalance significantly outperformed unadjusted models for sensitivity across all of the algorithms. The number of living children under five years, number of births in the last five years and gestation length were the most important variables identified by the feature importance analysis. When the mechanically outcome-related predictor, "number of living children under 5 years," was removed, the ROC-AUC dropped from 0.921 to 0.700, suggesting that a significant amount of the apparent predictive performance was due to information leakage and not to the predictor itself. These results show that the solution for class imbalance problem is more important than using different classification methods, and emphasize the importance of selecting predictors for temporal and definitional relationship with the outcome. The study gives methodological directions for creating accurate predictions to neonatal mortality and other rare health phenomena based on survey data nationally representative.

Background of the Study

The neonatal period (first 28 completed days of life) is still one of the biggest challenges in global child health [1]. While the overall under-five mortality rate has declined significantly in the world since 1990, progress against neonatal mortality has not been as rapid as that against post-neonatal and child mortality, and in recent years the burden of neonatal mortality has risen. This change is consistent with a simple epidemiological fact: post-neonatal deaths are driven by infections and nutrition, and can be addressed relatively rapidly by community-based interventions like immunization, oral rehydration, and control of vectors, while neonatal deaths are driven by intrapartum complications, prematurity, and severe infection, and require timely, more intensive clinical care around the time of birth. Consequently, the neonatal period is now the biggest and most persistent cause of under-five mortality, and the Sustainable

Citation: Erick Cheruiyot Kirui, Stephen Muthii Wanjohi. Predicting Neonatal Mortality Using Demographic and Health Survey Data: Comparing Logistic Regression and Machine Learning under Severe Class Imbalance. J Epidemiol Med Stat. 2026. 1(2): 1-10. DOI: doi.org/10.61440/JEMS.2026.v1.05

Development Goal (SDG) target of 12 or fewer neonatal deaths per 1,000 live births in all countries by 2030 has been set [2].

Kenya is a case in point of the scale and complexity of the challenge. Though the country has experienced significant advances in under five survival over the last two decades, the neonatal deaths remain the largest and highest proportion of under five deaths, and successive rounds of the Kenya Demographic and Health Survey (KDHS) have consistently identified the same set of demographic and socioeconomic factors as being associated with increased neonatal risk: maternal age, birth order, birth interval, multiple births, household wealth, maternal education and place of residence. The results are consistent with the general sub-Saharan African body of literature which also points to low socioeconomic status, high parity, low birth interval and poor access to skilled birth attendance as being common risk factors. There are three gaps identified in this review. First, although comparisons between machine learning and logistic regression are common for clinical data on higher prevalence neonatal mortality, such comparisons in lower prevalence, population-representative household survey data are much less prevalent and are limited to the type of data collected by DHS programs and analysed in this work. Second, while many remedies have been developed in the literature for the class imbalance issue, few neonatal mortality studies compare these remedies explicitly and side by side on a common evaluation set, nor do they justify or compare them in order to select the most appropriate remedy. Third, there is a lack of quantitative demonstration, to the authors' knowledge, in the Kenyan or sub-Saharan African DHS-based neonatal mortality literature, of the risk of outcome-endogenous predictors distorting predictive performance, these being highly common in the DHS literature on neonatal mortality. This study aims to fill in all three gaps in a coherent and integrated analytical framework.

This classic problem has been studied extensively and is motivated by two developments. First, the coded birth-history datasets produced by national household surveys, including 18,978 live births and 27 candidate predictors, are of the type analyzed in this study, and are well suited to modern predictive modeling techniques beyond the multivariable or multilevel logistic regression that has traditionally dominated this literature. Second, machine learning techniques like random forest and gradient boosting have been used more and more in clinical, facility based neonatal mortality data, usually in conjunction with logistic regression as a benchmark, but have not been widely used or tested in nationally representative household survey data, like the data produced by DHS programs, and with the rigor required in terms of class balance. This study directly tackles that gap, by testing a large birth-history extract against both classical and machine learning methods of predicting neonatal mortality under a severe class imbalance inherent in any rare, population-level health outcome.

In Kenya, and in much of sub-Saharan Africa, neonatal mortality rates are far greater than the SDG goals and conventional (multivariable or multilevel) logistic regression has been used extensively with DHS birth-history data to describe the demographic and socioeconomic factors associated with neonatal death [3]. This body of work has been helpful in determining those factors that are statistically significant and identifying adjusted odds ratios, but there are three related limitations that need to be addressed by the present study.

Second, in conventional regression-based studies, models are not typically compared with more recent predictive algorithms that have the ability to model non-linear relationships and interactions among predictors without the analyst needing to specify them a priori, so it is not known, for data on neonatal mortality from households surveys as opposed to clinical settings, whether tree-based ensembles like random forest and gradient boosting meaningfully outperform logistic regression or whether, as some clinical comparisons have suggested, a well-specified logistic regression can be competitive.

Second, the neonatal death prevalence rate characteristic of most population-based survey data (2.22% neonatal death prevalence, as compared to higher-risk cohorts of neonates in clinical NICUs, where the prevalence of neonatal deaths is often in the range of 10-100%) results in a large class imbalance that is not explicitly considered in most existing DHS-based regression studies. Fitting without balancing the data for the overall accuracy of the model might result in a substantial underestimation of the proportion of true deaths that is not identified by the model, and thus in a low performance rating of the model, while the overall accuracy appears good.

Third, and most importantly when considering the fit of any model to this type of birth-history data, several common survey variables are mechanically related to the outcome of the neonatal death itself (most importantly, the number of surviving children under 5 years of age recorded at the time of the survey). A model of this type can seem very predictive, and if they are included without critical examination, they may be a definitional artefact and not a true actionable risk factor. This leads to credible risk that the results of well-conceived studies may mislead the research community and policy-makers as to what is actually predictive.

