

Application of Machine Learning to Predict Teenage Pregnancy in Zambia: Evidence from 2024 Zambia Demographic and Health Surveys

Teebeny Zulu

teebenyzulu@gmail.com

The University of Zambia

Nasson Nathan Tembo

Mulungushi University

Patrick Musonda

The University of Zambia

Research Article

Keywords: Teenage pregnancy, Zambia Demographic and Health Survey, Predictors, Machine learning, Logistic regression, XGB

Posted Date: February 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8754908/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

Teenage pregnancy remains a significant public health issue in Zambia, contributing to high maternal and child morbidity and mortality, school dropout, and socioeconomic vulnerability. Understanding the most influential predictors of teenage pregnancy is a key initial step in developing targeted interventions to address teenage pregnancy. Therefore, this study is aimed at applying machine learning (ML) and ordinary logistic regression to determine and explain the most influential predictors of teenage pregnancy in Zambia.

Methods

This cross-sectional study used data from the 2024 Zambia Demographic and Health Survey (ZDHS). Feature selection was performed using mutual information scores and the Boruta algorithm in 5-fold cross-validation (CV). Data was split into 80% training set and 20% held-out testing set. Seven ML models were trained in python on 80% of the training set using 5-fold CV and average metrics were obtained on whom the best performing model was identified which was later tested on the held-out testing set. Model performance was evaluated using accuracy and the area under the receiver operating characteristic curve (AUC-ROC). The Shapley Additive Explanation (SHAP) method was applied to determine feature importance. Weighted logistic regression in STATA was then used to estimate odds ratios for key predictors.

Results

A total of 3292 adolescents were included in the analysis, of whom 26.3% had experienced teenage pregnancy. The Extreme Gradient Boosting (XGB) model outperformed other models, achieving an accuracy of 0.806 (0.789–0.822) and AUC-ROC 0.839 (0.821–0.858). On the held-out test set, the XGB model achieved the accuracy of 0.746 an AUC-ROC of 0.824. The most influential predictors of teenage pregnancy were participant's current age, followed by age of the household head, knowledge of ovulatory cycle, wealth status, education, type of place of residence, frequency of watching television, internet use, and province. A one-year increase in participant's age (OR = 2.20; 95% CI = 2.03–2.37) and knowledge of ovulatory cycle (OR = 2.70; 95% CI = 2.12–3.42), increased the odds of teenage pregnancy in Zambia. Conversely, a one-year increase in the age of the household head (OR = 0.98; 95% CI = 0.98–0.99), belonging to the richer wealth quintile (OR = 0.58; 95% CI = 0.38–0.90), belonging to the richest quintile (OR = 0.18; 95% CI = 0.10–0.31), having attained secondary or higher education (OR = 0.42; 95% CI = 0.25–0.71), watching television at least once a week (OR = 0.67; 95% CI = 0.47–0.95), watching television almost every day (OR = 0.72; 95% CI = 0.52–0.99), internet use in the last 12 months (OR = 0.67; 95% CI = 0.49–0.92), reduced the odds of teenage pregnancy.

Conclusion

Applying ML models and conventional modelling could help identify and explain the most influential predictors of teenage pregnancy to support evidence-based, targeted interventions.

1. Introduction

Teenage pregnancy constitutes a major public health and socio-economic concern globally, more so in developing regions, where an estimated 21 million girls aged 15–19 years become pregnant annually, resulting in an estimated 12 million births [1]. An estimated 50% of these annual pregnancies are unintended, with more than half ending in abortions, which are often unsafe [2, 3]. In Zambia, although the general fertility trend has been on the decline since 2007, about 28% of teenage girls have a history of pregnancy across the country [4, 5].

Teenage pregnancy presents negative economic, psychosocial, and health consequences for teenage girls, their infants, their families, and communities. Teenage mothers face higher risks of eclampsia, puerperal endometritis, obstructed labor, obstetric fistula, sexually transmitted infections, and systemic infections than women aged 20–24 years [1, 4, 6]. Additionally, teenage pregnancy has been associated with low economic productivity, high maternal and neonatal morbidity and mortality [6]. Consequently, babies born to teenage mothers face higher risks of low birth weight, preterm birth, intrauterine growth retardation, and severe neonatal conditions [4, 7].

The specific factors influencing teenage pregnancy are complex and context-specific. Findings using conventional modeling techniques in Zambia and beyond have shown that several social, economic, and cultural factors may influence teenage pregnancy [1, 3, 4, 6–10]. However, these techniques assume linear relationships between the predictors and the outcome and could be limited in the case of complex non-linear relationships. In addition, these techniques lack the ability to rank the predictors in order of their predictive influence on the outcome [11, 12]. In such cases where the stability of estimates from conventional methods could be unreliable, ML models can be employed to simultaneously capture complex non-linear relationships, rank features based on influence, and optimize model performance [12, 13].

Despite ML modelling being shown to be a promising modelling approach to modelling health outcomes, research that combines conventional modelling approaches and ML techniques to predict the risk of teenage pregnancy in Zambia remains scanty, with the majority of studies conducted outside the country largely focusing on applying ML techniques in predicting pregnancy outcomes rather than teenage pregnancy [14–21]. Therefore, this study aims to apply ML techniques and conventional weighted logistic regression to national-level data to predict the risk of teenage pregnancy among girls in Zambia to inform targeted measures in mitigating the burden.

2. Materials and Methods

Machine learning modelling steps implemented in this study are presented in Fig. 1. The 2024 ZDHS raw data was cleaned where we checked for missing values, renaming variables and checking for outliers. The data was later split into 80% of training set and 20% was held-out for final testing and model validation. Five-fold CV was implemented on the training set. The preprocessing step was implemented during CV in 4-folds and fitted on the remaining 1-fold (testing fold). Feature selection was performed in each fold using the mutual information scores and the Boruta algorithm whose scores were averages of scores across 5-folds for each feature. Similarly, ML models were trained on 4-folds (80% of the training set) and tested on the remaining fold (20% of the training set) at each iteration. Performance metrics were averaged across 5-folds. Based on the averaged metrics, the best performing ML model was selected. This model was then tested on the held-out test set which remained untouched throughout model training to avoid data leakage. Finally, feature importance using SHAP was performed and odds ratios through conventional logistic regression were obtained to help quantify the influence of predictors on teenage pregnancy.

2.1 Study Design.

This analytical cross-sectional study utilized data from the 2024 ZDHS in the women's file (*IR file*).

2.2 Data source and population

This study analyzed the women's data from the 2024 ZDHS. The Demographic and health survey is a national representative survey used to collect information from women aged 15–49 years and men aged 15–59 years about demographic and health status every five years. This survey collected demographic, socio-economic, health and nutritional history, from participants in household interviews. In this study 3292 girls below 20 years with complete information on pregnancy history were included for analysis. Details of the Demographic and health survey are published in [22].

2.3 Outcome variable

The outcome variable in this study was teenage pregnancy and was defined as girls aged 15–19 years who were currently pregnant or had at least one child at the time of the survey [27]. Teenage pregnancy was a binary composite variable created

from two questions in the survey; 1) are you currently pregnant (*no, yes*); 2) total number of children ever born. The variable teenage pregnancy was coded 1 (*yes*) if a girl answered yes to the first question or if she ever had a child, otherwise teenage pregnancy was coded 0 (*no*).

2.4 Explanatory variables

Several factors reported to be associated with teenage pregnancy in the literature were assessed in this study and these included; respondent's current age, wealth quintile, frequency of watching television, age of the household head, frequency of reading newspapers, highest level of education, type of place of residence, employment status, knowledge of ovulatory cycle, internet usage, province, having a mobile phone, sex of the household head, and the frequency of listening to radio [27–31].

2.5 Handling of Missing Data.

To start with, missing values in the outcome (teenage pregnancy) were dropped and this reduced the sample size to 3292. Missing values in the features were imputed using the mean for continuous variables since both variables (respondent's current age and age of the household head) were normally distributed and the mode for categorical variables.

2.6 Descriptive statistics and statistical software.

Weighted frequencies and percentages were presented for categorical variables. Associations between teenage pregnancy and categorical predictors were explored using the weighted chi-squared (Rao-Scott) test and the T-test was used for quantitative predictors. Data cleaning and exploratory analysis such as checking for missing values, feature engineering, univariate analysis and logistic regression were all performed in STATA version 17. The Python version 3.12.8 was used for ML model building [23].

2.7 Handling class Imbalance

To address class imbalance in the outcome variable, class weighting was applied during model training. Weights were assigned inversely proportional to class frequencies, enabling improved detection of the minority class in the outcome while preserving the original data structure [24].

2.8 Validation Strategy

This study used the Hold-out method and 5-fold CV strategy to validate the models. To start with, data was split into 80% training and 20% held-out testing sets. The 20% testing set was set aside for testing after selecting the best performing ML model. The 80% training set was partitioned into 5-folds where at each iteration 4-folds were used for training and the remaining 1-fold was used for testing, the performance metrics were saved at each iteration, and the average metric was reported for each metric. The best ML model was then selected based on the average metric and it was tested on the 20% reserved set. This strategy has been shown to be sufficient in assessing the generalizability of ML models [25]. The ML models used were optimized for performance using the GridSearchCV hyper parameter tuning strategy [26].

