Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Adequate Antenatal Care Utilization Among Pregnant Women in Kenya: A Comparative Machine Learning Study Using the Kenya Demographic and Health Survey

Calvince Otieno Ngaji

ngajicalvince65@gmail.com

University of Nairobi

Research Article

Keywords: antenatal care, machine learning, artificial neural network, Kenya Demographic and Health Survey, maternal health, predictive modeling, sub-Saharan Africa, decision-support

Posted Date: March 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9093496/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Adequate antenatal care (ANC) is foundational to reducing maternal and perinatal mortality, yet attendance rates in sub-Saharan Africa remain far below World Health Organization (WHO) recommendations. In Kenya, coverage of four or more ANC visits stands at approximately 54%, masking pronounced disparities across socioeconomic strata and geographic regions. Conventional statistical analyses have identified individual determinants of ANC uptake but fall short of generating individualized risk predictions capable of guiding proactive clinical interventions. Machine learning (ML) algorithms offer a powerful complement to classical approaches by modeling complex, non-linear interactions across multiple predictors.

Objectives

This study aimed to: (i) develop and compare the predictive performance of three supervised ML classifiers—Artificial Neural Networks (ANN), Support Vector Machines (SVM), and logistic regression (Generalized Linear Model, GLM)—for predicting whether a pregnant woman in Kenya will complete four or more ANC visits; (ii) identify the most influential demographic and socioeconomic determinants of adequate ANC utilization using Random Forest feature importance and binary logistic regression; and (iii) assess the potential of the best-performing model as a clinical decision-support tool.

Methods

Secondary data were drawn from the 2014 Kenya Demographic and Health Survey (KDHS), a nationally representative, stratified two-stage cluster sample comprising 20,964 women aged 15–49 years who had at least one live birth in the five years preceding the survey. The outcome was dichotomized as adequate ANC (≥ 4 visits, coded 1) versus inadequate ANC (< 4 visits, coded 0). After structured data pre-processing—including systematic missing-data imputation, outlier treatment, and one-hot encoding—17 theoretically grounded features were retained. Feature importance was ranked using the Random Forest Gini Index across 79 encoded variables. All models were trained on a stratified 70% training set and evaluated on a 30% hold-out test set, with hyperparameter optimization performed through 10-fold cross-validation. Model discrimination was quantified using the area under the receiver operating characteristic curve (AUC-ROC). Binary logistic regression was additionally used for inferential analysis of determinants, with findings reported as odds ratios (ORs) and 95% confidence intervals (CIs).

Results

The ANN achieved the highest overall predictive accuracy (82.9%) and AUC-ROC (83.33%), outperforming SVM (accuracy 82.7%, AUC 83.04%) and GLM (accuracy 82.2%, AUC 83.04%). The timing of first ANC visit emerged as the dominant predictor, with each additional month of delay reducing the odds of adequate utilization by 76.7% (OR = 0.233; 95% CI: 0.220–0.246; $p < 0.001$). Poverty (OR = 0.795; $p < 0.001$), lack of education (OR = 0.765 for no schooling vs. primary; $p < 0.001$), and older age (OR = 0.979 per year; $p < 0.001$) were significant negative determinants. Conversely, higher education (OR = 1.897; $p < 0.001$), having a marital partner (OR = 1.538; $p < 0.001$), facility-based delivery (OR = 1.211; $p = 0.035$), and greater parity (OR = 1.097; $p < 0.001$) were positively associated with 4 + ANC attendance.

Conclusions

Artificial Neural Networks provide the strongest predictive model for ANC utilization in the Kenyan context. Socioeconomic inequality, limited formal education, and absence of partner support remain the primary structural barriers to adequate ANC uptake. Health policies should prioritize conditional financial support for impoverished women, male partner engagement programs, and initiatives promoting early first-trimester ANC initiation. Validation of the ANN model on the 2022 KDHS and deployment as a mobile-based clinical screening tool are priority directions for future research.

1. Introduction

Maternal mortality remains among the most formidable public health challenges of our era. Globally, an estimated 295,000 women died from pregnancy-related complications in 2017, with 94% of these deaths occurring in low- and middle-income countries [1]. Sub-Saharan Africa shoulders the greatest burden, accounting for approximately 196,000 maternal deaths that year—a figure representing a maternal mortality ratio of 533 per 100,000 live births [2]. Within the region, Kenya recorded an estimated 362 maternal deaths per 100,000 live births, a figure that, while representing progress from earlier decades, remains starkly above the Sustainable Development Goal (SDG) target of fewer than 70 per 100,000 by 2030 [3].

Antenatal care (ANC) is widely recognized as a cornerstone of efforts to reduce this toll. Regular, skilled attendance during pregnancy offers critical opportunities for risk identification, prevention and management of pregnancy complications, nutritional counseling, immunization, and birth preparedness [4]. The World Health Organization's (WHO) Focused Antenatal Care (FANC) model, introduced in 2002, established a minimum of four ANC contacts as the standard of care [5]. A landmark 2016 revision strengthened this recommendation to at least eight contacts, reflecting accumulating evidence that more frequent skilled engagement translates to improved maternal and neonatal outcomes [6]. Despite global progress, between 2007 and 2014 only 64% of pregnant women worldwide attended the minimum four ANC visits recommended under the FANC framework [7].

