International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 25(2), November, 2025, pp 81-99
DOI: https://doi.org/10.3329/ijss.v25i2.85769

Identifying Risk Factors of Early Marriage among Women in Bangladesh Using Machine Learning Algorithms

Md. Fahim, Papia Sultana, Md. Rezaul Karim, Dulal Chandra Roy and Md. Mahfuz Uddin*

Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
*Correspondence should be addressed to Md. Mahfuz Uddin
(Email: mahfuz.ru.stat.58@gmail.com)

[Received July 28, 2025; Accepted September 10, 2025]

Abstract

Early marriage, defined as marriage before age 18, is a human rights violation with serious consequences for women's health and well-being, and remains a major public health issue, particularly in South Asia and Bangladesh. The objective of this study is to identify the key socio-demographic and household decision making risk factors associated with early marriage among women in Bangladesh by applying various machine learning algorithms, and to evaluate the predictive performance of these models for effective policy formulation using data from the nationally representative BDHS 2022. Chi-square tests assessed associations between respondent characteristics and early marriage, while three advanced feature selection methods Boruta, LASSO, and Information Gain were employed for selection of relevant features. Eight machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, AdaBoost, LogitBoost, and Neural Network, were evaluated using 5-fold cross-validation. Model performance was assessed by sensitivity, specificity, precision, accuracy, FDR, and AUC. The analysis revealed that the prevalence of early marriage was 67.02%. Combining significance tests and feature selection, Division, Wealth Index, Reading Newspaper, Religion, Residence, Household Purchases, and Age consistently emerged as the most influential predictors. Among all models, Decision Tree provided the best balanced performance on the testing set (sensitivity: 0.248, specificity: 0.896, precision: 0.540, accuracy: 0.682, AUC: 0.690), indicating its suitability for generalizable early marriage prediction. Feature importance analysis highlighted Wealth Index and Division as primary drivers. This study guides policymakers to target interventions by pinpointing high-risk regions and socioeconomic groups driving early marriage. Strengthening girls' education, economic support, and community awareness can effectively reduce its prevalence in Bangladesh.

Keywords: Early marriage, Feature selection, Machine learning algorithms, Socio-economic and household decision making factors, Bangladesh

AMS Classification: 62P25, 68T09.

1. Introduction

Early marriage, defined as marriage before the age of 18, remains a critical issue globally with more than 650 million women alive today who were married as children[1,2]. This practice disproportionately affects girls in developing countries, where poverty, low education, and deep-rooted gender norms make them especially vulnerable. South Asia alone accounts for nearly 45% of all child brides, and Bangladesh ranks among the highest in the region[3]. It remains a critical human rights violation and public health concern, particularly in South Asian countries like Bangladesh. It contributes to early pregnancies and adverse maternal and child health outcomes. Despite ongoing efforts, the practice remains widespread due to varying socio-economic and demographic factors. Early marriage (EM) and early childbearing (ECB) remained pressing public health and social concerns in South Asia, with serious implications for women's health, education, and overall well-being. The study[4] analyzed nationally representative data from multiple DHS rounds (2017-18 and earlier) across Bangladesh, Nepal, India, and Pakistan, revealing that early marriage (EM) and early childbirth (ECB) remain prevalent especially among poor, uneducated rural women despite some progress. Understanding the drivers of change was essential for designing effective multi-sectoral strategies to eliminate these practices and improve maternal and child health outcomes[5]. This systematic review, following PRISMA guidelines, analyzed studies from 2014 to 2024 across 26 countries on adolescent girls' role in child marriage decisions. It found that economic hardship, parental pressure, and social norms restrict girls' autonomy, while education and family support offer protection. Ending child marriage requires holistic strategies to empower girls and change harmful societal norms[6]. Early marriage limited women's autonomy and weakened their bargaining power within households, restricting their participation in key decision-making processes. The causal impact of age at marriage on women's empowerment in Bangladesh, using age at menarche as an instrumental variable. Drawing on nationally representative data, the analysis revealed that delaying marriage significantly enhanced women's empowerment both in domestic and productive spheres. Later marriage increases autonomy in agriculture, control over income and assets, freedom of movement, and decision-making power, mainly through better education and labor market participation[7]. However, most of the existing studies in Bangladesh utilized data from the 2014 and 2017 Demographic and Health Surveys (BDHS), which didn't adequately reflect the present situation[8,9,10]. There remains a significant lack of research of utilizing advanced feature selection techniques such as Boruta algorithm, Regularized Random Forest (RRF), LASSO, Information Gain, Relief, etc. Additionally, previous studies rarely used machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, AdaBoost, LogitBoost, and Neural Network, etc.[4,5,6,7,8,9,10]. In our study, we utilized data from Bangladesh Demographic and Health Survey (BDHS) 2022 and examined key factors that reflect the present context. We apply three advanced feature selection techniques including Boruta algorithm, LASSO, and Information Gain for selecting relevant features. We also apply eight machine learning (ML) algorithms including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, AdaBoost, LogitBoost, and Neural Network using 5-fold cross-validation to compare the predictive capabilities of each ML model. The objective of this study is to identify the key socio-demographic and household decision making risk factors associated with early marriage among women in Bangladesh by applying various machine learning algorithms, and to evaluate the predictive performance of these models for effective policy formulation using data from the nationally representative BDHS 2022.