2.9 Machine Learning Algorithms

Seven ML algorithms were trained in this study and these included the Logistic Regression (LR), decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), Extreme Gradient Boosting (XGB) and Artificial Neural Networks (ANN). These algorithms are discussed in detail in other studies [22].

2.10 Performance Evaluation Metrics

The metrics used to evaluate the performance of models in this study were accuracy, precision, recall, F1-score, balanced accuracy, and the AUC-ROC. The accuracy and AUC-ROC were prioritized.

2.11 Feature importance analysis

Feature importance was performed using the SHAP values to rank the predictors according to their influence on the prediction of teenage pregnancy in XGB model. Logistic regression was later performed to quantify the influence that each important predictor had on teenage pregnancy.

3. Results

3.1 Demographic characteristics of study participants

A total of 3292 adolescents were included in the analysis, of whom 26.3% (865) had experienced teenage pregnancy (Table 1). The mean age of the participants was 16.9 years (SD = 1.4), and the mean age of household heads was 45.7 years (SD = 13.2). Slightly more than half of participants resided in rural areas 50.8% (1672). The largest proportion of participants was from Lusaka 16.7% (550). Most adolescents lived in male-headed households 66.1% (2177).

Table 1
Demographic characteristics of study participants five years preceding the 2024 ZDHS

Variable	Overall 3292 (100%)	Teenage pregnancy		p-value
		No 2427 (73.7%)	Yes 865 (26.3%)	
Age (years), mean (SD)	16.9 (1.4)	16.6 (1.4)	17.82 (1.2)	< 0.001 ^t
Age of household head (years), mean (SD)	45.7 (13.2)	46.83 (12.02)	42.68 (15.5)	< 0.001 ^t
Place of residence				
Urban	1620 (49.2)	1345 (55.4)	275 (31.8)	
Rural	1672 (50.8)	1082 (44.6)	590 (68.2)	< 0.001 ^c
Province				
Central	403 (12.2)	282 (11.6)	121 (14.0)	
Copperbelt	457 (13.9)	380 (15.7)	77 (8.9)	
Eastern	389 (11.8)	232 (9.6)	156 (18.1)	
Luapula	258 (7.8)	198 (8.2)	59 (6.9)	
Lusaka	550 (16.7)	473 (19.5)	77 (8.9)	
Muchinga	142 (4.3)	95 (3.9)	47 (5.4)	
Northern	251 (7.6)	180 (7.4)	70 (8.1)	
North-Western	227 (6.9)	155 (6.4)	72 (8.3)	
Southern	400 (12.2)	274 (11.3)	126 (14.6)	
Western	216 (6.6)	157 (6.5)	60 (6.9)	< 0.001 ^c
Sex of household head				
Male	2177 (66.1)	1590 (65.5)	587 (67.9)	
Female	1115 (33.9)	837 (34.5)	278 (32.1)	0.2490 ^c
<i>t: T-test, c: weighted chi-squared test (Rao-Scott test)</i>				

3.2 Demographic characteristics of study participants by teenage pregnancy status

Adolescents who had experienced teenage pregnancy were significantly older than those who had not (17.8 vs. 16.6 years; $p < 0.001$) (Table 1). The mean age of household heads was also significantly lower among adolescents who had experienced pregnancy compared with those who had not (42.7 vs. 46.8 years; $p < 0.001$). Place of residence was significantly associated

with teenage pregnancy ($p < 0.001$), with a higher proportion of rural residents among adolescents with teenage pregnancy (68.2%) compared to those without (44.6%). Teenage pregnancy also varied significantly by province ($p < 0.001$), with higher proportions observed in Eastern (18.1%), Southern (14.6%), and Central (14.0%) provinces, and lower proportions in Lusaka (8.9%) and Copperbelt (8.9%).

3.3 Overall socio-economic characteristics of study participants by teenage pregnancy status

Overall, 22.7% (746) of adolescents belonged to the richest wealth quintile (Table 2). More than half of participants reported not watching television at all, 58.4% (1924). Approximately two-thirds of adolescents, 66.5% (2190), reported knowledge of the ovulatory cycle. Most adolescents had attained secondary or higher education, 57.5% (1894). A majority reported no internet use, 78.4% (2581), and were not employed, 82.0% (2699). Most participants did not read newspapers or magazines at all, 85.2% (2,805), did not own a mobile telephone, 66.4% (2187), and did not listen to the radio at all, 64.1% (2111).

Table 2
Socio-economic characteristics of study participants five years preceding the 2024 ZDHS

Variable	Overall 3292 (100%)	Teenage pregnancy		p-value
		No	Yes	
		2427 (73.7%)	865 (26.3%)	
Wealth status				
Poorest	558 (16.9)	323 (13.3)	235 (27.1)	
Poorer	569 (17.3)	372 (15.3)	197 (22.8)	
Middle	685 (20.8)	454 (18.7)	230 (26.6)	
Richer	735 (22.3)	583 (24.0)	152 (17.6)	
Richest	746 (22.7)	695 (28.6)	51 (5.9)	< 0.001 ^t
Frequency of watching television				
Not at all	1924 (58.4)	1268 (52.2)	656 (75.8)	
Less than once a week	113 (3.4)	80 (3.3)	33 (3.8)	
At least once a week	414 (12.6)	346 (14.3)	68 (7.9)	
Almost every day	842 (25.6)	733 (30.2)	109 (12.6)	< 0.001 ^t
Knowledge of ovulatory cycle				
No	1102 (33.5)	941 (38.8)	161 (18.6)	
Yes	2190 (66.5)	1486 (61.2)	704 (81.4)	< 0.001 ^t
Education level				
No education	100 (3.1)	59 (2.4)	42 (4.8)	
Primary	1298 (39.4)	864 (35.6)	434 (50.1)	
Secondary or higher	1894 (57.5)	1504 (62.0)	390 (45.1)	< 0.001 ^t
Internet use				
No	2581 (78.4)	1824 (75.2)	757 (87.5)	
Yes	711 (21.6)	602 (24.8)	109 (12.5)	< 0.001 ^t
Employment status				
No	2699 (82.0)	2043 (84.2)	656 (75.8)	
Yes	593 (18.0)	384 (15.8)	210 (24.2)	< 0.001 ^t
Frequency of reading newspaper/magazine				
Not at all	2805 (85.2)	2033 (83.8)	773 (89.3)	
Less than once a week	151 (4.6)	119 (4.9)	32 (3.7)	
At least once a week	195 (5.9)	153 (6.3)	42 (4.8)	
Almost every day	141 (4.3)	123 (5.1)	19 (2.2)	0.0008 ^t
t: T-test, c: weighted chi-squared test (Rao-Scott test)				

Variable	Overall	Teenage pregnancy		p-value
	3292 (100%)	No	Yes	
		2427 (73.7%)	865 (26.3%)	
Wealth status				
Owns a mobile telephone				
No	2187 (66.4)	1645 (67.8)	543 (62.7)	0.0147
Yes	1105 (33.6)	782 (32.2)	323 (37.3)	
Frequency of listening to the radio				
Not at all	2111 (64.1)	1514 (62.4)	597 (69.0)	0.0089
Less than once a week	235 (7.1)	185 (7.6)	49 (5.7)	
At least once a week	625 (19.0)	484 (19.9)	141 (16.3)	
Almost every day	322 (9.8)	244 (10.0)	78 (9.0)	
t: T-test, c: weighted chi-squared test (Rao-Scott test)				

3.4 Socio-economic characteristics by teenage pregnancy status

Wealth status showed a significant association with teenage pregnancy ($p < 0.001$), with the highest proportion observed in the poorest wealth quintile, 27.1% (235) as shown in Table 2. Similarly, most pregnant adolescents 75.8% (656), did not watch television at all compared with those without pregnancy ($p < 0.001$). Knowledge of the ovulatory cycle also differed significantly by pregnancy status ($p < 0.001$), with 81.4% (704) of adolescents with teenage pregnancy reporting correct knowledge. Education level was significantly related to teenage pregnancy ($p < 0.001$), with a greater proportion of affected adolescents having only primary education, 50.1% (434). Internet use was lower among adolescents who had experienced pregnancy, 12.5% ($n = 109$) ($p < 0.001$), while employment was more common in this group, 24.2% ($n = 210$) ($p < 0.001$). In addition, there was a significant difference in the frequency of listening to the radio ($p = 0.0089$) and reading newspapers or magazines ($p = 0.0008$) among teenagers who were pregnant as compared to those who were not pregnant. In fact, 597 (69.0) and 89.3% (773) of teenagers among those who were pregnant did not listen to the radio and read newspapers or magazines respectively. More teenagers among those who were pregnant 323 (37.3) owned mobile phones as compared to those who were not pregnant 32.2% (782) ($p = 0.0147$).

3.5 Feature selection

Feature selection was performed using Mutual information scores and the Boruta algorithm which were both implemented inside 5-fold CV. At each iteration feature selection was performed in 4-folds for five iterations in total since data was partitioned into 5 splits during 5-fold CV. Mutual information scores were calculated at each iteration and stored, then the averaged scored was computed across five iterations and plotted (Fig. 2). For Boruta algorithm, features were classified to be relevant or not at each iteration and the set was stored. The proportion of folds in which each feature emerged relevant was calculated and plotted (Fig. 3). During Boruta algorithm feature selection, six features were classified to be relevant in all the five iterations (5-fold CV).