This study thus sought to address the problem of comparing classical and machine learning classifiers in the context of a rare event outcome using survey-based neonatal mortality data, the evaluation and comparison of explicit strategies for handling data imbalance in that comparison, and the scrutiny of the set of candidate predictors for outcome-endogenous predictors that might impact predictive performance as well as substantive interpretation of feature importance. Unaddressed, these gaps will reduce the confidence and usefulness of predictive models based on neonatal outcomes from DHS.

Literature Review

Maternal age at birth, maternal education, household wealth, parity, birth interval, multiple pregnancy, birth order and place of delivery have been consistently shown to be predictive of early neonatal death through pooled analyses of Demographic and Health Survey data across multiple sub-Saharan African countries, often using multilevel logistic regression to account for the hierarchical clustering of births within mothers, households, and survey clusters [4]. Evidence review studies in the low and middle-income countries have also identified short intervals between births (specifically less than 18-24 months), extremes of maternal age, low household food security, low birth weight,

unimproved sanitation, rural living and low human development and literacy rates as consistent correlates of infant and neonatal mortality, with significantly different pooled risk estimates for regions; the highest are found for West Africa and South Asia.

The determinants of neonatal mortality were analysed across successive rounds of Kenya Demographic and Health Survey. At the bivariate level, maternal work status, household wealth index, birth weight, birth interval and perceived size at birth were significant factors associated with neonatal mortality, and maternal work status, wealth index, birth weight and size of the baby remained significant in adjusted regression models in the analysis of the 2008/09 KDHS round [5]. A subsequent study of 18,951 births with 356 neonatal deaths in a reference period very similar to that used in this study (1998–2013) found that low birth weight was linked to over three times the odds of a neonatal death (18,978 births and 422 neonatal deaths analyzed here). Geographical and temporal variation in neonatal mortality by county has been modeled, and additional covariates have been identified using 9-county subnational analyses that combine KDHS and routine health information system data, and are modeled using multilevel logistic regression. A machine learning analysis of the 2022 KDHS, explicitly recommending the inclusion of family planning programs to lower infant mortality rates by decreasing the number of young children per house and the number of births per woman in the last 5 years, highlighted short birth intervals, multiple births, second birth order, poor and middle wealth quintiles, unimproved toilet facilities, and extremes of maternal age as significant predictors; a recommendation that, when combined with the endogeneity concern, underscores the need for careful distinction between variables that are actually modifiable risk factors and those that are a definitional consequence of the outcome being studied [6].

The application of machine learning classifiers to the prediction of neonatal and infant mortality has been gaining momentum, with most of the studies using clinical, facility-based (usually NICU) data in which event rates are higher and strong clinical predictors such as birth weight, gestational age and Apgar score are available [7,8]. Five machine learning models were compared in a population based study that used 33 predictors ranging from birth facility to prenatal care, pregnancy history, and newborn characteristics, to predict infant death, with a gradient boosting model (XGBoost) providing the highest discrimination (area under the ROC curve 0.93, average precision 0.55), as compared with random forest (AUC 0.88); a simple model of just four of these highly predictive clinical variables (gestational age, birth weight, 5-minute Apgar score, and number of prenatal visits) had only modestly lower discrimination (AUC 0.91) than the full 33-variable model (AUC 0.93), indicating that predictive power tends to be highly concentrated among a small number of good clinical variables in that setting.

Five teams of neonatologists used logistic regression, CatBoost, random forest, XGBoost, and neural networks to model the same dataset of more than 6,000 NICU admissions, and a modelled audience vote before the models were released showed that an audience-favored CNN was not the best discriminator on the held-out test set, but that logistic regression is a strong, interpretable benchmark that does not need to be as complex [2]. Similar results were seen in a comparison of logistic regression with artificial neural network, random forest and support vector machines models, using 11 neonatal and maternal clinical predictors in a study of 7,472 very-low-birth-weight infants in a nationwide Korean neonatal network, which again underscored the importance of the algorithm but in this study, discrimination was similar between the artificial neural networks and the random forest models and worse for the support vector machines models. Several machine learning models, such as random forest and variants of neural networks, were also compared with logistic regression in predicting bronchopulmonary dysplasia or death, and similarly exhibited suboptimal calibration (even if they had similar discrimination), a result that is pertinent to the calibration analysis of the current study [9].

A systematic review of studies using machine learning to tackle neonatal mortality revealed that most frequently used algorithms included neural networks, random forest and logistic regression, and reported varying levels of performance, influenced by the specific characteristics of the data used, including sample size, how missing data were handled, and data collection period and context [10,2]. None of the studies reviewed, however, covered a class imbalance as severe as the 2.22% incidence of the population sample in the present study; most of the studies were conducted in the NICU, where the populations admitted are generally at higher risk, with significantly higher mortality rates, and so mortality is a less salient concern that motivates much of the methodology of the present study.

Materials and Methods

The study uses a quantitative, cross-sectional, secondary analysis of retrospective birth histories in a national representative, household and maternal health survey of Demographic and Health Survey (DHS) type. The design includes two complementary statistical approaches: (a) inferential analysis (bivariate chi-square tests of association and analytic (Wald) standard errors) with the aim of testing hypotheses on which predictors are significantly associated with neonatal mortality; and (b) predictive modeling and algorithm comparison (a train/test evaluation protocol, with logistic regression versus two machine learning classifiers, under three classification problems scenarios of the presence of class imbalance). This integrated design enables the study to investigate both types of research questions (explanatory and predictive) and to examine how well and in what way survey information can accurately predict neonatal death, while simultaneously providing answers to the explanatory research questions of what factors are associated with neonatal death and how strongly.