2. Materials and Methods

2.1 Data source

This study utilizes data from the Bangladesh Demographic and Health Survey (BDHS) 2022, a nationally representative survey conducted by the National Institute of Population Research and Training (NIPORT) in collaboration with ICF. The BDHS 2022 provides detailed information on demographic, health, and social indicators, including women's age at first marriage and household decision-making power variables such as participation in decisions about health care, large household purchases, and visits to relatives. These variables enable the use of machine learning algorithms to identify key predictors of early marriage among women aged 15-49[11]. More details about the dataset can be accessed: https://www.dhsprogram.com/data/.

2.2 Data preprocessing

To ensure the quality and reliability of the analysis, the BDHS 2022 data underwent meticulous preprocessing. This involved excluding observations with missing or inconsistent values for key variables, removing outliers where applicable, and addressing any remaining missing data. As a result of these steps, the sample size was refined from the original 8,784 to 2,080 observations. This cleaned dataset was then utilized for subsequent statistical analyses to examine the prevalence and determinants of early marriage.

2.3 Outcome variable

We examined the prevalence of early marriage among women as the outcome variable, which is defined as a dichotomous variable (1 = yes, 0 = no). In the BDHS -2022, women were asked, "How old were you when you first started living with your husband?" The outcome variable was derived from this question and categorized as early marriage (1 = yes) if the women cohabited with their husband before turning 18 years old; otherwise, it was classified as not early marriage (0 = no).

2.4 Explanatory variables

The study takes into account some independent variables that can affect the outcome variable. The independent variables are all categorical. The overview of the independent factors is presented in Table 1.

Table 1: Description of explanatory variables included in the dataset

Explanatory variables	Description	Categories	Measurement scale
Socio-demographic variables			
Age	Age group of women respondents in years	15-19 20-24 25-29	Ordinal
Residence	Type of residence of women	Urban Rural	Nominal
Division	Geographical division of the respondent's residence	Barishal Chattogram Dhaka Khulna Mymensingh Rajshahi Rangpur Sylhet	Nominal

Religion	Religious affiliation	Islam Hinduism Buddhist Christianity	Nominal
Reading newspaper	Frequency of newspaper reading	Rare or no exposure Regular media exposure	Nominal
Watching television	Frequency of watching television	Rare or no exposure Regular media exposure	Nominal
Sex of household head	Gender of the head of the household	Male Female	Nominal
Husband education level	Educational level of the husband	Low education High education	Nominal
Respondent currently working	Employment status of the respondent at the time of the survey	Yes No	Nominal
Husband's age	Categorical variable representing the husband's age in years	≤24 25-34 ≥35	Ordinal
Highest educational level	Highest level of education achieved by the respondent	Low education High education	Nominal
Wealth index	Composite measure of household socioeconomic status	Poorest Poorer Middle Richer Richest	Ordinal
Household decision-making variables			
Respondent's health care	Involvement of the woman in decisions regarding her own healthcare	Women not involved Women involved	Nominal
Household purchases	Respondent's participation in major household purchase decisions	Women not involved Women involved	Nominal
Visits to family or relatives	Involvement in decision-making regarding visits to family or relatives	Women not involved Women involved	Nominal

2.5 Statistical methods

Descriptive analysis is conducted to examine the characteristics of the respondents. Chi-square tests are employed to assess the association between respondents' characteristics and early marriage. Three important features selection techniques e.g., Boruta algorithm, LASSO, and Information Gain, are employed for selecting important features to produce better performance in the model. Eight machine learning models such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Extreme Gradient Boosting, AdaBoost, LogitBoost, and Neural

Network are applied to predict early marriage and its associated risk factors among women in Bangladesh. All statistical analyses are performed using the R-programming language.

2.6 Feature selection

2.6.1 Boruta algorithm

The Boruta algorithm, based on Random Forest, is a notable method of feature selection [12]. It works by introducing shadow features randomized duplicates of the original variables into decision tree models. It evaluates feature importance by measuring how much these shadow features reduce the model's predictive performance. To quantify this, it computes Z-scores by dividing the average decrease in accuracy by its standard deviation, making this Z-score the primary indicator of a variable's relevance. The distribution of importance among shadow features serves as a benchmark to assess the true variables. Ultimately, each original feature's importance is compared to the maximum importance observed among the shadow features to identify which predictors are truly significant [13]. Required steps of this algorithm is given in[14].

2.6.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is a regularization method that serves both feature selection and coefficient shrinkage in predictive modeling. It enhances the loss function by adding a penalty proportional to the sum of the absolute values of the coefficients, effectively placing an upper limit on this total. By applying this penalty, LASSO drives some regression coefficients toward zero, with many ultimately reduced exactly to zero. As a result, only variables with non-zero coefficients after this shrinkage process are retained in the model, effectively performing variable selection. This strategy helps to minimize prediction error by discouraging overfitting. More details are outlined in[15,16].