Out of 14 features known to be associated with teenage pregnancy from literature and were available in the dataset, 9 of them were selected for model training. The 9 features selected in this study included; age, age of the household head, knowledge of ovulatory cycle, wealth status, level of education, type of place of residence, frequency of watching television, internet use, and province. This list was a union of features from prior knowledge, mutual information scores (Fig. 2) and Boruta algorithm (Fig. 3).

The figure shows the ranking of predictors based on mean mutual information with the outcome. Higher values indicate stronger dependency between the predictor and teenage pregnancy. Error bars represent variability across resampling iterations across 5-folds.

3.6 Machine learning model selection

Seven ML models which included Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM) Gradient Boosting (GB) Extreme Gradient Boosting (XGB) and Artificial Neural Networks (ANN) were trained. Hyperparameters were tuned for optimal performance using GridSearchCv (Table 3). The performance of these models was evaluated using accuracy precision, recall, F1 score balanced accuracy, and AUC-ROC as shown in Table 4 and Fig. 4. All models performed well with the minimum accuracy being 0.736 (0.697–0.775) for DT and the maximum being 0.808 (0.792–0.824) for SVM. In terms of AUC-ROC, the minimum was 0.810 (0.786–0.834) for DT algorithm and the maximum was 0.839 (0.821–0.858) for XGB. The best performing model was selected based on the accuracy and AUC-ROC. The XGB algorithm with accuracy 0.806 (0.789–0.822) and AUC-ROC 0.839 (0.821–0.858) was selected as the best model for this study though other models performed comparatively well.

Table 3
hyperparameter tunings for ML models trained in 5-fold CV

Model	Key optimized hyperparameters
LR	C = 0.13
DT	Max depth = 6; Min samples split = 20; Min samples leaf = 16
RF	No. trees = 330; Max depth = 10; Min samples split = 5; Min samples leaf = 7
GB	No. trees = 357; Max depth = 2; Learning rate = 0.020; Subsample = 0.868
XGB	No. trees = 286; Max depth = 2; Learning rate = 0.040; Subsample = 0.660; Column subsample = 0.736; Gamma = 1.05; Min child weight = 12
SVM	C = 0.31; Gamma = 0.071
ANN	Hidden layers = (32); L2 regularization (α) = 0.0089; Initial learning rate = 0.00012

Table 4
performance evaluation metrics for machine learning models

Model	ROC AUC (95% CI)	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)	Balanced Accuracy (95% CI)
XGB	0.839 (0.821–0.858)	0.806 (0.789–0.822)	0.716 (0.669–0.763)	0.560 (0.500–0.620)	0.597 (0.559–0.636)	0.717 (0.693–0.740)
GB	0.839 (0.822–0.856)	0.798 (0.786–0.809)	0.704 (0.658–0.751)	0.504 (0.450–0.559)	0.574 (0.559–0.590)	0.703 (0.693–0.712)
ANN	0.836 (0.817–0.855)	0.803 (0.788–0.818)	0.696 (0.646–0.746)	0.547 (0.498–0.597)	0.604 (0.584–0.625)	0.721 (0.710–0.733)
SVM	0.836 (0.818–0.853)	0.808 (0.792–0.824)	0.695 (0.655–0.736)	0.584 (0.535–0.634)	0.623 (0.595–0.650)	0.734 (0.717–0.750)
RF	0.835 (0.815–0.856)	0.760 (0.738–0.782)	0.556 (0.523–0.590)	0.754 (0.705–0.804)	0.632 (0.607–0.658)	0.752 (0.733–0.771)
LR	0.831 (0.810–0.851)	0.750 (0.727–0.774)	0.541 (0.507–0.575)	0.767 (0.721–0.813)	0.628 (0.602–0.655)	0.750 (0.730–0.770)
DT	0.810 (0.786–0.834)	0.736 (0.697–0.775)	0.524 (0.466–0.582)	0.783 (0.736–0.830)	0.614 (0.578–0.650)	0.738 (0.712–0.765)

1. Values are mean (95% confidence interval) derived from nested CV.

2. Bold values indicate the best-performing model per metric

3.7 Performance of the best ML (XGB) Model on the held-out test set

On the held-out test set, the XGB model achieved an accuracy of 0.746 and an AUC-ROC of 0.824 indicating good discriminative performance (Fig. 5). The precision–recall curve yielded an average precision (AP) of 0.676, reflecting the model's ability to maintain a favorable balance between precision and recall across different classification thresholds (Fig. 6).

Calibration analysis showed that the predicted probabilities were not perfectly aligned with the observed outcome proportions (Fig. 7). In particular, the model tended to overestimate risk at lower and intermediate predicted probability levels, while calibration improved at higher predicted probabilities, where the observed proportions more closely followed the ideal calibration line.

3.8 Feature importance

The Shapley additive explanation in Fig. 8 show participant's current age to be the most influential predictor of teenage pregnancy in the XGB model followed by age of the household head, knowledge of ovulatory cycle, wealth status, education, type of place of residence, frequency of watching television, internet use, and province.

3.9: Quantifying the effect of the important predictors of teenage pregnancy using odds ratios

Important predictors of teenage pregnancy identified and ranked using ML models were included in a LR model to obtain odds ratios to quantify the influence of each predictor (Table 5). It was found that a one-year increase in participant's age (OR = 2.20; 95% CI = 2.03–2.37; $P < 0.0001$), knowledge of ovulatory cycle (OR = 2.70; 95% CI = 2.12–3.42; $P < 0.0001$), increased the odds of teenage pregnancy in Zambia

On the other hand, a one-year increase in the age of the household head (OR = 0.98; 95% CI = 0.98–0.99; $P < 0.0001$), belonging to the richer wealth quintile (OR = 0.58; 95% CI = 0.38–0.90; $P = 0.014$), belonging to the richest quintile (OR = 0.18; 95% CI = 0.10–0.31; $P < 0.0001$), having attained secondary or higher education (OR = 0.42; 95% CI = 0.25–0.71; $P = 0.001$), watching television at least once a week (OR = 0.67; 95% CI = 0.47–0.95; $P = 0.025$), watching television almost every day (OR = 0.72; 95% CI = 0.52–0.99; $P = 0.040$), internet use in the last 12 months (OR = 0.67; 95% CI = 0.49–0.92; $P = 0.012$), reduced the odds of teenage pregnancy.

Predictor	COR (95% CI)	p-value	AOR (95% CI)	p-value
Age (years)	1.97 (1.84–2.11)	< 0.0001	2.20 (2.03–2.37)	< 0.0001
Age of household head (years)	0.98 (0.97–0.98)	< 0.0001	0.98 (0.98–0.99)	< 0.0001
Wealth status				
Poorest	ref		ref	
Poorer	0.73 (0.58–0.91)	0.006	0.92 (0.70–1.20)	0.534
Middle	0.70 (0.55–0.89)	0.003	0.91 (0.67–1.24)	0.555
Richer	0.36 (0.27–0.47)	< 0.0001	0.58 (0.38–0.90)	0.014
Richest	0.10 (0.07–0.15)	< 0.0001	0.18 (0.10–0.31)	< 0.0001
Knowledge of ovulatory cycle				
No	ref		ref	
Yes	2.76 (2.26–3.39)	< 0.0001	2.70 (2.12–3.42)	< 0.0001
Highest level of education				
No education	ref		ref	
Primary	0.71 (0.46–1.08)	0.108	0.82 (0.51–1.31)	0.406
Secondary or higher	0.36 (0.24–0.56)	< 0.0001	0.42 (0.25–0.71)	0.001
Region				
Central	ref		ref	
Copperbelt	0.47 (0.32–0.70)	< 0.0001	0.71 (0.46–1.09)	0.116
Eastern	1.57 (1.09–2.26)	0.017	1.33 (0.84–2.10)	0.219
Luapula	0.70 (0.48–1.02)	0.060	0.38 (0.25–0.59)	< 0.0001
Lusaka	0.38 (0.24–0.59)	< 0.0001	0.73 (0.44–1.19)	0.203
Muchinga	1.15 (0.79–1.67)	0.464	1.08 (0.71–1.65)	0.726
Northern	0.91 (0.62–1.34)	0.632	0.66 (0.42–1.04)	0.073
North Western	1.08 (0.75–1.54)	0.682	0.88 (0.59–1.30)	0.510
Southern	1.07 (0.75–1.53)	0.699	0.97 (0.66–1.41)	0.867
Western	0.89 (0.62–1.26)	0.499	0.80 (0.54–1.19)	0.263
Frequency of watching television				
Not at all	ref		ref	
Less than once a week	0.79 (0.52–1.20)	0.265	1.34 (0.84–2.14)	0.221
At least once a week	0.38 (0.28–0.51)	< 0.0001	0.67 (0.47–0.95)	0.025
Almost every day	0.29 (0.22–0.37)	< 0.0001	0.72 (0.52–0.99)	0.040
Type of place of residence				
Urban	ref		ref	
Rural	2.67 (2.18–3.27)	< 0.0001	1.09 (0.81–1.47)	0.550

Predictor	COR (95% CI)	p-value	AOR (95% CI)	p-value
Internet use in last 12 months				
No	ref		ref	
Yes	0.43 (0.34–0.56)	< 0.0001	0.67 (0.49–0.92)	0.012

COR = crude odds ratio; AOR = adjusted odds ratio; CI = 95% confidence interval; ref=Reference category.