The study adopts a pre-coded, anonymized birth-history extract from 18,978 live births, a binary predictor for neonatal survival, and 27 potential predictor variables. The structure of the extract, such as the treatment of its coding scheme, the category definitions and sample size, follows the standard format of DHS birth-recode construction where each row is the birth history of 1 live birth for the 5 years prior to the survey. The observed neonatal death count and prevalence in this extract (422 deaths, 2.22%) are similar to published analyses of the 2014 KDHS birth recode (18,951 births, 356 deaths, 1.88%), reinforcing the face validity of this extract as a true and authentic national

data source, rather than a synthetic or convenience sample. No additional primary data was collected; this is secondary statistical analysis of available and identified survey data. The minimum sample size n to estimate this neonatal death proportion p within a margin of error e at 95% confidence can be checked against the Cochran formula:

$$n = \frac{z^2 p(1-p)}{e^2} \quad (1)$$

where Z = 1.96 for a 95% confidence level. Additionally, it is found that the sample size n_0 = 3,300 is sufficiently large to estimate the neonatal mortality rate accurately, and that those 18,978 births can be appropriately subdivided for a train/test partition, as shown by the calculation of n_0. The outcome variable is a binary variable that takes the value 1 if the index birth dies within the first 28 days after birth and 0 if it survives and is interviewed or is outside the neonatal period, whichever comes first:

$$Y_i \sim \text{Bernoulli}(p_i) = \begin{cases} Y_i = 1 \, if \; neonate \, i \, died \, within \, 28 \, days \, of \, birth \\ Y_i = 0 \, otherwise \end{cases} \quad (2)$$

The status variable was recoded from the source variable "status" (coded 1 = alive, 0 = dead) to ensure that 1 always indicates the event of interest (death) when used throughout the modelling process, consistent with the usual coding convention used for binary classification.

Predictor variables

A total of 27 candidate predictors were selected and grouped into four conceptual categories following the theoretical framework proposed in the literature by Mosley-Chen.

Table 1: Predictor Variables Grouped by Conceptual Category

Group	Variables
Reproductive history (proximate/ maternal factors)	Age at first birth; birth order; preceding birth interval; pregnancy losses; ever had a terminated pregnancy; multiple birth status; number of births in the last five years; gestation length (months)
Household & socioeconomic context (distal factors)	Wealth index quintile; household floor material; household toilet facility; drinking water source; cooking fuel; radio listening frequency; place of residence (urban/rural); province
Maternal & household-head characteristics	Mother's age; mother's education; mother's occupation status; union status; sex of household head; religion
Child & birth characteristics	Sex of child; day of birth (within month); number of living children under five in the household; mosquito net use

All predictors were coded in pre-coded categorical or ordinal integer form (coded 0–3 for no education through higher education for mother's education or coded 1–5 for poorest through richest for the wealth index) as is common with DHS variables. No cases would be lost due to listwise deletion or missing data, as there were no missing values in the predictor variable(s) used in the extract.

Data Preparation

The variable names were transformed into descriptive labels appropriate for the analyses and the outcome variable was recoded. Two parallel feature representations have been constructed using the same data. The nominal categorical predictor had c levels: (c - 1) binary indicator (dummy) variables were used (with the omitted reference level j = 0):

$$\begin{cases} X_{ji} = 1 \, if \; \text{var} \, iable \, X \, takes \, level \; j \; for \, observayion \, i \\ X_{ji} = 0 \, otherwise \end{cases}$$

j = 1,…,c - 1

This type of reference category coding is useful in avoiding the dummy variable problem (perfect collinearity with the intercept) and in estimating a log-odds effect of each non-reference category in comparison to the reference category. As is typical for tree-based classifiers, the random forest model and the gradient boosting model did not use the dummy coding of the features, but instead kept the original ordinal/integer coding of the features (as opposed to dummy coding) because these classifiers split on the raw values of the features using threshold cuts. There were no continuous variables that needed to be scaled and all three algorithms did not need input scaling to be estimated correctly because all predictors were categorical or small range ordinal counts.

Data Partitioning

Stratified sampling was used to randomly divide the entire sample of N = 18,978 births into a training set and a held-out test set, with the death prevalence π = 422/18,978 = 0.0222 being preserved (within stratified sampling precision) in both sets.

n_{train} = 0.75 N=14,233,n_{test} = 0.25 N=4,745

The death rate was found to be consistent across the partitions as evidenced by the number of deaths 422 in the sample, which was split into two partitions: 316 and 106 deaths, respectively, within 0.75 × 422 = 316.5 and 0.25 × 422 = 105.5 of the sample size, respectively. All imbalance-handling strategies were used only on the training partition and the test partition was not resampled, reweighted, or otherwise changed from its original setup, so that reported performance statistics are based on the true, real-world class distribution. A single fixed random seed was used for the entire process of partition and all subsequent model fits to ensure exactly the same partition and model fits every time.

Logistic regression

Logistic regression models the conditional probability that neonate i dies, $p_i = P(Y_i = 1 \mid X_i)$, via the logistic (sigmoid) link function applied to a linear predictor:

Equivalently, the model is linear in the log-odds (logit) of death:

$$\log it(p_i) = In\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \quad (3)$$

$$pi = \frac{1}{(1+\exp(-n_i))}, n_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \quad (4)$$

where $X_{1i},\ldots,X_{ki}$ are the dummy-coded predictor values for observation i. The coefficient vector $\beta=\beta_0,\beta_1,\ldots,\beta_k$ is estimated by maximum likelihood: since each Y_i is Bernoulli(p_i), the likelihood of the observed training data is

$$L(\beta)=\prod_{i=1}^{n} p_i^{Yi}(1-p_i)^{1-Y_i} \quad (4)$$

$$l(\beta)=\sum_{i=1}^{n}[Y_i \ln(p_i)+(1-Y_i)(1-p_i)] \quad (5)$$

and β is obtained by maximizing the log-likelihood, numerically, via iteratively reweighted least squares (equivalently Newton-Raphson on l(β)), since no closed-form solution exists for logistic MLE. The odds ratio for each prediction is interpreted as the multiplicative change in the odds of a neonatal death that would result from a one-unit (or 1 level) change in predictor j while holding all other predictors constant. Logistic regression was used in two ways in this study: as one of the three predictive classifiers tested on the test set, and as a baseline model without weights in the unweighted form to which the odds ratios, standard errors, and hypothesis tests of individual predictors were calculated.