2.6.3 Information Gain

Information Gain (IG) is an entropy-based method widely applied for feature selection in classification problems. It evaluates how much the uncertainty (entropy) of the outcome variable decreases when a particular feature is taken into account. Essentially, IG quantifies the improvement in predicting the target achieved by knowing the value of a given predictor[17]. The procedure starts by determining the overall entropy of the target variable, which reflects its inherent disorder in the absence of any predictors. Subsequently, for each feature, the conditional entropy of the target given that feature is calculated, allowing the reduction in uncertainty attributable to each feature to be assessed. Additional insights are discussed in[18,19].

2.7 Machine learning algorithms

2.7.1 Logistic Regression

Logistic regression is a supervised machine learning technique used for classification tasks where the primary goal is to ascertain whether an instance belongs to a specific class or not. Binary classification is accomplished via the sigmoid function, which takes input as independent variables and produces a probability value between 0 and 1. Using logistic regression, a function from the dataset's attributes is mapped to the targets in order to determine the probability that a new example will belong to one of the target classes[20]. More information can be found in[21,22].

2.7.2 Decision Tree

A decision tree is a supervised learning method employed for both classification and regression tasks. Structurally resembling a flowchart, it facilitates decision-making and prediction by systematically splitting data[23]. The tree comprises internal nodes that represent tests or

conditions on features, branches that denote the outcomes of these tests, and terminal leaf nodes that provide the final prediction either a class label or a numerical value. The root node encapsulates the entire dataset and initiates the first split. From there, each internal node applies a decision rule, with branches directing to subsequent nodes based on the outcome. The process continues until reaching leaf nodes, where no further splitting occurs, delivering the ultimate prediction. In-depth steps are given in[24].

2.7.3 Random Forest

Random forest is a supervised machine learning method that may be applied to challenges involving both classification and regression[25]. It constructs multiple decision trees and aggregates their output to increase forecast accuracy and stability. It uses bootstrap sampling, which uses replacement random sampling to produce several subsets of the original dataset. Instead of using all available features, a random subset of attributes is chosen at each split to reduce the correlation between trees[26]. Each decision tree is then constructed according to a preset splitting criterion. Comprehensive details are provided in[27].

2.7.4 Gradient Boosting

Gradient Boosting is a robust machine learning technique primarily applied to solve regression and classification problems. It sequentially combines numerous weak models, usually decision trees, to create a strong prediction model. Using a technique known as gradient descent, each new model is trained to rectify the mistakes caused by the earlier models by minimizing a loss function, such as the mean squared error or cross-entropy of the prior model[28]. It frequently produces very high accuracy, is quite adaptable, and can handle a variety of data formats. More information is covered in[29,30].

2.7.5 Extreme Gradient Boosting (XGBoost)

XGBoost is a popular machine learning method that minimizes loss functions used for both classification and regression tasks. An ensemble of decision trees is progressively generated, each of which fixes the errors of the one before it. To increase efficiency and decrease overfitting, it adds a number of improvements, including regularization, effective computing, and highly complex optimization methods[31]. There is a detailed explanation in[32,29].

2.7.6 Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble machine learning algorithm that builds a sequence of weighted decision trees, typically using simple one-level trees known as “stumps”. This approach incrementally combines these weak learners to improve overall predictive performance. Each tree is trained on the entire dataset, but adaptive sample weights are utilized to add weight to samples that were previously erroneously classified. High-performing trees have more influence on the results since it aggregates the trees for classification tasks using a weighted voting technique. The strength of the model lies in its adaptive learning process; whilst each basic tree may be a "weak learner" that marginally outperforms random guessing, the weighted collection of trees becomes a "strong learner" that progressively focuses on and corrects errors[33]. Refer to[34,35] for a more thorough discussion.

2.7.7 LogitBoost

LogitBoost is a boosting technique that tackles classification problems by incrementally fitting additive logistic regression models in a stage-wise manner. With the logistic loss function, it works

especially well for binary classification[36]. It is called the Additive Logistic Regression Model. The core idea of LogitBoost is to use boosting when building a logit model. Because it continuously employs different training examples, it is classified as a "weak" or "base" learning algorithm. This results in several rounds because the base learning algorithm generates a new weak prediction rule. These weak rules must then be combined via the boosting algorithm to create a single strong prediction rule, which is usually significantly more accurate than a weak rule. Additional details are available in[37].

2.7.8 Neural Network

Neural Networks are machine learning models modeled after the architecture and operations of the human brain, designed to capture complex patterns in data. Each layer is made up of interconnected nodes, or neurons, which process input data before sending the results to the layer below. An input layer, an output layer, and one or more hidden layers are common components of neural networks. In order to decrease the error between anticipated and actual values, the model learns by modifying the weights of the connections between neurons through a process known as backpropagation[37]. Because of their great adaptability, neural networks can identify intricate, non-linear relationships in data. The sigmoid, ReLU, Tanh, and softmax activation functions are frequently used in it[38]. Details explanation can be found in[39,40].

2.8 Model evaluation metrics

The model evaluation metrics, along with their computation formulae, are briefly described in Table 2. In this table, TP = True Positive, FN = False Negative, FP = False Positive, TN = True Negative[41,42].

Table 2: Model evaluation metrics

Metrics	Formula	Description
Sensitivity	$\frac{TP}{TP + FN}$	Indicates the proportion of true positives accurately identified.
Specificity	$\frac{TN}{FP + TN}$	Indicates the proportion of true negatives accurately identified.
Precision	$\frac{TP}{TP + FP}$	Determines how many predicted positives are actually correct.
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	Calculates the overall correctness of the model.
FDR	$\frac{FP}{TP + FP}$	Determines the proportion of false positives among predicted positives.