4. Discussion

Extreme gradient boosting was the best performing ML model in this study with accuracy of 0.806 (0.789–0.822) and AUC-ROC of 0.839 (0.821–0.858). When this model was tested on the held-out set to check generalizability, the AUC-ROC was 0.824 and accuracy was 0.746. The performance of this model was good since most ML models struggle to capture complex human behaviours with high accuracy [27].

An increase in the age of a teenager was associated with an increase in the risk of teenage pregnancy and this predictor was the most influential. This pattern is largely driven by the progressive likelihood of sexual initiation as girls grow older, coupled with increased exposure to situations that facilitate unprotected sexual activity [10]. Older adolescents may also face heightened social pressures to engage in relationships, marry, or demonstrate fertility, particularly in rural areas where cultural norms emphasize early childbearing [28, 29]. This age-related gradient underscores the importance of targeting older adolescents with tailored interventions that combine sexual education, contraceptive access, and social support to delay first pregnancies and reduce associated health and social risks.

The second most influential predictor of teenage pregnancy in this study was the age of the household head, with adolescents living in households headed by older individuals being less likely to become pregnant. This finding suggests that household maturity and stability play a protective role in shaping adolescent reproductive outcomes. Studies have shown that family structure, parental presence, and household supervision are important determinants of adolescent pregnancy [30, 31]. Older household heads may provide stronger guidance, monitoring, and enforcement of social norms that discourage early sexual activity and childbearing. Studies have demonstrated that household contextual factors such as media exposure, family environment, and socioeconomic stability are associated with lower odds of teenage pregnancy [33, 34]. Moreover, supportive home environments have been shown to enhance school continuation and improve communication about sexual and reproductive health, both of which are protective against early pregnancy [35]. These findings underscore the importance of family-centered interventions and caregiver engagement strategies in efforts to reduce teenage pregnancy in Zambia.

Knowledge of ovulatory cycle ranked third from the most influential predictor of teenage pregnancy. Teenagers who reported having knowledge of their ovulatory cycle had a higher likelihood of teenage pregnancy in Zambia. Although this may appear counterintuitive, similar findings have been reported in other studies using Demographic and Health Survey data from sub-Saharan Africa [36–38]. Some studies have reported that adolescents with knowledge of their menstrual cycle may fall pregnant on purpose to prove their femininity [38]. On the other hand, the reported “knowledge” may not necessarily indicate accurate understanding of the fertility within the menstrual cycle, many adolescents misunderstand the timing of fertility within the menstrual cycle. This finding highlights the need to strengthen comprehensive sexuality education in Zambia, ensuring that young people not only have access to reproductive health information but also develop accurate, practical understanding of fertility and contraceptive use.

The fourth most influential predictor of teenage pregnancy was wealth quintile where wealthier teenagers had a lower risk of being pregnant compared to their poor counterparts. This finding aligns with previous studies [10, 30, 39]. Economic well-being often provides better access to education, reproductive health information, and contraceptive services, which collectively lower the likelihood of teenage pregnancy. In contrast, adolescents from poorer households may face limited access to education, financial insecurity, and social vulnerability that increase their susceptibility to early marriage or transactional sex [40]. Moreover, wealthier families are more likely to prioritize girls’ education which was the fifth most influential predictor of

teenage pregnancy and was found to reduce the risk of teenage pregnancy in this study. Education delay marriage, empower girls to better face future challenges, increases girls' ability to negotiate for safer sex and also increases their involvement in reproductive health decision making, thereby reducing exposure to early pregnancy[41]. This socioeconomic gradient underscores the importance of poverty reduction and equitable access to education and reproductive health services as key strategies to prevent teenage pregnancy in Zambia.

The sixth and seventh most influential predictor of teenage pregnancy were the frequency of watching television and internet use respectively. Teenagers who reported watching television at least once a week or every day had a lower risk of teenage pregnancy in Zambia. In addition, teenagers who used the internet in the last 12 months prior to the 2024 survey were less likely to become pregnant. These findings are consistent with previous studies which have shown that exposure to mass media, including newspapers, radio, internet, and television, is associated with improved reproductive health knowledge and lower rates of early childbearing [36]. Access to mass media serves as an important channel for disseminating information about family planning, contraception, and the consequences of early pregnancy [42]. Therefore, strengthening media-based health communication strategies, especially those targeting adolescents in rural and low-literacy communities, could contribute to reducing teenage pregnancy rates in Zambia.

5. Strengths and limitations

The strength of this study lies in the use of the national wide representative sample which allows us to generalize these findings to all teenager girls in Zambia. In addition, this study combines two data analysis approaches namely, ML modelling and conventional statistical analysis. This hybrid approach ensures that each method complement each other's weakness and makes the story complete. For instance, ML models are robust algorithms with the ability to rank predictors, however, these models do not quantify the influence of each predictor in terms of odds ratios hence conventional statistical analysis comes in handy. Despite the methodological rigor of this study, causality could not be claimed since this study was a cross-sectional study. Furthermore, the outcome of this study was self-reported and could suffer from recall bias.

Conclusion

In trying to understand predictors of teenage pregnancy, we applied machine learning models and the top nine predictors of teenage pregnancy ranked in order of their influence were; participant's current age, age of the household head, knowledge of ovulatory cycle, wealth status, education, type of place of residence, frequency of watching television, internet use, and province. Therefore, it is essential to align statutory and customary laws regarding the legal age of marriage in Zambia and implement behavioural change communication strategies that discourage early sexual debut, particularly targeting older adolescents. Additionally, the role of media platforms, particularly television, in disseminating information about family planning in Zambia is crucial. Furthermore, efforts to promote comprehensive education on the correct use of contraception while ensuring consistent access to contraception among adolescents should be encouraged.

Declarations

Author Contribution

T.Z data analysis, and writing the main manuscript textN.N.T writing and reviewing the manuscriptP.M reviweing the manuscript and providing overall supervision of the project

Ethical Approval, Human ethics and Consent to Participate declarations

This study was a secondary analysis of the ZDHS dataset which is publicly available upon registration online as there was no need for consent to participate. The ICF Institutional Review Board approved the ZDHS data survey protocols with the ICF Project Number: 132989.0. 000. ZM. DHS.02. Permission to use the dataset was acquired from ICF Macro, and the ZDHS dataset is available for download upon registration at <https://www.dhsprogram.com/data>. The user carefully adhered to the

guidelines given, highlighting the sensitive nature of the information and the significance of not trying to uncover the identity of any household or individual surveyed for the study (maintaining anonymity)

Consent to Publish declarations

All authors of this manuscript consented to submit and publish this work to the journal

Data availability statement

The datasets generated and/or analyzed during the current study are available in the Demographic and health survey (DHS) repository, <https://www.dhsprogram.com/data>.

Clinical trial number

Not applicable

Funding declaration

No specific funding was received for this work

References

1. WHO. Adolescent pregnancy [Internet]. 2024 [cited 2025 Oct 5]. Available from: <https://www.who.int/news-room/fact-sheets/detail/adolescent-pregnancy>
2. Sully EA, Biddlecom A, Darroch JE, Riley T, Ashford LS, Lince-Deroche N et al. Adding It Up: Investing in Sexual and Reproductive Health 2019. N Y Guttmacher Inst [Internet]. 2020; Available from: <https://www.guttmacher.org/report/adding-it-up-investing-in-sexual-reproductive-health-2019>
3. Malunga G, Sangong S, Saah FI, Bain LE. Prevalence and factors associated with adolescent pregnancies in Zambia: a systematic review from 2000–2022. Arch Public Health [Internet]. 2023 Feb 20 [cited 2025 Oct 5];81(1):27. Available from: <https://doi.org/10.1186/s13690-023-01045-y>
4. Tembo NN, Jacobs C, Maliko CN, Musonda P. Effect of Exposure to Boarding Schooling on Pregnancy among School-Going Teenagers: A Retrospective Case-Control Study in Eastern Province, Zambia. Adolescents [Internet]. 2023 Sept [cited 2025 Oct 5];3(3):446–56. Available from: <https://www.mdpi.com/2673-7051/3/3/31>
5. Zambia Statistics Agency, Ministry of Health (MoH) [Zambia], ICF. Zambia Demographic and Health Survey 2024: Key Indicators Report. Lusaka Zamb Rockv Md USA Zamb Stat Agency MoH ICF [Internet]. 2024; Available from: www.DHSprogram.com.
6. Phiri M, Kasonde ME, Moyo N, Sikaluzwe M, Simona S. A multilevel analysis of trends and predictors associated with teenage pregnancy in Zambia (2001–2018). Reprod Health [Internet]. 2023 Jan 18 [cited 2025 Oct 5];20(1):16. Available from: <https://doi.org/10.1186/s12978-023-01567-2>
7. Crooks R, Bedwell C, Tina L. Adolescent experiences of pregnancy in low-and middle-income countries: a meta-synthesis of qualitative studies | BMC Pregnancy and Childbirth [Internet]. BMC Pregnancy and Childbirth. 2022 [cited 2025 Oct 5]. Available from: <https://link.springer.com/article/10.1186/s12884-022-05022-1>
8. Ochen AM, Chi PC, Lawoko S. Predictors of teenage pregnancy among girls aged 13–19 years in Uganda: a community based case-control study. BMC Pregnancy Childbirth [Internet]. 2019 June 24 [cited 2025 Oct 5];19(1):211. Available from: <https://doi.org/10.1186/s12884-019-2347-y>
9. Nduhuye F, Kubana E, Matutina S, Mwesigye D, Munyaneza A, Nyirazinyoye L. The factors associated with teenage pregnancy among young women aged between 15 and 19 years in Rwanda: a retrospective cross-sectional study on the Rwanda Demographic Health Survey 2019–2020. Front Reprod Health. 2024;6:1453933.
10. Mbele W, Jassey B, Nyirenda N. Factors associated with adolescent pregnancy in Zambia: a secondary data analysis of the 2018 Zambia demographic and health survey. BMC Pregnancy Childbirth [Internet]. 2025 Oct 2 [cited 2025 Oct