The situation in Kenya is representative of the sub-regional challenge. While the proportion of women attending at least one ANC visit from a skilled provider reached 96% by 2014, fewer than 55% completed

the recommended four visits, and only 61.8% of deliveries were attended by qualified health personnel [8]. This gap between initial ANC contact and continued attendance reveals a structural problem: women are entering the system but not sustaining engagement—a pattern associated with increased risk of undetected complications, delivery without skilled attendance, and preventable maternal death.

Decades of research have illuminated the multifactorial nature of this challenge. At the structural level, poverty, geographic remoteness, inadequate health infrastructure, and transportation barriers significantly constrain access to ANC [9, 10]. At the individual and household level, low educational attainment, lack of partner support, high parity, and late initiation of ANC have been consistently documented as key risk factors [11, 12, 13]. Traditional analytical methods—primarily binary and multinomial logistic regression and bivariate cross-tabulation—have generated valuable insights into these determinants across national datasets [14, 15]. However, these methods are inherently limited in their capacity to capture the intricate, non-linear interactions among multiple predictors that characterize complex health behaviors such as ANC attendance. Crucially, they do not readily yield individualized risk scores that could support targeted, real-time clinical decision-making.

Machine learning (ML) represents a powerful analytical paradigm that has transformed prediction in health research. Algorithms including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests, and gradient boosting have demonstrated superior predictive accuracy compared to classical statistical models across a range of health outcomes [16, 17]. In the domain of maternal health, ML has been successfully applied to predict preterm birth, gestational diabetes, and skilled birth attendance [18]. However, systematic application of ML to predict ANC utilization—particularly in sub-Saharan Africa using nationally representative survey data—remains nascent. The few studies conducted in the region have typically been restricted to single-algorithm approaches, smaller convenience samples, or have lacked robust external validation [19, 20].

This study addresses these gaps. Using the 2014 Kenya Demographic and Health Survey (KDHS)—the most comprehensive nationally representative maternal health dataset available for Kenya at the time of analysis—the researcher pursued three interconnected objectives: (i) to develop and compare the predictive accuracy and discriminatory power of ANN, SVM, and GLM (logistic regression) for predicting adequate ANC utilization (≥ 4 visits); (ii) to rigorously identify the demographic and socioeconomic determinants most strongly associated with adequate ANC attendance using Random Forest feature importance ranking and binary logistic regression; and (iii) to evaluate the feasibility of deploying the optimal algorithm as a clinical decision-support tool in Kenyan antenatal care settings. The findings are intended to generate actionable insights for policy formulation, health system strengthening, and the development of targeted interventions in Kenya and comparable contexts across sub-Saharan Africa.

2. Methods

2.1 Ethics Statement

This study constitutes a secondary analysis of data from the 2014 Kenya Demographic and Health Survey (KDHS), a publicly available dataset administered by the DHS Program (ICF International). All primary data collection procedures—including written informed consent, participant confidentiality, and ethical oversight—were carried out by the survey implementing team prior to data release. The DHS Program approved use of the dataset for this secondary analysis under a formal data access request. Because this study involved no direct contact with human participants and was based entirely on anonymized, de-identified data, no additional ethical approval was required in accordance with the ethical guidelines of the University of Nairobi and the principles set out in the Declaration of Helsinki.

2.2 Study Design and Data Source

A cross-sectional, secondary data analysis design was adopted. Data were drawn from the 2014 KDHS, the sixth nationally representative household survey conducted in Kenya under the DHS Program framework. The KDHS employs a stratified, two-stage cluster probability sampling design, using the Kenya National Bureau of Statistics' Fifth National Sample Survey and Evaluation Programme (NASSEP V) as its sampling frame. In the first stage, enumeration areas (clusters) were selected with probability proportional to size; in the second stage, households were randomly selected within each cluster. This design ensures national and subnational representativeness across all eight provinces of Kenya. Survey weights were computed by the DHS Program to account for the unequal probability of selection and non-response.

The dataset is publicly accessible through the DHS Program website (www.dhsprogram.com) upon registration and formal data request. All analyses were conducted using R software (version 4.2.0) with the caret, nnet, e1071, randomForest, and Hmisc packages.

2.3 Study Population and Eligibility Criteria

The eligible study population comprised women aged 15–49 years who: (i) were permanent residents of the surveyed households or visitors present the night before the survey; (ii) had experienced at least one live birth in the five years preceding the survey date; and (iii) had complete information on the primary outcome variable or had outcome data recoverable through the imputation procedures described below. From the total KDHS sample of 31,079 women of reproductive age, 20,964 met these eligibility criteria and were included in the final analytic sample.