3. Results

3.1 Preliminary data analysis

Table 3 presents the socio-demographic and household decision making related characteristics of the respondents. The majority of participants were aged 25-29 years (40.96%), followed by 20-24 years (35.87%). Over two-thirds of respondents (65.72%) resided in rural areas, and the highest

proportion came from the Dhaka (15.29%) and Chattogram (14.37%) divisions. Most participants identified as Muslim (90.48%), with small proportions adhering to Hinduism (8.51%) or other religions. Exposure to mass media was limited: 97.36% reported little or no exposure to newspapers, and 52.07% had limited access to television. Women's involvement in household decision-making varied: 72.64% were involved in healthcare decisions, 67.50% in household purchases, and 74.57% reported freedom to visit family or relatives. However, only 24.66% of women were currently working. Male-headed households predominated (84.23%). Educational attainment was low among women, with 81.54% having low education, compared to 43.89% of husbands. Most husbands (67.60%) were aged 35 years or older. The wealth distribution was relatively balanced across quintiles, although the highest proportions were in the richest (20.96%) and middle (20.72%) categories. Notably, 67.02% of women reported experiencing early marriage, underscoring the continued prevalence of this practice.

Table 3: Characteristics of the study subjects

Covariates	Frequency	Percent frequency
Socio-demographic variables		
Age of women		
15-19	482	23.17
20-24	746	35.87
25-29	852	40.96
Residence of women		
Urban	713	34.28
Rural	1367	65.72
Division		
Barishal	203	9.76
Chattogram	299	14.37
Dhaka	318	15.29
Khulna	268	12.88
Mymensingh	220	10.58
Rajshahi	276	13.27
Rangpur	266	12.79
Sylhet	230	11.06
Religion		
Islam	1882	90.48
Hinduism	177	8.51
Buddhist	17	0.82
Christianity	4	0.19
Reading newspaper		
Rare or no exposure	2025	97.36
Regular media exposure	55	2.64
Watching television		
Rare or no exposure	1083	52.07
Regular media exposure	997	47.93
Sex of household head		
Male	1752	84.23
Female	328	15.77
Husband education level		
Low education	913	43.89
High education	1167	56.11

Respondent currently working		
No	1567	75.34
Yes	513	24.66
Husband's age		
<=24	104	5.00
25-34	570	27.40
35 and above	1406	67.60
Highest educational level		
Low education	1696	81.54
High education	384	18.46
Wealth index		
Poorest	392	18.85
Poorer	418	20.09
Middle	431	20.72
Richer	403	19.38
Richest	436	20.96
Household decision-making variables		
Respondent's health care		
Woman not involved	569	27.36
Woman involved	1511	72.64
Household purchases		
Woman not involved	676	32.50
Woman involved	1404	67.50
Visits to family or relatives		
Woman not involved	529	25.43
Woman involved	1551	74.57
Outcome variable		
Early Marriage		
No	686	32.98
Yes	1394	67.02

[Wealth Index was calculated by the Demographic and Health Surveys (DHS) Program using Principal Component Analysis (PCA) on household asset ownership, housing characteristics, and access to basic services. Households were then ranked and categorized into five wealth quintiles: Poorest, Poorer, Middle, Richer, and Richest.]

Table 4 presents the bivariate associations between respondents' characteristics and early marriage. Statistically significant associations (p-value < 0.05) were observed for several variables. Age was significantly associated with early marriage (χ^2 = 11.450, p-value = 0.003). Women currently aged 25-29 years reported the highest proportion of early marriage (43.11%), followed by those aged 20-24 (33.50%). Place of residence showed a strong association (χ^2 = 24.473, p-value < 0.001); early marriage was more prevalent among rural women (69.37%) than urban counterparts (30.63%). Geographic division also demonstrated a significant association (χ^2 = 87.942, p-value < 0.001). Higher rates of early marriage were observed in Dhaka (15.28%), Rajshahi (15.06%), and Khulna (14.56%), while lower rates were reported in Sylhet (7.53%) and Mymensingh (10.12%). Religion was significantly related to early marriage (χ^2 = 17.534, p-value = 0.001), with the highest prevalence among Muslim women (92.04%) and the lowest among Christians (0.07%). Media exposure, particularly through newspapers, was significantly associated (χ^2 = 25.465, p-value < 0.001); early marriage was more common among women with rare or no exposure (98.64%). Household decision-making on purchases was significantly associated with early marriage (χ^2 = 6.989, p-value = 0.008), with higher prevalence among women not involved in

such decisions. Wealth status exhibited a strong and significant association ($\chi^2 = 67.945$, p-value < 0.001). Women from the middle quintile experienced the highest rate of early marriage (21.74%), whereas the lowest prevalence was observed among the richest (15.92%). No statistically significant associations were found for watching television (p-value = 0.066), sex of household head (p-value = 0.766), women's involvement in healthcare decisions (p-value = 0.216), visits to family or relatives (p-value = 0.283), current employment status (p-value = 0.221), husband's age (p-value = 0.874), or educational levels of either respondent (p-value = 0.144) or husband (p-value = 0.598).