- 5];25(1):1005. Available from: <https://doi.org/10.1186/s12884-025-08063-4>
11. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. *Perspect Clin Res* [Internet]. 2017 [cited 2025 Oct 5];8(3):148–51. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5543767/>
 12. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina (Mex)* [Internet]. 2020 Sept 8 [cited 2025 Oct 5];56(9):455. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7560135/>
 13. Palomares-Salas JC, Aguado-González S, Sierra-Fernández JM. Robustness of Machine Learning and Deep Learning Models for Power Quality Disturbance Classification: A Cross-Platform Analysis. *Appl Sci* [Internet]. 2025 Jan [cited 2025 Oct 5];15(19):10602. Available from: <https://www.mdpi.com/2076-3417/15/19/10602>.
 14. Liu L, Jiao Y, Li X, Ouyang Y, Shi D. Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. *Comput Methods Programs Biomed* [Internet]. 2020 Nov 1 [cited 2025 Oct 5];196:105624. Available from: <https://www.sciencedirect.com/science/article/pii/S0169260720314577>
 15. Morang'a CM, Amenga-Etego L, Bah SY, Appiah V, Amuzu DSY, Amoako N et al. Machine learning approaches classify clinical malaria outcomes based on haematological parameters. *BMC Med* [Internet]. 2020 Nov 30 [cited 2025 Mar 29];18(1):375. Available from: <https://doi.org/10.1186/s12916-020-01823-3>
 16. Islam MN, Mustafina SN, Mahmud T, Khan NI. Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. *BMC Pregnancy Childbirth* [Internet]. 2022 Apr 22 [cited 2025 Oct 5];22:348. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9097057/>
 17. Mamo DN, Gebremariam YH, Adem JB, Kebede SD, Walle AD. Machine learning to predict unintended pregnancy among reproductive-age women in Ethiopia: evidence from EDHS 2016. *BMC Womens Health* [Internet]. 2024 Jan 23 [cited 2025 Oct 5];24:57. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10804521/>
 18. Setegn GM, Dejene BE. Explainable artificial intelligence models for predicting pregnancy termination among reproductive-aged women in six east African countries: machine learning approach. *BMC Pregnancy Childbirth* [Internet]. 2024 Sept 16 [cited 2025 Oct 5];24(1):600. Available from: <https://doi.org/10.1186/s12884-024-06773-9>
 19. Pi X, Wang J, Chu L, Zhang G, Zhang W. Prediction of high-risk pregnancy based on machine learning algorithms. *Sci Rep* [Internet]. 2025 May 4 [cited 2025 Oct 5];15(1):15561. Available from: <https://www.nature.com/articles/s41598-025-00450-3>
 20. Sylvain MH, Nyabyenda EC, Uwase M, Komezusenge I, Ndikumana F, Ngaruye I. Prediction of adverse pregnancy outcomes using machine learning techniques: evidence from analysis of electronic medical records data in Rwanda. *BMC Med Inform Decis Mak* [Internet]. 2025 Feb 12 [cited 2025 Oct 5];25(1):76. Available from: <https://doi.org/10.1186/s12911-025-02921-z>
 21. Yehuala TZ, Fente BM, Wubante SM. Exploring machine learning algorithms to predict short birth intervals and identify its determinants among reproductive-age women in East Africa. *BMC Pregnancy Childbirth* [Internet]. 2025 May 9 [cited 2025 Oct 5];25(1):551. Available from: <https://doi.org/10.1186/s12884-025-07668-z>
 22. Corsi DJ, Neuman M, Finlay JE, Subramanian SV. Demographic and health surveys: a profile. *Int J Epidemiol*. 2012;41(6):1602–13.
 23. Pedregosa F, Pedregosa F, Varoquaux G, Varoquaux G, Org N, Gramfort A et al. Scikit-learn: Machine Learning in Python. *Mach Learn PYTHON*.
 24. Kuhn M, Johnson K. *Applied Predictive Modeling* [Internet]. New York, NY: Springer New York; 2013 [cited 2026 Jan 21]. Available from: <http://link.springer.com/10.1007/978-1-4614-6849-3>
 25. Berrar D. (2019) Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, 1, 542–545. - References - Scientific Research Publishing [Internet]. [cited 2025 Sept 9]. Available from: <https://www.scirp.org/reference/referencespapers?referenceid=3814225>
 26. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13(null):281–305.
 27. Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC Med Res Methodol*. 2021 July 31;21(1):159.

28. 2023 Annual Results and, Phase II. (2020–2023) Report | Amplifying Change: Harnessing Collective Power to End Child Marriage | UNICEF [Internet]. [cited 2025 Oct 24]. Available from: <https://www.unicef.org/aboutgp2023AAR>
29. Kabelka H, Kok M. Drivers of Child Marriage in Eastern and Western provinces in Zambia and effective preventive interventions.
30. Yakubu I, Salisu WJ. Determinants of adolescent pregnancy in sub-Saharan Africa: a systematic review. *Reprod Health*. 2018;15(1):15.
31. Kassa GM, Arowojolu AO, Odukogbe AA, Yalew AW. Prevalence and determinants of adolescent pregnancy in Africa: a systematic review and Meta-analysis. *Reprod Health*. 2018;15(1):195.
32. Mmari K, Sabherwal S. A review of risk and protective factors for adolescent sexual and reproductive health in developing countries: an update. *J Adolesc Health Off Publ Soc Adolesc Med*. 2013;53(5):562–72.
33. Malunga G, Sangong S, Saah FI, Bain LE. Prevalence and factors associated with adolescent pregnancies in Zambia: a systematic review from 2000–2022. *Arch Public Health*. 2023;81:27.
34. Mbele W, Jassey B, Nyirenda N. Factors associated with adolescent pregnancy in Zambia: a secondary data analysis of the 2018 Zambia demographic and health survey. *BMC Pregnancy Childbirth*. 2025;25:1005.
35. Hindin MJ, Kalamar AM, Thompson TA, Upadhyay UD. Interventions to Prevent Unintended and Repeat Pregnancy Among Young People in Low- and Middle-Income Countries: A Systematic Review of the Published and Gray Literature. *J Adolesc Health Off Publ Soc Adolesc Med*. 2016 Sept;59(3 Suppl):S8–15.
36. (PDF) Prevalence and associated factors of adolescent pregnancy (15–19 years) in East Africa: a multilevel analysis [Internet]. [cited 2025 Oct 24]. Available from: https://www.researchgate.net/publication/350416501_Prevalence_and_associated_factors_of_adolescent_pregnancy_15-19_years_in_East_Africa_a_multilevel_analysis
37. Ayele BG, Gebregzabher TG, Hailu TT, Assefa BA. Determinants of teenage pregnancy in Degua Tembien District, Tigray, Northern Ethiopia: A community-based case-control study. *PLoS ONE*. 2018;13(7):e0200898.
38. A multilevel analysis of trends and predictors associated with teenage pregnancy in Zambia. (2001–2018) | Reproductive Health | Full Text [Internet]. [cited 2025 Oct 24]. Available from: <https://reproductive-health-journal.biomedcentral.com/articles/10.1186/s12978-023-01567-2>
39. Neal S, Matthews Z, Frost M, Fogstad H, Camacho AV, Laski L. Childbearing in adolescents aged 12–15 years in low resource countries: a neglected issue. New estimates from demographic and household surveys in 42 countries. *Acta Obstet Gynecol Scand*. 2012 Sept;91(9):1114–8.
40. Adolescent pregnancy | United Nations Population Fund [Internet]. [cited 2025 Oct 24]. Available from: <https://www.unfpa.org/adolescent-pregnancy>
41. Zulu T, Musukuma M, Jacobs C, Musonda P. Factors associated with the ability of adolescent girls and young women (AGYW) in sexual unions to negotiate for safer sex. An analysis of data from the 2018 Zambia Demographic and Health Survey (ZDHS). *PLOS Glob Public Health* 2024 Sept 4;4(9):e0003614.
42. Phiri M, Kasonde ME, Moyo N, Sikaluzwe M, Simona S. A multilevel analysis of trends and predictors associated with teenage pregnancy in Zambia (2001–2018). *Reprod Health*. 2023;20:16.