Random Forest

Random forest is an ensemble learning method that builds B classification trees, independently trained on a different bootstrap sample of the training data, and using only a random sample of predictors during each candidate split. Each individual tree splits a node by selecting the predictor and threshold that cause the largest decrease in Gini impurity, which for a node t with observed class proportions p_0 (t) (survival) and p_1 (t) (death) is defined as:

$$Gini(t)=1-p_0(t)^2-p_1(t)^2=2p_o(t)\,p_i(t) \quad (6)$$

A candidate split into child nodes t_L and t_R is chosen to maximize the impurity reduction:

$$\Delta Gini = Gini(t)-\left[\frac{n_L}{n_t}Gini(t_L)+\frac{n_R}{n_t}Gini(t_R)\right] \quad (6.1)$$

where n_t is the number of observations in the parent node, n_L the number of observations in the left node, and n_R the number of observations in the right node.

The forest generates a predicted probability for each observation (predicted probability $p_i^{(b)}$, a bagging, or bootstrap-aggregating, procedure that will benefit the forest by bringing down the variance of any single tree, while preserving the ability of trees to capture non-linear relationships or interactions without prespecification:

$$\widehat{p}_i^{RF}=\frac{1}{B}\sum_{b=1}^{B}\widehat{p}_i^{(b)} \quad (7)$$

In this study, a random forest of B = 500 trees with a maximum depth of 8 was fit to the numeric-coded training data, with each split considering a random subset of √k candidate predictors, following the standard default for classification forests.

Gradient boosting

Gradient boosting is an ensemble approach that trains a series of M shallow decision trees, each tree being trained to the negative gradient (functional residual) of a given loss function applied to the ensemble's prediction at the time of training. In binary classification, the ensemble is additive with regard to the log-odds:

$$F_M(x)=F_0(x)+\eta\sum_{m=1}^{M}h_m(x) \quad (8)$$

where F_0 (x) is an initial constant (log-odds) prediction, h_m (x) is the m-th shallow regression tree, and η is the learning rate that shrinks each tree's contribution to guard against overfitting. Each tree h_m is fit to the negative gradient of the binomial deviance (log) loss with respect to the current predictions F_{M-1} (x):

$$L(Y,F)=\sum_{i=1}^{n}\left[Y_i ln\left(1+exp\left(-F(x_i)\right)\right)+(1-Y_i)ln\left(1+exp\left(F(x_i)\right)\right)\right] \quad (9)$$

The ensemble is updated as $F(x_i)=F_{M-1}$ (x)+ηh_m (x) for m = 1,…,M and then the sigmoid transform Equation (5) is applied to the final $F(x_i)$. The numeric-coded training data was used to train a gradient boosting classifier with a histogram, with 300 boosting iterations, tree depth of 4, and learning rate η = 0.05; these hyperparameters are standard default settings for shallow, well-regularized boosting, where they trade some of the learning speed of a single tree for a higher number of trees to mitigate the risk of overfitting in a moderate-sized tabular dataset.

Handling Class Imbalance

To evaluate the impact of imbalance-handling strategy, each of the three algorithms was fitted under three conditions, all operating on the training partition.

Baseline (unadjusted)

This algorithm is trained on observed data without any reweighting or resampling and is used as a baseline for the cost of imbalance-handling; the loss function assumes that a misclassified death is as costly as a misclassified survival, even though deaths are about 44 times less likely.

Class-weighting (cost-sensitive learning)

The weights assigned to each training observation are inversely proportional to the frequency of the class to which the observation belongs and the classes contribute equally to the overall weighted loss:

$$w_k=\frac{n_{train}}{2.n_k},\quad k\in\left\{0\left(survival\right),1\left(death\right)\right\}$$

With n_{train} = 14,233, n_1 = 316 deaths, and n_0 = 13,917 survivals, this yields w_1 = 14,233/(2×316) ≈ 22.52 for deaths and w_0 = 14,233/(2×13,917) ≈ 0.51 for survivals; an effective weighting ratio of approximately 44:1 in favor of the minority class,

exactly offsetting the observed class imbalance. For logistic regression and random forest, class weights enter the weighted log-likelihood/impurity criterion directly (e.g., replacing $\ell(\beta)$ with $\sum_i w_{Yi}\,[Y_i \ln(p_i)+(1-Y_i)\ln(1-p_i)]$; for gradient boosting, the same weights w_i enter as per-observation multipliers on the loss.

Synthetic minority oversampling (SMOTE)

Training observations for synthetic minority-class (death) are generated through interpolation in the feature space, as done in [11]. For each minority observation x_i requiring synthesis, one of its k = 5 nearest minority-class neighbors, $x_{i(nn)}$ is selected at random, and a new synthetic observation is constructed as:

$$x_{new}=x_i+\delta.(x_{i(nn)}-x_i) \quad (10)$$

The process is repeated until parity is reached in the training partition but not in the testing partition:

$n_{needed}= n_0 - n_1 = 13{,}917 - 316 = 13{,}601$ synthetic observations where n_{needed} is the number of synthetic minority-class observation required, n_0 is the survival observations and n_1 is the number of deaths observations.

The algorithms were then refitted on an augmented training set of 27,834 observations (13,917 in each class). This yielded a total of nine model by imbalance-strategy combinations (three algorithms × three strategies), each of which was tested exactly the same way on the untouched held-out test set [12].