Table 4: Association analysis among different characteristics with early marriage

Covariates	Early Marriage		Chi-square value	P-value
	No (686) N (%)	Yes (1394) N(%)		
Socio-demographic variables				
Age of women				
15-19	156 (22.74)	326 (23.39)	11.450	0.003
20-24	279 (40.67)	467 (33.50)		
25-29	251 (36.59)	601 (43.11)		
Residence of women				
Urban	286 (41.69)	427 (30.63)	24.473	<0.001
Rural	400 (58.31)	967 (69.37)		
Division				
Barishal	52 (7.58)	151 (10.83)	87.942	<0.001
Chattogram	124 (18.08)	175 (12.56)		
Dhaka	105 (15.31)	213 (15.28)		
Khulna	65 (9.48)	203 (14.56)		
Mymensingh	79 (11.52)	141 (10.12)		
Rajshahi	66 (9.62)	210 (15.06)		
Rangpur	70 (10.20)	196 (14.06)		
Sylhet	125 (18.22)	105 (7.53)		
Religion				
Islam	599 (87.32)	1283 (92.04)	17.534	0.001
Hinduism	73 (10.64)	104 (7.46)		
Buddhist	11 (1.60)	6 (0.43)		
Christianity	3 (0.44)	1 (0.07)		
Reading newspaper				
Rare or no exposure	650 (94.75)	1375 (98.64)	25.465	<0.001
Regular media exposure	36 (5.25)	19 (1.36)		
Watching television				
Rare or no exposure	337 (49.13)	746 (53.52)	3.376	0.066
Regular media exposure	349 (50.87)	648 (46.48)		
Sex of household head				
Male	575 (83.82)	1177 (84.43)	0.088	0.766
Female	111 (16.18)	217 (15.57)		
Husband education level				
Low education	295 (43.00)	618 (44.33)	0.278	0.598
High education	391 (57.00)	776 (55.67)		
Respondent currently working				
No	505 (73.62)	1062 (76.18)	1.497	0.221
Yes	181 (26.38)	332 (23.82)		

Husband's age				
<=24	36 (5.25)	68 (4.88)		
25-34	191 (27.84)	379 (27.19)	0.269	0.874
35 and above	459 (66.91)	947 (67.93)		
Highest educational level				
Low education	572 (83.38)	1124 (80.63)	2.132	0.144
High education	114 (16.62)	270 (19.37)		
Wealth index				
Poorest	100 (14.58)	292 (20.95)		
Poorer	118 (17.20)	300 (21.52)		
Middle	128 (18.66)	303 (21.74)	67.945	<0.001
Richer	126 (18.37)	277 (19.87)		
Richest	214 (31.19)	222 (15.92)		
Household decision-making variables				
Respondent's health care				
Woman not involved	200 (29.15)	369 (26.47)	1.534	0.216
Woman involved	486 (70.85)	1025 (73.53)		
Household purchases				
Woman not involved	250 (36.44)	426 (30.56)	6.989	0.008
Woman involved	436 (63.56)	968 (69.44)		
Visits to family or relatives				
Woman not involved	185 (26.97)	344 (24.68)	1.154	0.283
Woman involved	501 (73.03)	1050 (75.32)		

3.2 Feature selection

Figure 1 presents the results of feature selection using three different methods: Boruta algorithm, LASSO regression, and Information Gain. Across all three methods, Division, Wealth Index, and Reading newspaper consistently emerged as the most influential predictors of early marriage. The Boruta algorithm confirmed these variables, along with Religion, Residence, and Household Purchases, as strongly relevant features. Variables such as Sex of household head, Husband's education, and current employment were rejected or found to have minimal influence. The LASSO method highlighted Religion, Division, and Reading newspaper as the top contributors based on absolute coefficient values. This method also emphasized Wealth Index and Age, though other variables contributed marginally. Using Information Gain, Division and Wealth Index again ranked highest, followed by Reading Newspaper and Religion, while most other features demonstrated low information scores. The consistency across methods underscores the critical role of geographic, socioeconomic, and informational factors in predicting early marriage, suggesting that interventions should prioritize these domains. By integrating the results from the significance test and multiple feature selection techniques, the variables Division, Wealth Index, Reading Newspaper, Religion, Residence, Household Purchases, and Age consistently emerged as both statistically significant and highly ranked. These variables represent the most robust predictors of early marriage across all analytical approaches and were used as input features in the application of machine learning algorithms, with Early Marriage defined as the outcome variable.