Figures

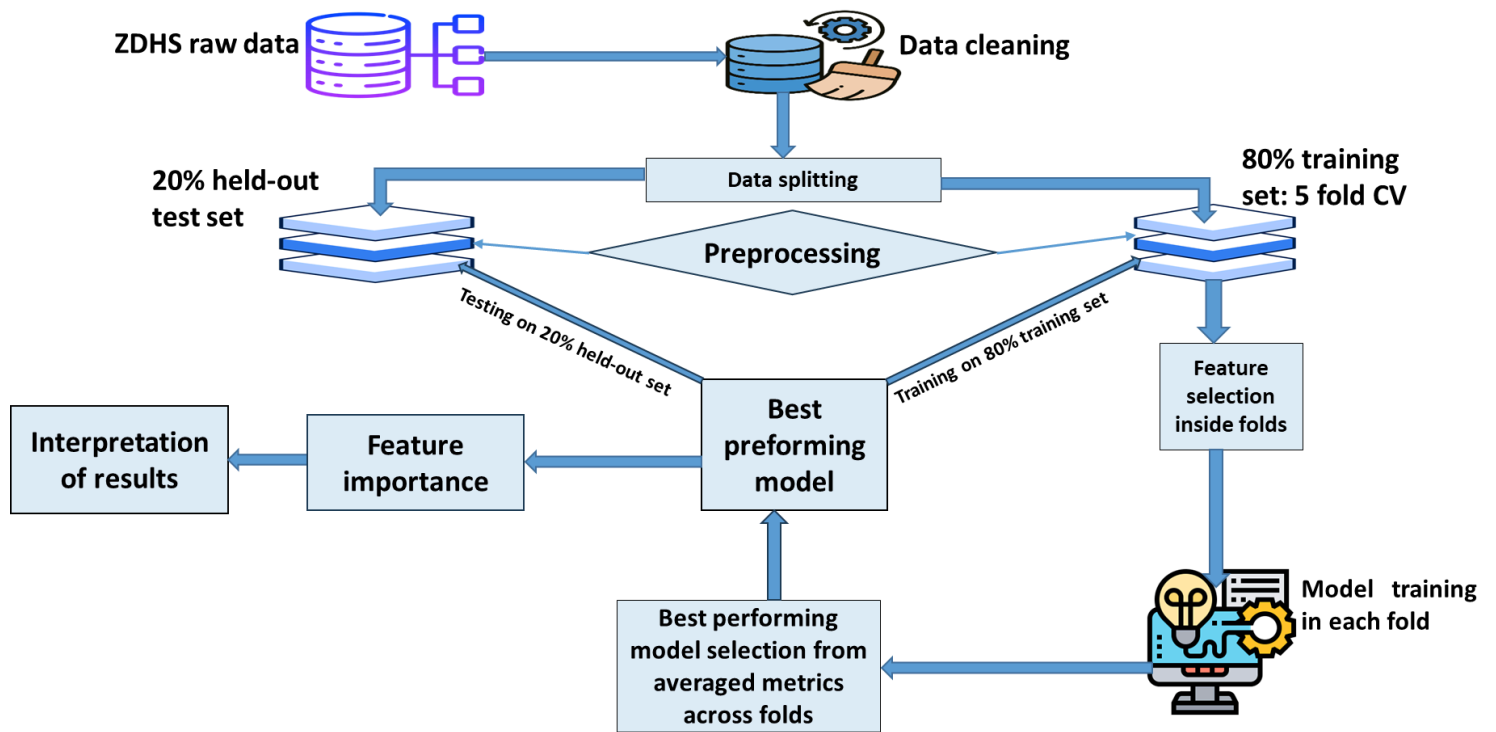


Figure 1

Machine learning modelling workflow for predicting teenage pregnancy

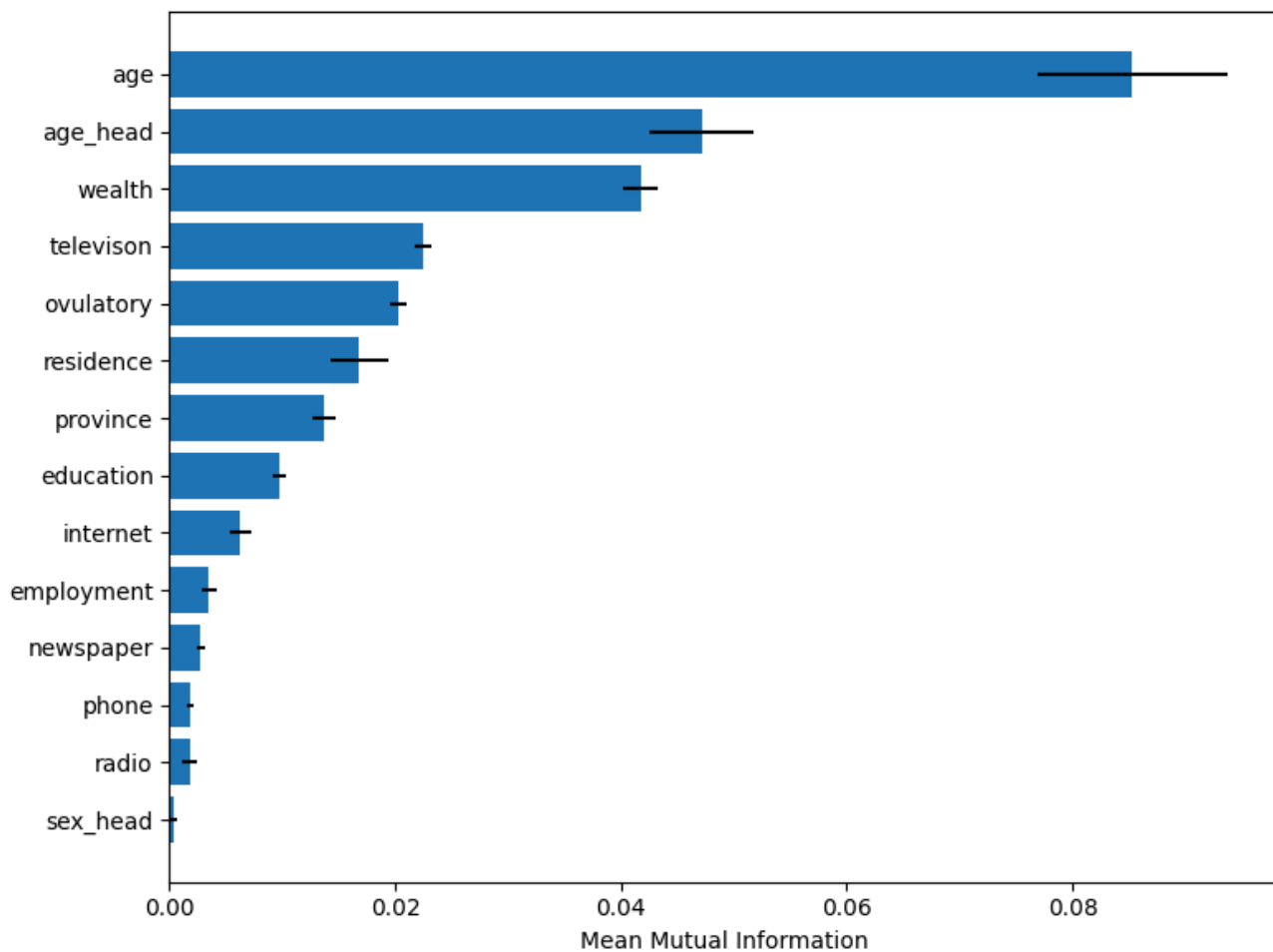


Figure 2

Mutual information–based feature importance for predictors of teenage pregnancy

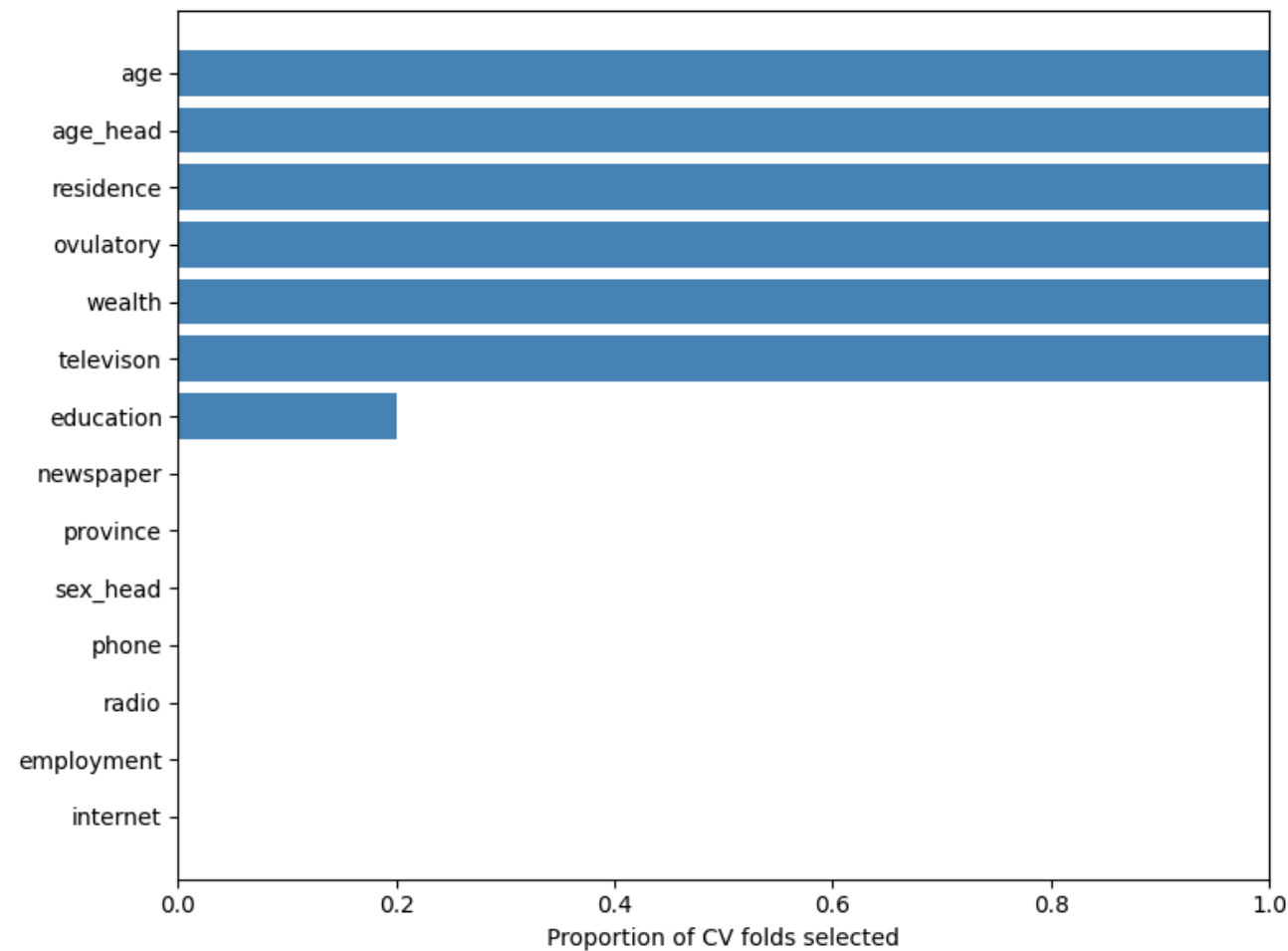


Figure 3

The Boruta algorithm stability of feature selection across cross-validation folds. Feature selection was performed independently within each CV fold to avoid data leakage. Proportions