Statistical Inference for Logistic Regression

Analytic standard errors were computed for the unweighted baseline logistic regression as the inverse of the observed Fisher information matrix at the maximum-likelihood estimates $\hat{\beta}$:

$$\mathrm{Var}(\hat{\beta})=(X^T WX)' \quad (11)$$
$$W=\mathrm{diag}[p_i (1-p_i)] \quad (12)$$

where X is the n × (k+1) training design matrix (including an intercept column) and W is the n × n diagonal matrix of fitted Bernoulli variances. The square root of the diagonal elements of Var(β) are the standard errors $SE(\beta_j)$. The Wald test statistic for testing the null hypothesis H_0: $\beta_j = 0$ is:

$$Z_j = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)} \quad (13)$$

$$\text{p-value}=2[1-\phi(|Z\_j|)] \quad (14)$$

whereas $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution, Z_j is Wald Statistic test. The odds ratio and 95% Wald CI are:

$OR= \exp(\beta_j)$,
and
$95\%\ CI = \exp[\ \beta_j \pm 1.96 \cdot SE(\beta_j)]$

These quantities (Z_j, p-values, OR_j, and confidence intervals) were calculated for each of the predictors and used to inform the hypothesis

Bivariate Screening

Each candidate predictor was then cross-tabulated with the outcome and tested for independence of the predictor from the outcome using Pearson's chi-square statistic prior to multivariable modeling. Prior to multivariable modeling, each candidate predictor was cross tabulated with the outcome and tested for independence with Pearson's chi-square statistic. A predictor with r levels against the 2-level outcome, data for which are observed in O_{ij} cells and are expected under the null hypothesis of independence under E_{ij}= (row total × column total)/n.

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{2} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (15)$$

The resulting statistic is compared to a χ^2 distribution with (r − 1) degrees of freedom to get the p-value listed for each predictor. The multivariable logistic regression and machine learning models that are multivariable models that adjust for all predictors at once; they follow this bivariate pattern.

Model Evaluation Metrics

The performance was evaluated using a 0.5 decision threshold as is common with the class imbalance literature presented in Section 2.5, and using a 2 × 2 confusion matrix of predicted versus observed class, where TP (true positive / correctly predicted death), FN (false negative), FP (false positive), and TN (true negative):

Sensitivity (Recall) = TP / (TP + FN)
Specificity = TN / (TN + FP)
Precision = TP / (TP + FP)

$$F1=\frac{2\cdot\left(Precision\cdot Sensitivity\right)}{\left(Precision+Sensitivity\right)}$$

Two threshold-independent discrimination metrics were also computed. The receiver operating characteristic (ROC) curve is a graph of true positive rate (sensitivity) versus false positive rate (1 − specificity) for all possible classification thresholds T; the area under this curve is equal to the probability that a randomly selected death will be classified as death with a higher threshold than a randomly selected survival:

$$ROC\ ?\ AUC=P(\hat{p}_{death}>\hat{p}_{survival})=\int_0^1 TPR_{(T)}\left|dFPR_{(T)}\right| \quad (16)$$

The precision-recall curve plots precision versus recall (sensitivity) for every threshold; in a severe class imbalance setting, the area of the precision-recall curve, the PR-AUC (average precision), is not inflated by the large number of true negatives in a rare event confusion matrix unlike ROC-AUC:

$$PR-AUC=\sum_k \left(R_k - R_{k-1}\right)P_k \quad (17)$$

The sum over thresholds k ordered by decreasing recall R_k, where P_k is the precision at threshold k. Lastly, the probabilistic calibration was described by the Brier score, which is the mean squared difference between the predicted probabilities ($\hat{p}_i$) and actual outcomes (Y_i):

$$Brier = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i - Y_i\right)^2 \quad (18)$$

The lower the value, the better the probabilistic predictions are calibrated; this was also displayed in a plot showing the mean probability of the event (over deciles of predicted risk) against the frequency of the event.

Feature Importance

In logistic regression, variable importance is measured directly by the Wald-based odds ratios, confidence intervals and p-values. Permutation importance was used for the random forest model and the gradient boosting model, since the coefficients from these models cannot be interpreted directly. For predictor j, its values across the n_{test} held-out test observations are randomly permuted (shuffled) R = 20 times; for each permutation r, the fitted model's ROC-AUC is recomputed on the permuted test set, $AUC_j^{(r)}$, and importance is the mean decrease relative to the unpermuted (baseline) test ROC-AUC, AUC_0:

$$Importance_j = \frac{1}{R}\sum_{r=1}^{R}\left(AUC_0 - AUC_j^{(r)}\right) \quad (19)$$

Importance estimates are accompanied by the associated standard deviation as shown around each value as an error bar, across the R permutations. It was adopted as the preferred method over impurity based (Gini) importance, which is intrinsic to the tree-based models, as permutation importance is calculated on the held-out data and is not systematically favoring high cardinality categorical predictors.

Robustness Analysis

To address the endogeneity concern mentioned in Sections 1.2 and 2.6, a dedicated robustness check was performed in which the best performing model configuration (class-weighted gradient boosting) was refitted using the same train/test split but omitting X_{u5}:

$$F_M^{(restricted)}(x) = F_M(X \backslash X_{u5}) \quad (20)$$

holding every other modeling choice (algorithm, hyperparameters, imbalance strategy, train/test partition, and random seed) constant. $F_M(.)$ Is the original predictive model M, $F_M^{(restricted)}(x)$ is the predictions from the model fitted without excluding predictor, X_{u5} is the predictor vector representing the number of living children under 5 years, $X \backslash X_{u5}$ is predictor vector after removing X_{u5}. Comparing ROC-AUC, PR-AUC, and sensitivity with and without X_{u5} gives a direct, quantitative measure of the extent to which the apparent predictive power of the model comes from this single mechanically-outcome-linked predictor, and provides a more defensible estimate of the predictive ceiling that can be achieved from the remaining, exogenous predictors.