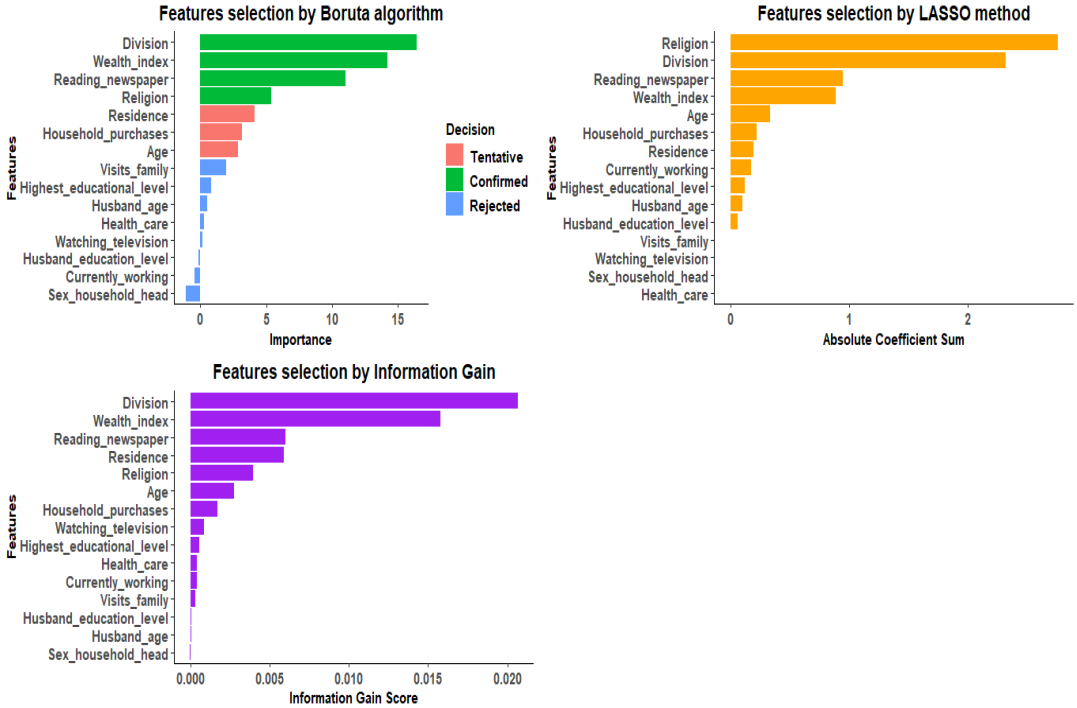

Figure 1: Features selection by Boruta, LASSO, and Information Gain

3.3 Evaluation of model performance

Table 5 presents the performance metrics of eight machine learning algorithms in predicting early marriage using the selected covariates in feature selection and significance test. Initially, the dataset was partitioned into training (80%) and testing (20%) subsets. The performance of each model was assessed based on both training and testing set. Overall, model performance varied across sensitivity, specificity, precision, accuracy, false discovery rate (FDR), and area under the curve (AUC), with notable discrepancies between training and testing sets. Among the models, AdaBoost demonstrated the highest training performance, with the highest sensitivity (0.617), precision (0.717), and AUC (0.883), and the best overall accuracy (0.793). However, its generalizability was limited on the testing set, where sensitivity dropped to 0.270 and accuracy to 0.617. Random Forest achieved strong training performance (AUC = 0.765), but also suffered from reduced sensitivity (0.153) and precision (0.467) on the testing set, suggesting possible overfitting. Similarly, Gradient Boosting and XGBoost showed reasonable performance on training data (AUC = 0.747 and 0.715, respectively), but their test performance declined, particularly in sensitivity and FDR. LogitBoost showed the highest specificity (0.977) and accuracy (0.727) on the testing set, despite a very low sensitivity (0.060). The Decision Tree and Logistic Regression models provided balanced yet moderate performance, with test accuracies of 0.682 and 0.670, respectively, and AUCs near 0.690. Finally, the Neural Network model offered relatively stable performance across both datasets, with test accuracy of 0.680 and AUC of 0.638, though sensitivity remained low (0.168). These findings suggest that while some models excel in training environments, their performance may not translate effectively to unseen data. Among all,

the Decision Tree model provided the best balanced performance between sensitivity (0.248), specificity (0.896), precision (0.540), accuracy (0.682), and AUC (0.690) on the testing set, indicating its suitability for generalizable early marriage prediction.

Table 5: Evaluation of model performance metrics for eight ML algorithms

Model names	Dataset	Performance metrics					
		Sensitivity	Specificity	Precision	Accuracy	FDR	AUC
Logistic Regression	Train	0.295	0.909	0.616	0.710	0.384	0.694
	Test	0.234	0.885	0.500	0.670	0.500	0.630
Decision Tree	Train	0.311	0.911	0.633	0.714	0.366	0.694
	Test	0.248	0.896	0.540	0.682	0.460	0.690
Random Forest	Train	0.293	0.952	0.749	0.735	0.251	0.765
	Test	0.153	0.914	0.467	0.663	0.533	0.625
Gradient Boosting	Train	0.357	0.895	0.626	0.718	0.374	0.747
	Test	0.212	0.863	0.433	0.648	0.567	0.545
XGBoost	Train	0.253	0.922	0.615	0.702	0.385	0.715
	Test	0.204	0.892	0.483	0.665	0.517	0.618
AdaBoost	Train	0.617	0.880	0.717	0.793	0.283	0.883
	Test	0.270	0.788	0.385	0.617	0.615	0.534
LogitBoost	Train	0.099	0.984	0.708	0.737	0.292	0.662
	Test	0.060	0.977	0.500	0.727	0.500	0.683
Neural Network	Train	0.199	0.952	0.669	0.703	0.331	0.685
	Test	0.168	0.932	0.548	0.680	0.452	0.638

Figure 2 highlights that on the training set, AdaBoost achieved the highest AUC (0.883), followed by Random Forest and Gradient Boosting, demonstrating strong in-sample fit. However, Figure 3 shows a shift on the testing set, where the Decision Tree (AUC = 0.690) performed best, suggesting greater stability and generalization. These results indicate that while complex ensemble methods excelled on training data, simpler models maintained more reliable predictive power on unseen data. The Decision Tree model is the best overall performer, offering the most balanced and interpretable performance for predicting early marriage.