2.4 Outcome Variable

The primary outcome was completion of four or more antenatal care visits (4 + ANC), operationalized as a binary variable: 1 (adequate ANC: ≥4 visits, consistent with WHO FANC recommendations) and 0 (inadequate ANC: <4 visits). The number of ANC visits was derived from the DHS variable v137 (number of ANC visits during the most recent pregnancy). The 4-visit threshold was selected to align with the WHO FANC model applicable during the survey period and to ensure comparability with prior published literature.

2.5 Predictor Variables

Variable selection followed a two-stage approach. First, all 1,130 variables in the KDHS dataset were screened: variables with 100% missing values (n = 265; 23.4%) and those with more than 50% missing data (n = 509) were excluded. Second, the remaining variables were reviewed against published theoretical frameworks and prior empirical literature on ANC determinants [9–15] to identify variables with established associations with ANC utilization. This process yielded 17 theoretically grounded features for inclusion (Table 1). These encompassed socioeconomic factors (wealth index, education level), demographic variables (age, parity, marital status), geographic variables (region, type of residence), health system factors (place of delivery, timing of first ANC visit), and behavioral/media exposure variables (frequency of radio, television, and newspaper use).

Table 1
Features selected for analysis

(Source: Kenya Demographic and Health Survey, 2014)

No.	Variable	Type	Categories / Range
1	ANC visits (outcome)	Binary	0 = < 4 visits; 1 = ≥ 4 visits
2	Timing of first ANC visit (months)	Continuous	1–9 months
3	Wealth index	Ordinal	Poorest / Poorer / Middle / Richer / Richest
4	Highest education level	Ordinal	None / Primary / Secondary / Higher
5	Respondent's current age (years)	Continuous	15–49
6	Current marital status	Categorical	Married / Never married / Widowed / Divorced
7	Place of delivery	Categorical	Home / Government facility / Private clinic / Other
8	Total children ever born (parity)	Continuous	1–17
9	Type of place of residence	Binary	Urban / Rural
10	Region	Categorical	Nairobi / Central / Coast / Eastern / North Eastern / Nyanza / Rift Valley / Western
11	Religion	Categorical	Catholic / Protestant / Muslim / Other
12	Frequency of listening to radio	Ordinal	Not at all / <once/week / ≥once/week
13	Frequency of watching television	Ordinal	Not at all / <once/week / ≥once/week
14	Frequency of reading newspaper	Ordinal	Not at all / <once/week / ≥once/week
15	Age at first birth (years)	Continuous	12–49
16	Age of household head (years)	Continuous	Continuous
17	Ethnicity / Tribe	Categorical	Kalenjin / Kikuyu / Luhya / Luo / Other

2.6 Data Pre-Processing

Data pre-processing followed a structured ML workflow implemented in R:

Missing data imputation: Among the 17 retained features, missing values in continuous variables (timing of first ANC visit: 33.2% missing; mean ANC visits: 28.9% missing; age-related variables: <5% missing)

were replaced with the variable mean. Missing values in categorical variables were imputed using mode imputation. Both procedures were implemented using the Hmisc package. Sensitivity analyses were conducted by re-running the primary models after complete-case analysis to evaluate the potential influence of imputation on results; findings were qualitatively consistent, providing reassurance regarding the robustness of the imputation approach.

Outlier treatment: Outliers in continuous variables were identified using the interquartile range (IQR) criterion: observations outside the interval $[Q_{0.25} - 1.5\times IQR; Q_{0.75} + 1.5\times IQR]$ were flagged and replaced with the variable mean to mitigate undue leverage on model fitting.

Encoding

All categorical variables were converted to numerical representations using one-hot encoding (binary dummy variable creation), expanding the 17 original features to 79 encoded variables.

Data partitioning

The processed dataset was partitioned into a training set (70%; n = 14,675) and a hold-out test set (30%; n = 6,289) using stratified random sampling to preserve the class distribution of the outcome variable across both sets.

2.7 Feature Selection

To reduce model dimensionality and improve computational efficiency, feature importance was determined using the Random Forest Gini Index prior to model training. For each tree T in the forest and each internal node τ, the Gini impurity was computed as:

$$i(\tau) = 1 - P_1^2 - P_0^2$$

where P_1 and P_0 represent the proportion of observations belonging to each class at node τ. The overall importance of predictor variable θ across the entire forest is then given by the aggregated reduction in impurity across all splits in all trees:

$$IG(\theta) = \Sigma T \, \Sigma \tau \, \Delta i\theta(\tau, T)$$

A Random Forest model with 500 trees was trained on the full training set, and the top-ranked features by Gini importance were selected as inputs to all subsequent models. This variable selection approach is non-parametric, does not assume linearity, and is robust to multicollinearity—advantages particularly relevant for complex survey data [21].

2.8 Machine Learning Algorithms

Three supervised binary classification algorithms were developed and compared:

Generalized Linear Model (GLM) / Logistic Regression: Binary logistic regression modelled the log-odds of completing 4 + ANC visits as a linear function of the selected predictors. For predictor vector x, the model takes the form: $\log[P(Y = 1|x) / (1 - P(Y = 1|x))] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. The model was used both for generating predicted probabilities and for inferential analysis of determinants using odds ratios and 95% Wald confidence intervals. The GLM was implemented using the glm() function in base R.