Software and Tools

Data preparation, analysis and modelling were done in Python 3, using pandas, NumPy for data handling, SciPy for performing chi-square tests and Wald and scikit-learn (version 1.8) for train/test partition, logistic regression, random forest, gradient boosting and permutation importance. To preserve the complete analytical pipeline and ensure its reproducibility, the SMOTE-style synthetic oversampling and Wald standard-error routines have been implemented directly within scikit-learn's nearest-neighbor and linear-algebra utilities, respectively. Figures were produced using Matplotlib.

Ethical Considerations

This study relies on a secondary and de-identified data set, which does not include personally identifiable data; no primary data collection with human subjects was conducted. The use of the data for statistical and methodological research is regarded as presenting minimal risk, and is considered to be a standard form of secondary analysis of de-identified survey extracts of this type. However, analysts using the microdata provided by this original survey program, for other research purposes than methodological studies of this nature are expected to observe the registration and usage conditions of the data provider.

Results and Discussions

The empirical findings of this study, based on the methodology, is summarized in this chapter, applied to the study sample of 18,978 births. The Pearson chi-square statistic and p-value for each of the 27 candidate predictors, versus the outcome (neonatal death), is reported in Table 4.1. At the 5% level, six variables were associated with the outcome: number of living children under five, length of gestation, multiple birth status, number of births in last five years, preceding birth interval, and previous terminated pregnancy. Union status, age at first birth, mother's education and pregnancy losses were associated at a more liberal 10% level. In this data, no bivariate associations were found for any household-level infrastructure variables (toilet, drinking water, cooking fuel, household floor) or household-head sex and neonatal death.

Table 2: Bivariate Chi-Square Tests of Association with Neonatal Death, Sorted By P-Value

Variable	Levels	x^2	p-value
Number of living children under 5	6	953.22	< 0.001
Gestation length (months)	6	710.64	< 0.001
Multiple birth status	4	212.44	< 0.001
Number of births, last 5 years	5	137.33	< 0.001
Preceding birth interval	5	36.13	< 0.001
Ever had a terminated pregnancy	2	15.01	< 0.001
Pregnancy losses	4	15.85	0.001
Age at first birth	4	9.13	0.028
Mother's education	4	8.71	0.033
Union status	3	6.75	0.034
Radio listening frequency	3	5.58	0.062
Mother's age	7	10.51	0.105
Sex of child	2	1.71	0.190
Day of birth	6	6.58	0.254
Place of residence	2	1.03	0.309
Province	8	8.19	0.316
Age of household head	4	2.86	0.414

Wealth index	5	2.90	0.575
Birth order	7	4.76	0.575
Household toilet	2	0.29	0.592
Religion	4	1.61	0.657
Mother's occupation	2	0.10	0.754
Cooking fuel	3	0.55	0.759
Mosquito net use	2	0.09	0.761
Drinking water source	2	0.01	0.935
Household floor	2	0.01	0.939
Sex of household head	2	0.00	0.990

Inferential Logistic Regression

The statistically significant predictors ($p < 0.05$) of the baseline (unweighted) logistic regression are presented in Table 4.2 and the standard errors, Z-statistics, odds ratios, and the 95% confidence intervals. Three variables were significant with very large effects and two more variables were significant with more modest effects.

Table 3: Statistically Significant Predictors from Baseline Logistic Regression (Wald Test).

Predictor	Odds Ratio	95% CI	p-value
Number of living children under 5	0.097	0.079 – 0.119	< 0.001
Gestation length (months)	0.344	0.280 – 0.423	< 0.001
Number of births, last 5 years	10.140	8.103 – 12.690	< 0.001
Sex of household head (female)	0.608	0.435 – 0.850	0.004
Union status (currently in union)	0.579	0.344 – 0.977	0.040

The odds ratio for number of living children under 5 (OR = 0.097) means that, with the other predictors of the index birth's death fixed, the chance of the index birth's death decreases by 90.3% for every increase in the number of living children under 5 in the household; and this is not a purely causal association. The odds ratio for number of births in last five years (OR = 10.140) suggests that the higher the recent parity (number of births in past 5 years), the greater the odds of death, consistent with the short-birth-interval and high parity risk pathway. The gestation length was associated with an expected protective direction (OR = 0.344 for each additional band, consistent with the established clinical literature of the relationship between preterm birth and neonatal survival.

Predictive Performance Comparison

The three imbalance-handling strategies are then compared with each of the three algorithms of and the resulting comparison is reported in Table 4.3, which uses the same held-out test set (n = 4,745; 106 deaths).

There are three patterns to note. First, ROC-AUC is very stable across algorithms and imbalance strategies, ranging from 0.898 to 0.921, suggesting that all three algorithms discriminate in the same way across the entire range of births. Second, the unweighted baseline random forest (0.371) and unweighted gradient boosting (0.347) have the highest PR-AUC, but their sensitivity is very low (6.6% and 12.3% respectively): this is consistent with the concern about the accuracy paradox mentioned above, as these models classify nearly none of the true deaths as neonatal deaths, despite their high precision when they do make a classification. Third, class-weighted and SMOTE dramatically move this tradeoff toward sensitivity: class-weighted gradient boosting correctly classifies 83.0% of deaths in the test set (vs. 12.3% for unweighted), has the best overall ROC-AUC (0.921) of any model, and produces a competitive PR-AUC (0.311).).