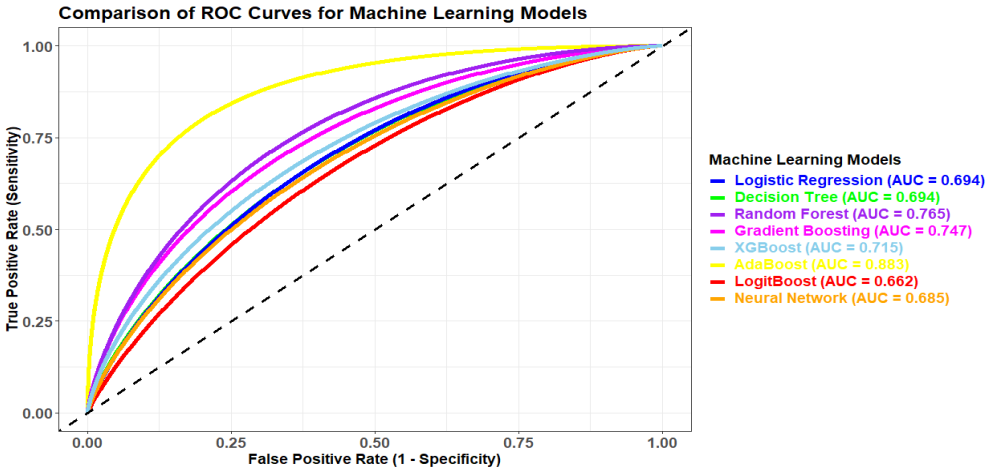

Figure 2: Comparison of ROC curves of eight ML models using training set

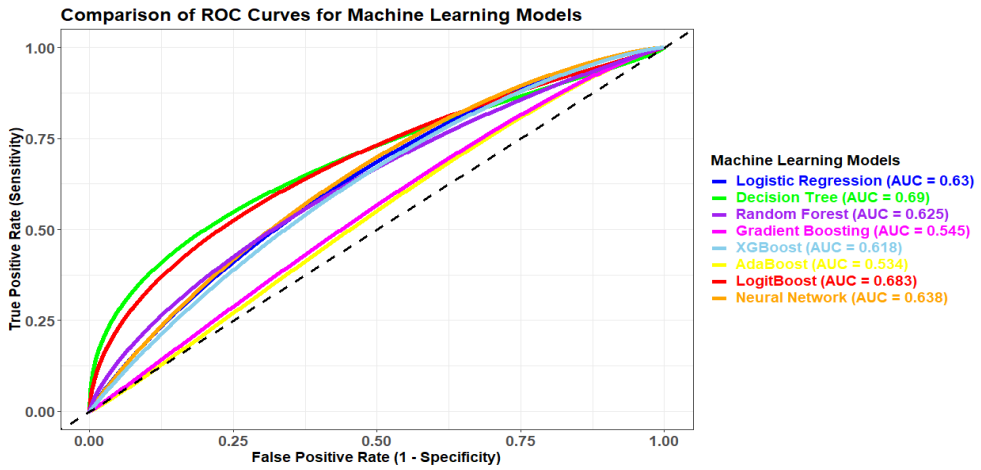

Figure 3: Comparison of ROC curves of eight ML models using testing set

3.4 Covariates effect and contribution for Decision Tree model

The Figure 4 illustrates the relative importance and contribution of covariates in predicting early marriage based on the Decision Tree model. The Wealth index emerged as the most influential predictor, closely followed by Division, indicating that economic status and regional disparities are critical determinants of early marriage in the study population. Age of the respondent also showed substantial predictive power, reflecting its direct relationship with marriage timing. Other important covariates included Residence, Religion, Household purchases decision-making, and Reading newspaper, though their contributions were comparatively modest. Overall, the model highlights the dominant role of socioeconomic and demographic factors, emphasizing the need to target wealth-related and regional inequalities in policies aiming to reduce early marriage prevalence.