represent the frequency with which a variable was retained to be relevant across all folds.

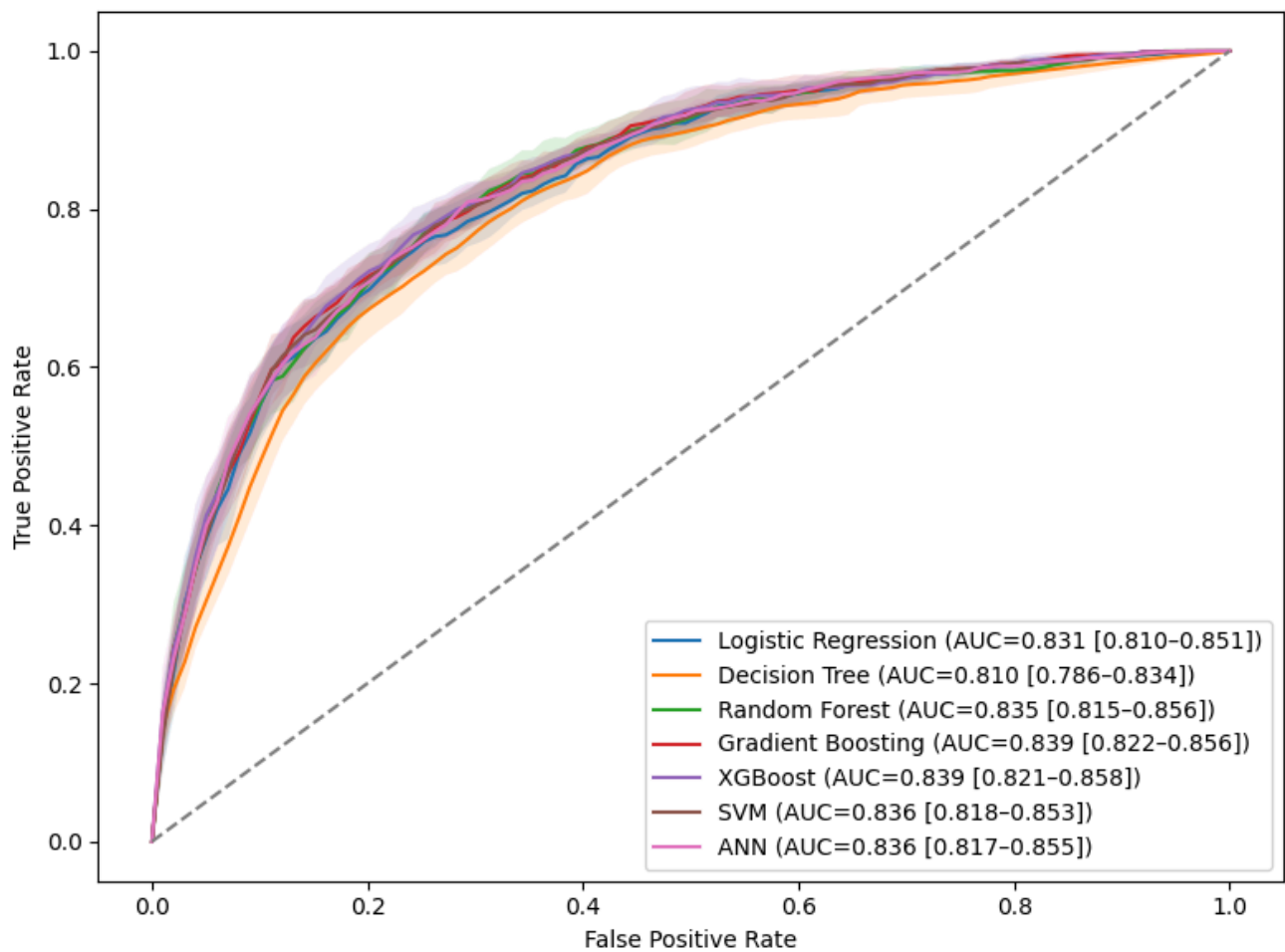


Figure 4

AUC-ROC curves for seven Machine learning models implemented inside CV. Shaded areas represent 95% confidence intervals for the AUC. The dashed diagonal line indicates no-discrimination performance (AUC = 0.50)

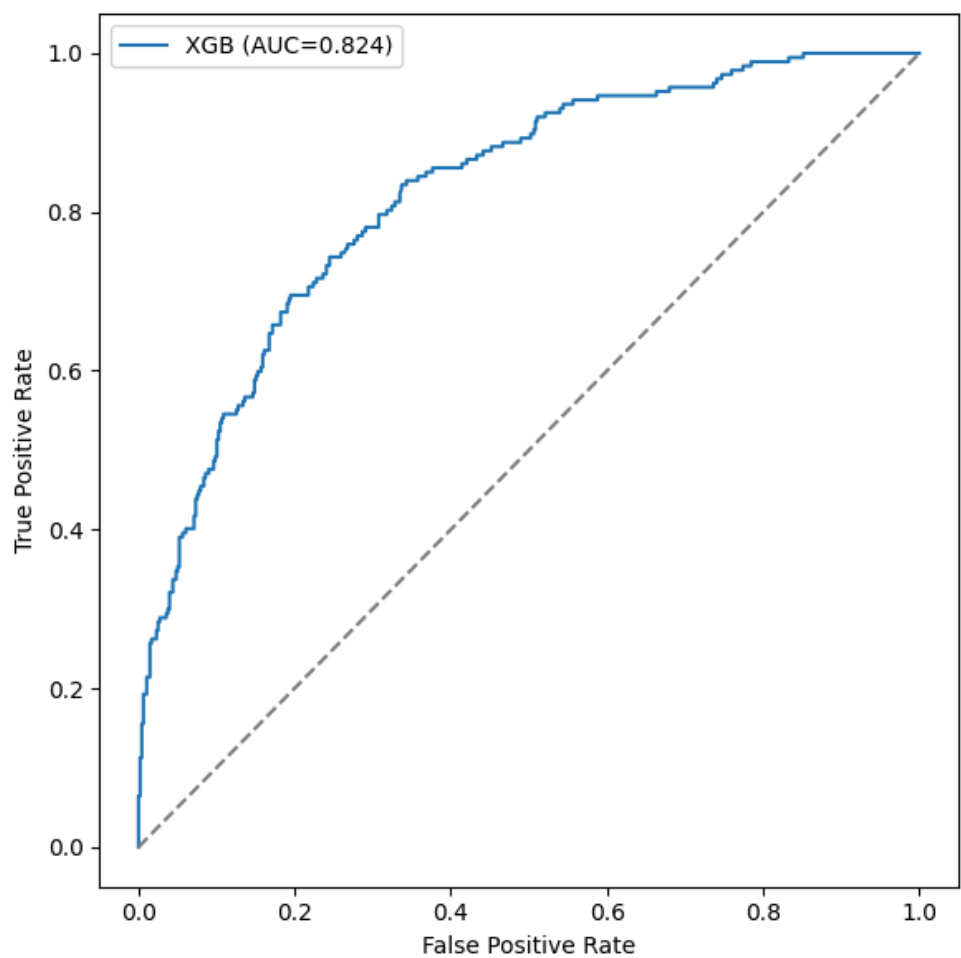


Figure 5

Receiver operating characteristic curve for the XGB model evaluated on the held-out test set