Table 4: Test-Set Performance for All Nine Model × Imbalance-Handling Combinations

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity	Precision	F1	Brier
Logistic Regression- Baseline	0.915	0.302	0.160	0.996	0.472	0.239	0.019
Logistic Regression – Weighted	0.920	0.274	0.858	0.891	0.153	0.260	0.097
Logistic Regression – SMOTE	0.919	0.281	0.840	0.908	0.173	0.287	0.083
Random Forest – Baseline	0.919	0.371	0.066	1.000	0.875	0.123	0.017
Random Forest – Weighted	0.898	0.267	0.481	0.964	0.232	0.313	0.067
Random Forest – SMOTE	0.911	0.270	0.283	0.984	0.286	0.284	0.033
Gradient Boosting – Baseline	0.916	0.347	0.123	0.998	0.591	0.203	0.017
Gradient Boosting – Weighted	0.921	0.311	0.830	0.913	0.179	0.294	0.076
Gradient Boosting – SMOTE	0.919	0.312	0.179	0.996	0.500	0.264	0.018

There are three patterns to note. First, ROC-AUC is very stable across algorithms and imbalance strategies, ranging from 0.898 to 0.921, suggesting that all three algorithms discriminate in the same way across the entire range of births. Second, the unweighted baseline random forest (0.371) and unweighted gradient boosting (0.347) have the highest PR-AUC, but their sensitivity is very low (6.6% and 12.3% respectively): this is consistent with the concern about the accuracy paradox mentioned above, as these models classify nearly none of the true deaths as neonatal deaths, despite their high precision when they do make a classification. Third, class-weighted and SMOTE dramatically move this tradeoff toward sensitivity: class-weighted gradient boosting correctly classifies 83.0% of deaths in the test set (vs. 12.3% for unweighted), has the best overall ROC-AUC (0.921) of any model, and produces a competitive PR-AUC (0.311).).

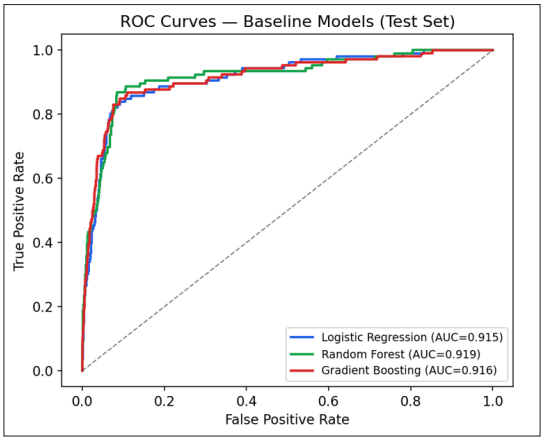

Figure 1: ROC curves for the three baseline (unadjusted) classifiers on the held-out test set.

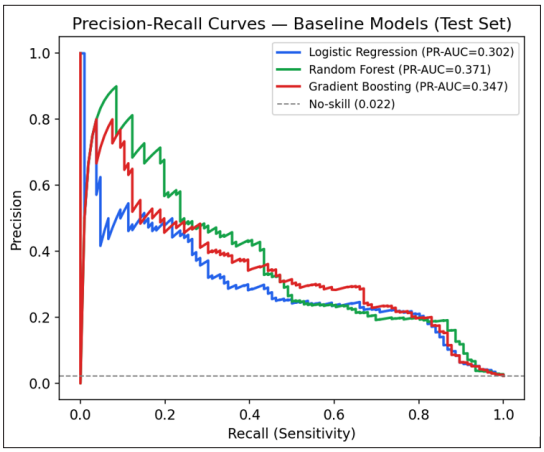

Figure 2: Precision-recall curves for the three baseline classifiers, with the no-skill (prevalence) baseline shown for reference.

Effect of Imbalance-Handling Strategy

Figure 4.3 summarizes how sensitivity and PR-AUC change across the three imbalance-handling strategies within each algorithm. Both class-weighting and SMOTE produce large, consistent sensitivity gains for all three algorithms relative to their unadjusted baselines, confirming that the choice of imbalance-handling strategy has a materially larger effect on the ability to flag at-risk births than the choice of algorithm itself, echoing the theoretical expectation that accuracy-oriented loss functions default toward the majority class absent explicit correction.

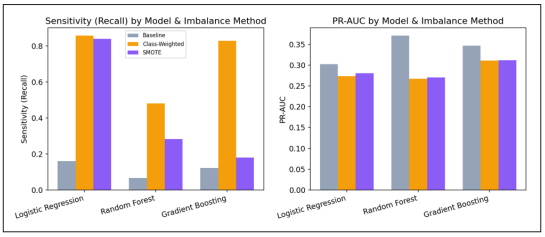

Figure 3: Sensitivity and PR-AUC by algorithm and imbalance-handling method.

Calibration

The sensitivity and the PR-AUC for the three imbalance-handling strategies are summarized in Figure 4.3 for each algorithm. The large size and consistency of the sensitivity increase for all three algorithms over their unweighted equivalents confirms that the algorithm's capability to find at-risk births is more strongly influenced by the way in which the classes are imbalanced than by the choice of algorithm, as predicted theoretically that an accuracy-oriented loss function would by default favor the majority class.

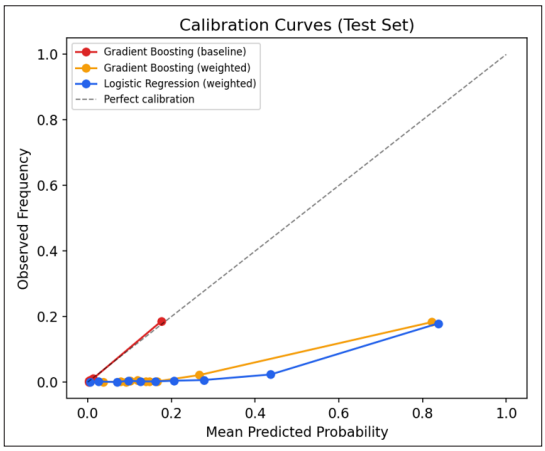

Figure 4: Calibration curves (deciles of predicted probability) for baseline and class-weighted models.