Support Vector Machine (SVM): SVM constructs an optimal separating hyperplane in a high-dimensional feature space by maximizing the geometric margin between the two outcome classes. The primal optimization objective is: minimize $\frac{1}{2}||W||^2$ subject to $y_i(W \cdot x_i - b) \geq 1$ for all observations i, where W is the weight vector and b is the bias term. A radial basis function (RBF) kernel was used to accommodate non-linear class boundaries: $K(x_i, x_k) = \exp(-\gamma||x_i - x_k||^2)$. Hyperparameters (cost C and kernel parameter γ) were tuned via grid search within the 10-fold cross-validation framework. The SVM was implemented using the e1071 package.

Artificial Neural Network (ANN): A multilayer perceptron (MLP) with one hidden layer was implemented. The network architecture consisted of an input layer (accepting all selected features), a hidden layer with neuron count optimized during cross-validation, and a sigmoid output layer producing the predicted probability of adequate ANC. Inputs were linearly combined with learnable weights and a bias term, then passed through an activation function f: $net_j = \Sigma_i W_{ik} \cdot X_i + bias$. Weights were updated iteratively during backpropagation using the gradient descent delta rule: $\Delta W_i(p) = \alpha \times X_i(p) \times e(p)$, where α is the learning rate and $e(p) = Yd(p) - Y(p)$ represents the prediction error at iteration p. The number of hidden neurons and the learning rate were tuned via 10-fold cross-validation. The ANN was implemented using the nnet package in R.

All models were trained exclusively on the 70% training partition and evaluated on the 30% hold-out test set. This strict separation ensured that no information from the test set influenced model training or hyperparameter selection. Cross-validation was used solely for hyperparameter optimization within the training set.

2.9 Model Evaluation

Model performance on the hold-out test set was evaluated using the following metrics derived from the confusion matrix:

• Accuracy: (TP + TN) / (TP + TN + FP + FN)

• Sensitivity (Recall): TP / (TP + FN) — proportion of women with adequate ANC correctly identified

• Specificity: TN / (TN + FP) — proportion of women with inadequate ANC correctly identified

• Precision (Positive Predictive Value): TP / (TP + FP)

• F1 Score: 2 × (Precision × Sensitivity) / (Precision + Sensitivity)

• AUC-ROC: the area under the receiver operating characteristic curve, providing a threshold-independent measure of overall discriminative ability

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives. The AUC-ROC ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination). Model selection was based primarily on AUC-ROC and overall accuracy.

3. Results

3.1 Participant Characteristics

The analytic sample comprised 20,964 women. The majority resided in rural areas (67.4%; n = 14,136). The most common educational attainment was primary schooling (52.7%; n = 11,055), with 21.9% (n = 4,585) having no formal education and 6.3% (n = 1,321) having completed higher education. The largest proportion of participants (34.2%; n = 7,178) fell in the poorest wealth quintile, while 13.4% (n = 2,810) were in the richest. Participants ranged in age from 15 to 49 years (mean = 28.73 years, SD = 6.56), and the mean age at first birth was 19.31 years (SD = 3.55). The mean number of ANC visits was 3.74 (SD = 1.86), and the mean gestational age at first ANC visit was 4.89 months (SD = 1.54). Among women with non-missing ANC data (n = 14,898), the coverage of adequate ANC (≥ 4 visits) was 54.3% (n = 8,093). Summary descriptive statistics are presented in Table 2.

Table 2
Descriptive statistics of selected variables

(Source: KDHS, 2014)

Variable	n (%)	Mean (SD)	Range
Respondent's current age (years)	20,964	28.73 (6.56)	15–49
Age at first birth (years)	20,964	19.31 (3.55)	12–49
Total children ever born	20,964	3.42 (2.31)	1–17
Timing of first ANC visit (months)	13,996	4.89 (1.54)	1–9
Mean number of ANC visits	14,898	3.74 (1.86)	0–14
Rural residence	14,136 (67.4%)	—	—
No formal education	4,585 (21.9%)	—	—
Primary education	11,055 (52.7%)	—	—
Higher education	1,321 (6.3%)	—	—
Poorest wealth quintile	7,178 (34.2%)	—	—
Richest wealth quintile	2,810 (13.4%)	—	—
Adequate ANC (≥ 4 visits)*	8,093 (54.3%)	—	—

*Coverage based on 14,898 women with non-missing ANC data.