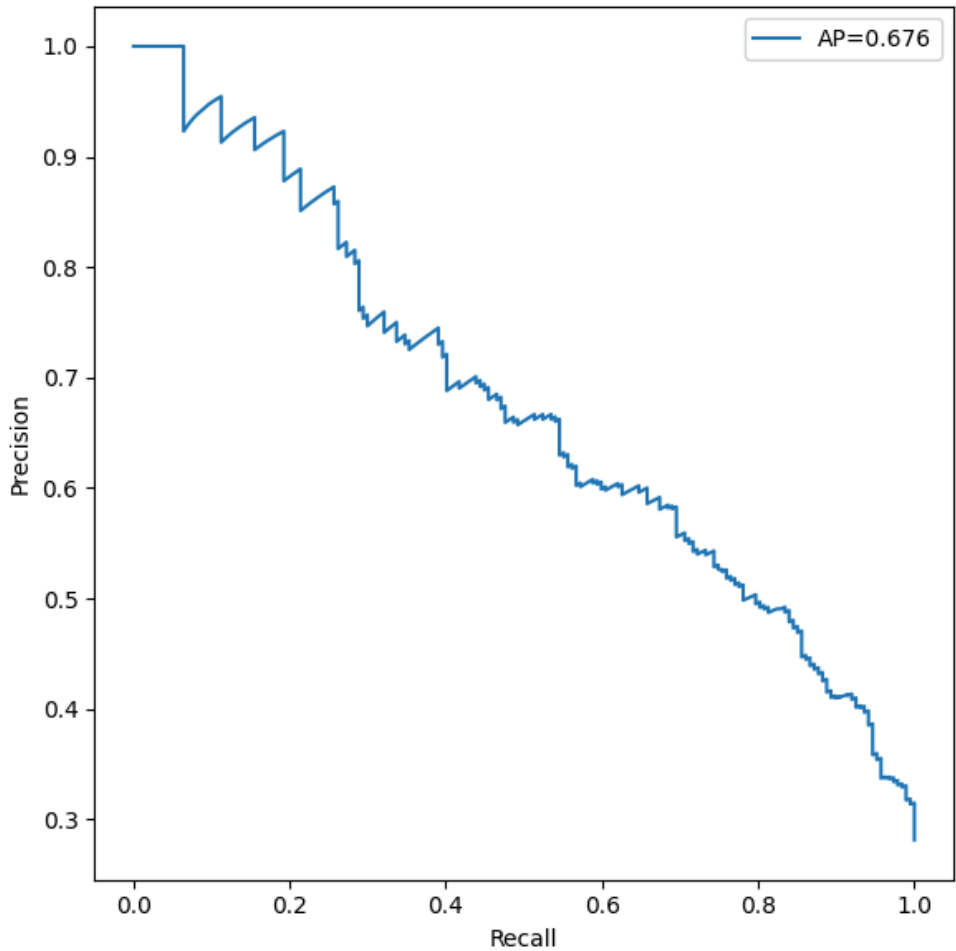


Figure 6

Precision–recall curve for the XGB model evaluated on the held-out test set

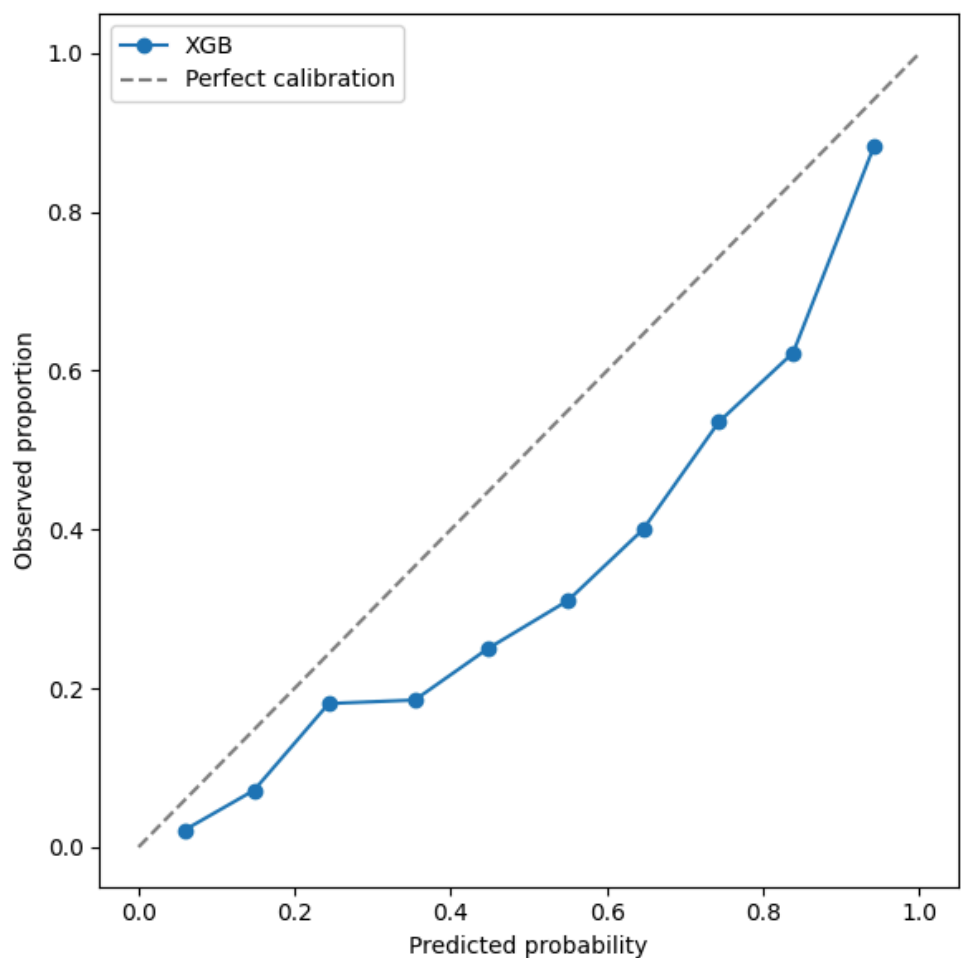


Figure 7

Calibration plot for the XGB model. The figure shows the relationship between predicted probabilities and observed outcome proportions for the XGB model on the held-out test set. The dashed diagonal line represents perfect calibration

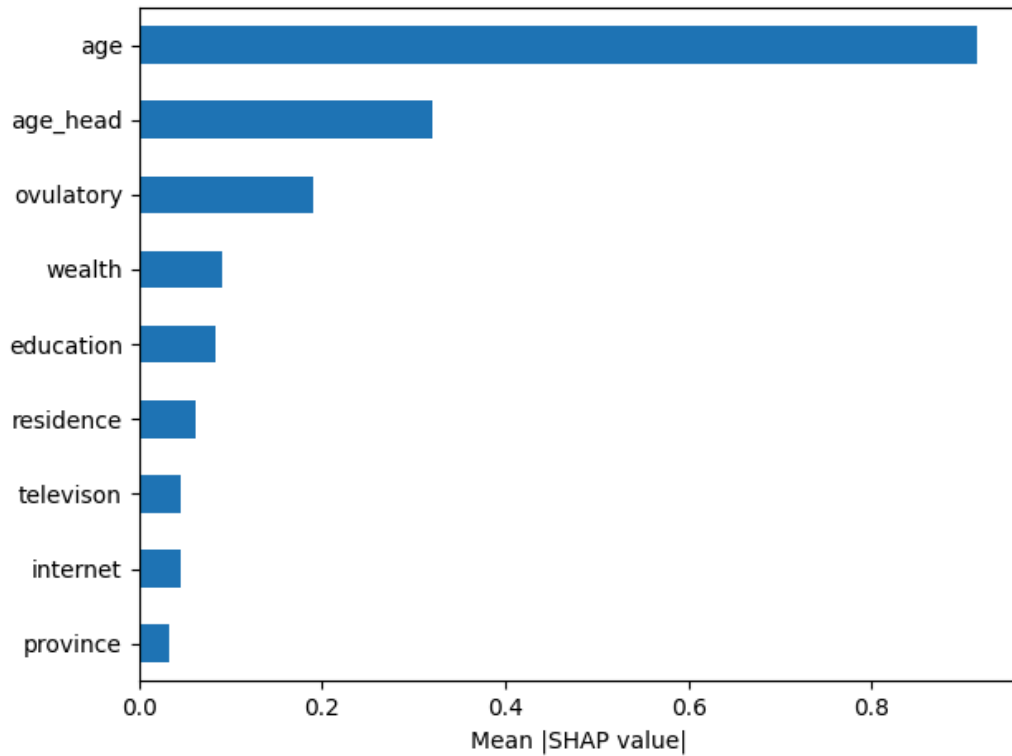


Figure 8

Grouped SHAP feature importance for the XGB model. The figure shows the mean absolute SHAP values for each predictor, summarizing their overall contribution to the XGB model predictions on the held-out test set. Higher values indicate greater importance