Feature Importance in Machine Learning Models

The mean reduction in ROC-AUC for the class-weighted gradient boosting model (Figure 4.5) confirms that three variables are most important: number of children alive under five, number of births in the last five years and gestation length, accounting for a significant proportion of the model's power, and all other variables have a contribution less than 0.5 percentage point individually. This is not a random forest model effect, as the best predictors are listed in the same order as the random forest model.

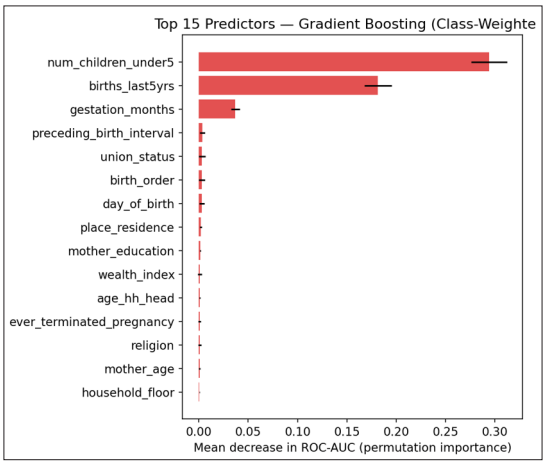

Figure 5: Permutation importance (mean decrease in test-set ROC-AUC over 20 permutations) for the top 15 predictors, class-weighted gradient boosting model.

Conclusion and Recommendations

This study showed that logistic regression, random forest and gradient boosting models had similar discrimination power for predicting neonatal mortality, with class-weighted gradient boosting having the highest level of both discrimination and sensitivity. The results indicated that dealing with the imbalanced classes had more impact on the performance of the models than

the classification algorithm, as the class weighting significantly improved the sensitivity of the models and did not introduce the complexity and potential bias of SMOTE. Living children under 5 years of age, number of births in the last 5 years and gestation length were the most important as determined by a feature importance analysis. After performing robustness analysis, however, the apparent predictive performance was greatly amplified by the inclusion of the number of living children under five years, a mechanically outcome-related variable which brought in information leakage. This predictor was successfully removed, lowering the model's ROC-AUC from 0.921 to 0.700, and emphasizing the need to screen the predictors for temporal and definitional relationships with the outcome. The study also found that proximal maternal and reproductive factors were better predictors of neonatal mortality than were distal socioeconomic factors. Going forward, predictive modelling with Demographic and Health Survey data should explicitly deal with class imbalance, measure evaluation metrics appropriate for rare events, be aware of potential predictor endogeneity, and favor class weighting over synthetic oversampling methods to increase the likelihood of model reliability, interpretability and usability in the public health decision-making process.

Author's Contribution

Erick Cheruiyot Kirui was the lead author and was responsible for the conception, study design, and implementation of the research. Stephen Wanjohi contributed to the study methodology and conducted the data analysis".

Funding Statement

This research did not receive any funding

Data Availability

The dataset utilized in this research can be provided by the author upon reasonable request.

Acknowledgement

The author would like to thank the department of Mathematics and Actuarial Science of Murang'a University of Technology for their support and guidance during the entire research process. Also, special thanks to my lovely wife and daughter Flossy and Natasha for their moral support.

Conflict of Interests

Authors declared no conflict of interest

References

1. Tamir TT, Mohammed Y, Kassie AT, Zegeye AF. Early neonatal mortality and determinants in sub-Saharan Africa: Findings from recent demographic and health survey data. PLoS ONE. 2024. 19: 0304065.
2. Sullivan BA, Moreira AG, McAdams RM, Knake LA, Husain A. Comparing machine learning techniques for neonatal mortality prediction: insights from a modeling competition. Pediatric Research. 2025. 98: 405-411.
3. Wainaina J, Gachohi J, Aluvaala, J. Outborn Neonates in Nairobi's Public Hospitals: Mortality Burden and Associated Risk Factors in a Retrospective Cohort Study. Wellcome Open Research. 2026. 11: 217
4. Asmare AA, Agmas YA. Multilevel multivariate modeling on the association between undernutrition indices of under-five children in East Africa countries: evidence from recent demographic health survey (DHS) data. BMC Nutrition. 2023. 9.
5. Fenta SM, Biresaw HB, Fentaw KD. Risk factor of neonatal mortality in Ethiopia: multilevel analysis of 2016 Demographic and Health Survey. Tropical Medicine and Health. 2021. 49.
6. Kirui EC, Luchemo E, Anapapa A. Modeling Infant Mortality Risk Factors using Logistic Regression Model and Spatial Analysis in Kenya. Asian Journal of Probability and Statistics. 2021. 13: 21-33.
7. Iqbal F, Chandra P, Khan AA, Edward S Lewis L. Prediction of mortality among neonates with sepsis in the neonatal intensive care unit: A machine learning approach. Clinical Epidemiology and Global Health. 2023. 24: 101414.
8. Mangold C, Zoretic S, Thallapureddy K, Moreira A, Chorath K. Machine learning models for predicting neonatal mortality: A systematic review. In Neonatology. 2021. 118: 394–405.
9. Zhang Z, Xiao Q, Luo J. Infant death prediction using machine learning: A population-based retrospective study. 2023.
10. Do HJ, Moon KM, Jin HS. Machine Learning Models for Predicting Mortality in 7472 Very Low Birth Weight Infants Using Data from a Nationwide Neonatal Network. Diagnostics. 2022. 12.
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. In Journal of Artificial Intelligence Research. 2002. 16.
12. Singh NS, Blanchard AK, Blencowe H, Koon AD, Boerma T. Zooming in and out: A holistic framework for research on maternal, late foetal and newborn survival and health. Health Policy and Planning. 2022. 37: 565-574.