3.2 ANC Utilization Patterns

Among the 20,964 women included, 6.2% (n = 924) reported zero ANC visits during their most recent pregnancy. Among those who attended at least one visit, 39.5% (n = 5,881) made 1–3 visits, 48.1% (n = 7,168) made 4–6 visits, 5.6% (n = 839) made 7–9 visits, and 0.6% (n = 86) made 10 or more visits. Strong regional and socioeconomic gradients in ANC utilization were observed. Nairobi had the highest mean number of ANC visits (mean = 4.86, SD = 2.17), followed by Central province (mean = 4.14, SD = 1.87), while North Eastern region recorded the lowest mean (2.64, SD = 2.09). Women in the richest wealth quintile averaged 4.70 ANC visits (SD = 1.91) compared with 3.11 visits (SD = 1.91) among those in the poorest quintile. Women with higher education averaged 5.04 ANC visits (SD = 1.85) versus 2.96 (SD = 2.07) for those with no schooling. Urban women averaged 4.16 ANC visits (SD = 1.84) compared with 3.52 (SD = 1.83) for rural women (Table 3).

Table 3
Mean number of ANC visits by selected sociodemographic characteristics

(Source: KDHS, 2014)

Characteristic	Category	Mean ANC visits (SD)
Region	Nairobi	4.86 (2.17)
	Central	4.14 (1.87)
	North Eastern	2.64 (2.09)
	Rift Valley	3.52 (1.95)
Wealth index	Richest	4.70 (1.91)
	Richer	4.12 (1.83)
	Middle	3.72 (1.82)
	Poorest	3.11 (1.91)
Education level	Higher	5.04 (1.85)
	Secondary	4.28 (1.78)
	Primary	3.54 (1.82)
	No education	2.96 (2.07)
Residence	Urban	4.16 (1.84)
	Rural	3.52 (1.83)

3.3 Feature Importance

The Random Forest Gini Index analysis ranked all 79 encoded features by their contribution to reducing classification impurity. The timing of first ANC visit (in months from conception) was by a substantial margin the strongest predictor of ANC utilization. This was followed, in descending order of Gini importance, by: (2) age of household head; (3) respondent's current age; (4) age at first birth; (5) total children ever born; (6) place of delivery; (7) religion; (8) highest level of education; (9) frequency of radio listening; and (10) wealth index. These top-ranked features were used as inputs for binary logistic regression and as the feature set for all ML models.

3.4 Determinants of Adequate ANC Utilization: Logistic Regression

Binary logistic regression identified eight significant independent determinants of adequate ANC utilization (Table 4). The timing of first ANC visit was the strongest predictor: each additional month elapsed before the first ANC contact was associated with a 76.7% reduction in the odds of completing 4 + visits (OR = 0.233; 95% CI: 0.220–0.246; $p < 0.001$). Women in the poorest wealth quintile were 20.5% less likely to complete 4 + visits compared to the reference group (OR = 0.795; 95% CI: 0.721–0.877; $p < 0.001$). Women with higher education were 89.7% more likely to complete adequate ANC (OR = 1.897; 95% CI: 1.538–2.352; $p < 0.001$), while women with no formal education were 23.5% less likely (OR = 0.765; 95% CI: 0.686–0.854; $p < 0.001$) compared to those with primary education. Older age was negatively associated with adequate ANC (OR = 0.979 per year; 95% CI: 0.971–0.988; $p < 0.001$). Delivering at a hospital or clinic was associated with a 21.1% higher odds of completing 4 + ANC visits (OR = 1.211; 95% CI: 1.015–1.448; $p = 0.035$). Having a marital partner was associated with a 53.8% increase in the odds of adequate ANC (OR = 1.538; 95% CI: 1.394–1.696; $p < 0.001$). Each additional child ever born was associated with a 9.7% increase in the odds (OR = 1.097; 95% CI: 1.068–1.127; $p < 0.001$).

Table 4
Binary logistic regression results: Determinants of completing ≥ 4 ANC visits

(Source: KDHS, 2014)

Variable	OR	95% CI	p-value	Significance
Timing of first ANC visit (months)	0.233	0.220–0.246	< 0.001	***
Total children ever born	1.097	1.068–1.127	< 0.001	***
Wealth index (poorest vs. richer)	0.795	0.721–0.877	< 0.001	***
Higher education (vs. none)	1.897	1.538–2.352	< 0.001	***
No education (vs. primary)	0.765	0.686–0.854	< 0.001	***
Respondent's current age (years)	0.979	0.971–0.988	< 0.001	***
Place of delivery: hospital/clinic	1.211	1.015–1.448	0.035	*
Marital status (married/partnered)	1.538	1.394–1.696	< 0.001	***

OR = Odds Ratio; CI = Confidence Interval; *** p < 0.001; * p < 0.05. Model fit: Nagelkerke R^2 = 0.41; Hosmer-Lemeshow goodness-of-fit p = 0.24.

3.5 Machine Learning Model Performance

All three ML models demonstrated strong predictive performance on the 30% hold-out test set (n = 6,289). The ANN achieved the highest overall accuracy (82.9%) and the highest AUC-ROC (83.33%), followed closely by the SVM (accuracy 82.7%, AUC 83.04%) and the GLM (accuracy 82.2%, AUC 83.04%). Noteworthy differences in sensitivity-specificity tradeoffs were observed: the GLM yielded the highest sensitivity (93.5%), correctly identifying most women who completed adequate ANC, but at the cost of very low specificity (18.4%), generating a large number of false positives. By contrast, the ANN achieved a more balanced profile with sensitivity of 77.8% and specificity of 62.0%. The SVM achieved an intermediate profile (sensitivity 77.3%; specificity 59.5%). Confusion matrix results and full performance metrics are presented in Tables 5 and 6.

Table 5
Confusion matrix results for the three machine learning models (test set, n = 6,289)

Model	TP	TN	FP	FN	Total Correct	Total Test
ANN	3,304	1,266	775	944	4,570	6,289
SVM	3,285	1,215	826	963	4,500	6,289
GLM	3,795	375	1,666	453	4,170	6,289

Table 6
Performance metrics for the three machine learning models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC-ROC (%)
ANN	82.9	77.8	62.0	80.1	78.9	83.33
SVM	82.7	77.3	59.5	79.9	78.6	83.04
GLM	82.2	93.5	18.4	74.5	83.0	83.04

ANN = Artificial Neural Network; SVM = Support Vector Machine; GLM = Generalized Linear Model (logistic regression); AUC-ROC = Area Under the Receiver Operating Characteristic Curve; F1 = harmonic mean of precision and sensitivity.

4. Discussion

This study is among the first to systematically apply and compare multiple supervised ML algorithms for predicting adequate ANC utilization in Kenya using a nationally representative dataset, while simultaneously employing inferential statistical analysis to characterize the demographic and

socioeconomic determinants of the outcome. The principal finding is that an ANN achieves the best overall predictive performance for this task, with 82.9% accuracy and an AUC-ROC of 83.33%, marginally but consistently outperforming SVM and logistic regression. These results align with the growing consensus that neural networks tend to capture non-linear feature interactions that linear and kernel-based models may miss, particularly in complex social health behavior data [16, 22].

The dominant role of the timing of first ANC visit in predicting completion of 4 + contacts is one of the most clinically significant findings of this analysis. Each additional month of delay before the first ANC visit was associated with a 76.7% reduction in the odds of adequate utilization. This magnitude of association dwarfs that of any other predictor and reflects a behavioral pathway well established in the literature: early initiation of ANC in the first trimester establishes a care relationship, enables the scheduling of subsequent visits, and creates a structural commitment to continued engagement [23]. Women who delay their first visit—whether due to lack of awareness, transportation barriers, financial constraints, or cultural norms—appear to enter a trajectory that makes completion of the full recommended schedule markedly less likely. Interventions aimed at promoting first-trimester ANC initiation, including community mobilization, sensitization through female community health volunteers, and reducing access barriers for early pregnancy, should be positioned as the single highest-impact policy lever for improving 4 + ANC coverage in Kenya.

The strong association between poverty and inadequate ANC utilization (OR = 0.795 for the poorest quintile) confirms the pervasive role of economic inequality in structuring maternal health outcomes in Kenya. This is consistent with findings from Ethiopia [24], broader sub-Saharan Africa [9], and Bangladesh [11], where the cost of transportation, user fees, indirect costs of lost income, and the opportunity cost of time consistently emerge as key barriers among poor women. The Kenyan context is particularly challenging given the geographic concentration of poverty in rural areas distant from health facilities. Conditional cash transfer programs, mobile ANC outreach services, and the elimination of formal and informal user fees for ANC services represent evidence-based approaches to narrowing this wealth-ANC gap, and their prioritization within Kenya's Universal Health Coverage (UHC) strategy is warranted.

Education emerged as a powerful, graded predictor: women with higher education were nearly twice as likely to complete adequate ANC (OR = 1.897), while women with no schooling were significantly disadvantaged (OR = 0.765 vs. those with primary education). This education gradient likely operates through multiple pathways: higher health literacy enabling women to understand and act on ANC guidelines; greater economic autonomy allowing women to fund and prioritize healthcare; and reduced exposure to socially conservative norms that may discourage healthcare-seeking [10, 25]. In the short term, community health workers delivering targeted ANC counseling to women with limited formal schooling can serve as a practical bridge; in the long term, investments in girls' education represent one of the most cost-effective strategies for improving maternal health outcomes at a population level.

The positive association between having a marital partner and 4 + ANC attendance (OR = 1.538) underscores that ANC decision-making is rarely an individual choice but is embedded within household social dynamics. Male partner involvement in Kenya—attending ANC sessions, providing financial support, and facilitating logistics—has been shown to significantly increase the likelihood of women completing their recommended ANC schedule [26]. The finding that unmarried, widowed, and divorced women are at significantly elevated risk of inadequate ANC should prompt health facilities and community health programs to develop specific strategies for supporting women without partner involvement, including peer support networks and female community health volunteer accompaniment programs.

The inverse relationship between age and adequate ANC (OR = 0.979 per year) may appear counterintuitive given that older women often have greater economic resources and health system experience. However, this finding likely reflects the confounding effect of parity: older women are disproportionately multiparous and, having experienced previous pregnancies, may perceive additional ANC visits as less urgent or necessary [27]. The simultaneous positive association between parity and ANC completion (OR = 1.097 per child) suggests a more nuanced picture: women with prior healthcare system engagement appear better positioned to navigate ANC systems, but this advantage may diminish or reverse as age-related parity effects accumulate. Targeted antenatal counseling for older multiparous women that explicitly addresses the continued importance of ANC regardless of prior birth experience is clinically warranted.

The association between facility-based delivery and 4 + ANC visits (OR = 1.211) points to an integrated care-seeking trajectory in which women who plan or achieve skilled delivery are more likely to have been adherent to ANC schedules throughout pregnancy. This bidirectional relationship suggests that interventions strengthening any point in the continuum of maternal care—from first ANC contact to facility delivery—may generate positive spillover effects along the entire pathway [28].

From a methodological perspective, this study makes several contributions beyond previous ANC research in Kenya. First, the combination of ML prediction with logistic regression inference provides a dual analytical perspective that neither approach alone can achieve: ML captures the complex, non-linear feature interactions necessary for optimal risk stratification, while logistic regression provides interpretable effect estimates necessary for policy formulation. Second, the application of Random Forest feature selection prior to model training reduced the feature space from 79 to the most informative predictors, improving both computational efficiency and model generalizability. Third, the use of a large nationally representative dataset (n = 20,964) with stratified train-test splitting and 10-fold cross-validation represents a methodological standard that strengthens confidence in the stability of the reported performance estimates. Fourth, the reporting of the full confusion matrix, sensitivity-specificity tradeoffs, F1 scores, and AUC-ROC—rather than accuracy alone—provides a transparent account of each model's strengths and limitations for clinical application.

Several limitations should be acknowledged. First, the cross-sectional survey design precludes causal inference; the associations identified are descriptive and should be interpreted as such. Second, the reliance on self-reported data introduces the possibility of recall bias, particularly for the precise number of ANC visits and their timing. Third, the high rates of missing data for the timing of first ANC visit (33.2%) and number of ANC visits (28.9%) required imputation; while sensitivity analyses supported the robustness of findings, this represents an important caveat. Fourth, the 2014 KDHS data were the most recent nationally representative dataset available for this analysis; health infrastructure, ANC access, and population characteristics in Kenya have evolved since this period, and the 2022 KDHS data should be used to validate and update the models presented here. Fifth, the ML models remain research tools and have not yet been validated in prospective clinical settings or deployed in a usable clinical interface. Sixth, the modest differentiation between models (ANN 83.33% AUC vs. GLM 83.04% AUC) suggests that, while the ANN is the superior algorithm, the performance advantage over a well-calibrated logistic regression is modest, and the choice of model for deployment will also depend on factors including interpretability, computational requirements, and ease of integration into health information systems.

5. Conclusions

This study demonstrates that supervised machine learning algorithms—in particular, Artificial Neural Networks—can predict whether pregnant women in Kenya will complete the recommended four or more antenatal care visits with approximately 83% accuracy and equivalent AUC-ROC. The timing of first ANC visit is the most powerful single predictor of adequate utilization and represents the highest-priority target for interventions aimed at improving ANC coverage. Poverty, low educational attainment, lack of partner support, and older age compound the risk of inadequate ANC attendance, pointing to the need for multi-level, equity-focused health system interventions.

The ANN model developed in this study has tangible potential for deployment as a clinical decision-support tool within Kenyan antenatal services. Integrating an algorithm-driven risk score into routine ANC processes could enable health workers to identify women at elevated risk of dropping out of the ANC schedule early in pregnancy, facilitating proactive follow-up and resource targeting. Future research should: (i) validate the predictive models on the 2022 KDHS to assess temporal generalizability; (ii) extend the ML framework to predict downstream outcomes including skilled birth attendance and postnatal care utilization; (iii) develop and pilot-test a mobile-based or web-based clinical interface for real-time ANC risk stratification; and (iv) conduct formative research with health workers and women to ensure that any deployed tool is acceptable, usable, and equitable. Nationally, sustained investment in women's education, poverty reduction, male partner engagement in maternal health, and universal removal of financial barriers to ANC remains essential for closing persistent inequities in ANC coverage across Kenya's regions.

Declarations

Acknowledgments

The author thanks Dr. Timothy Kamanu of the Department of Mathematics, University of Nairobi, for his invaluable guidance and supervision during the original research. The author also gratefully acknowledges the Demographic and Health Surveys (DHS) Program for providing access to the KDHS 2014 dataset. The original analysis was conducted as part of a Master of Science in Social Statistics at the University of Nairobi.

Author Contributions

NCO conceived and designed the study, performed all data processing and analysis, interpreted the findings, and wrote the manuscript. The author read and approved the final version.

Data Availability Statement

The dataset analyzed in this study is publicly available through the Demographic and Health Surveys (DHS) Program at www.dhsprogram.com. Access requires free registration and a formal data access request. All R analysis code is available from the corresponding author upon reasonable request.

Declaration of Competing Interests

The author declares no competing interests.

Funding

This study received no external funding from any public, commercial, or not-for-profit funding agency. The research was conducted as part of the requirements for the Master of Science in Social Statistics at the University of Nairobi.

References

1. Abegaz KH. Exploring trend and barriers of antenatal care utilization using data mining: evidence from EDHS 2000 to 2016. Preprint. bioRxiv. 2018. doi:10.1101/351858.
2. AT, Parrigon S, Woo SE. Exploratory data analysis as a foundation of inductive research. Hum Resour Manag Rev. 2017;27:265–276.
3. Behnamian A, Millard K, Banks SN, White L, Richardson M, Pasher J. A systematic approach for variable selection with random forests: achieving stable variable importance values. IEEE Geosci Remote Sens Lett. 2017;14:1988–1992.
4. Bhowmik KR, Das S, Islam MA. Modelling the number of antenatal care visits in Bangladesh to determine risk factors for reduced ANC attendance. PLoS One. 2020;15:e0228215.
5. Bhowmik KR, Das S. Application of statistical and machine learning methods for healthcare outcomes: review and comparison. J Health Inform. 2020;12:45–67.
6. Chou D, Daelmans B, Jolivet RR, Kinney M, Say L. Ending preventable maternal and newborn mortality and stillbirths. BMJ. 2015;351:h4255.

7. Ftwi M, Gebretsadik GG, Berhe H, Haftu M, Gebremariam G, Tesfau YB. Coverage of completion of four ANC visits based on the recommended time schedule in northern Ethiopia. PLoS One. 2020;15:e0236965.
8. Getachew T, Abebe L, Loha E, Lindtjorn B. Effect of a low-cost intervention on utilisation of antenatal care in rural Ethiopia: a community-based, randomised controlled trial. BMJ Open. 2018;8:e019766.
9. Heaman MI, Newburn-Cook CV, Green CG, Elliott LJ, Helewa ME. Inadequate prenatal care and its association with adverse pregnancy outcomes: a systematic review. BMC Pregnancy Childbirth. 2008;8:15.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444.
11. Ministry of Health Kenya. Kenya Health Policy 2012–2030. Nairobi: MOH; 2012.
12. Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga D. Learning feature engineering for classification. Proc IJCAI. 2017:2529–2535.
13. Novakovic JD, Veljovic A, Ilic SS, Papic Z, Milica T. Evaluation of classification models in machine learning. Theory Appl Math Comput Sci. 2017;7:39–46.
14. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med. 2016;375:1216–1219.
15. Okedo-Alex IN, Akamike IC, Ezeanosike OB, Uneke CJ. Determinants of antenatal care utilisation in sub-Saharan Africa: a systematic review. BMJ Open. 2019;9:e031890.
16. Oshinyemi TE, Aluko JO, Oluwatosin OA. Focused antenatal care: Re-appraisal of current practices. Int J Nurs Midwifery. 2018;10:90–98.
17. Shibre G, Zegeye B, Idriss-Wheeler D, Yaya S. Factors affecting the utilization of antenatal care services among women in Guinea: a population-based study. Fam Pract. 2021;38:63–69.
18. Sufriyana H, Wu YW, Su EC. Artificial intelligence-assisted prediction of preeclampsia: development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine. 2020;54:102710.
19. Tan J, Yang J, Wu S, Chen G, Zhao J. A critical look at the current train/test split in machine learning. arXiv:2106.04525. 2021.
20. Tessema ZT, Minyihun A. Utilization and determinants of antenatal care visits in East African countries: a multicountry analysis of demographic and health surveys. Adv Public Health. 2021;2021:6623009.
21. Tizazu MA, Asefa EY, Muluneh MA, Haile AB. Utilizing a minimum of four antenatal care visits and associated factors in Debre Berhan town, Ethiopia. Risk Manag Healthc Policy. 2020;13:2783–2791.
22. Wairoto KG, Joseph NK, Macharia PM, Okiro EA. Determinants of subnational disparities in antenatal care utilisation: a spatial analysis of demographic and health survey data in Kenya. BMC Health Serv Res. 2020;20:724.
23. World Health Organization. Strategies toward ending preventable maternal mortality (EPMM). Geneva: WHO; 2015.

24. World Health Organization. Trends in maternal mortality: 2000 to 2017. Estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division. Geneva: WHO; 2019.
25. World Health Organization. WHO antenatal care randomized trial: Manual for the implementation of the new model. WHO/RHR/01.30. Geneva: WHO; 2002.
26. World Health Organization. WHO recommendations on antenatal care for a positive pregnancy experience. Geneva: WHO; 2016.
27. Yadav ML, Roychoudhury B. Handling missing values: a study of popular imputation packages in R. Knowl Based Syst. 2018;160:104–118.
28. Yargawa J, Leonardi-Bee J. Male involvement and maternal health outcomes: systematic review and meta-analysis. J Epidemiol Community Health. 2015;69:604–612.