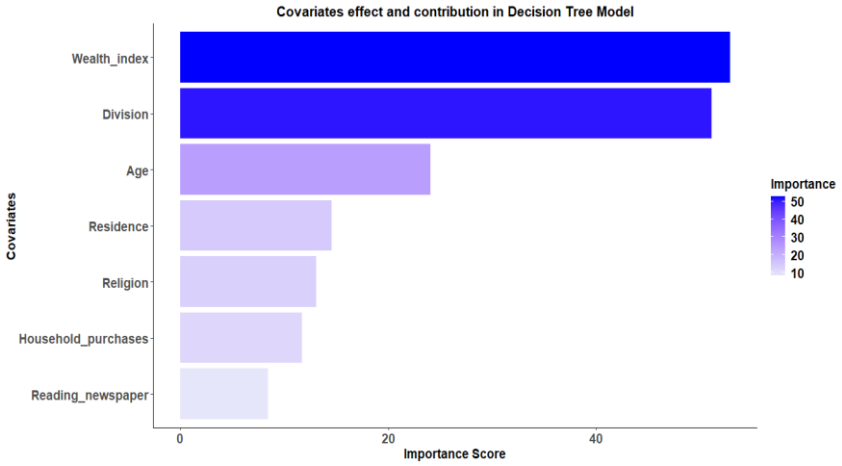

Figure 4: Covariates effect and contribution for Decision Tree model

4. Discussion

This study applies eight machine learning algorithms to identify the key socio-demographic and household decision making risk factors associated with early marriage among women in Bangladesh and to evaluate the predictive performance of these models' using data from the 2022 Bangladesh Demographic and Health Survey (BDHS). Key predictors identified through feature selection methods included Division, Wealth Index, and Reading newspaper consistently emerged as the most influential predictors of early marriage. Among all models, the Decision Tree provided the best balanced performance both on the training and testing set, indicating its suitability for generalizable early marriage prediction. Statistically significant associations (p-value < 0.05) were observed for most of the variables including Age, Residence, Division, Religion, Reading newspaper, Household purchases, and Wealth index. The important statistically significant correlates are similar the neighboring country India[43], Myanmar[44], Pakistan[45]. The Figure 4 revealed that wealth index, division, and age are the most influential correlates of early marriage, with residence, religion, household purchases, and reading newspaper contributing to a lesser extent, underscoring the pivotal role of socioeconomic and regional factors. Wealth index showed the highest significant correlates with early marriage. Women in the middle quintile faced the highest incidence of early marriage (21.74%), while the lowest rate was noted among the Richest (15.92%). Household wealth strongly influences early marriage, as poorer families may marry off daughters early to reduce economic burden or secure dowries, while wealthier families can afford to delay marriage for education and other opportunities. This aligns with widespread evidence linking poverty to early marriage practices[4]. Regional disparities (captured by division) reflect differences in cultural norms, enforcement of marriage laws, education access, and economic opportunities across administrative regions, making geographic location a critical driver of early marriage risk. Higher early marriage rates were noted in Dhaka (15.28%), Rajshahi (15.06%), and Khulna (14.56%), whereas lower rates were recorded in Sylhet (7.53%) and Mymensingh (10.12%). Women aged 25-29 years have reported the highest rate of early marriage at 43.11%, followed by those in the 20-24 age group at 33.50%[4]. Although used as a predictor here, younger women in reproductive age brackets are more susceptible to early marriage, inherently tying age to the probability of being married early. The place of residence (Urban vs. rural) plays a significant role; rural areas often exhibit higher prevalence of early marriage due to traditional practices, lower educational facilities, and less exposure to awareness campaigns. It indicated that early marriage was more common among rural women (69.37%) compared to their urban counterparts (30.63%)[8]. Religion has a notable connection to early marriage, with the highest occurrence found among Muslim women at 92.04%, while the lowest was observed among Christians at 0.07%[7]. Different religious communities may have varying norms, beliefs, or customary laws influencing acceptable marriage ages, thus contributing distinctly to early marriage risks. The decision-making process regarding household purchases was also linked to early marriage, with a higher occurrence observed among women who were not part of these decisions. Women's involvement in household purchasing decisions serves as a proxy for empowerment; limited decision-making capacity often coincides with patriarchal norms that favor marrying daughters early. Media exposure, especially reading newspapers, showed a strong correlation with early marriage, which was notably more prevalent among women who had little to no exposure (98.64%)[9]. Lower regular media exposure reduces awareness of early marriage risks and laws, increasing its likelihood, though its influence is less dominant. The consistency of these predictors across multiple selection methods strengthens the reliability of the results. However, the study includes the use of a large, representative dataset and the application of multiple robust ML

algorithms. Moreover, a constraint is that certain observations are absent, and those absent observations are removed. This study shows that machine learning algorithm can effectively predict early marriage and identify key factors. It actually highlights that geographic and socioeconomic disparities especially household wealth and administrative division are primary drivers of early marriage in Bangladesh. These insights can directly inform policy by enabling the use of practical models like Decision Trees to identify high-risk regions and communities for targeted action. To overcome early marriage, policymakers should prioritize tailored interventions in the most affected divisions, strengthen poverty reduction initiatives, and expand educational and economic opportunities for girls. Additionally, improving access to information through mass media campaigns, fostering women's participation in household decisions, and engaging religious and community leaders to shift social norms can collectively help delay marriage age. Thus, an integrated strategy addressing both economic vulnerabilities and socio-cultural practices is essential to reduce early marriage and achieve national and global development goals.

Strengths and limitations

Firstly, we analyzed a large and nationally representative sample of the most recent BDHS data (2022), ensuring the generalizability of our key findings in the Bangladeshi context. Second, we used three advanced feature selection techniques and eight machine learning algorithms to identify the key correlates associated with early marriage. The research addresses early marriage a pressing issue in Bangladesh with a novel analytical lens, offering data-driven evidence for policy intervention. The model considers multiple socio-demographic and household variables simultaneously, increasing robustness and controlling for confounders. Additionally, the research emphasized the differences in the prevalence of early marriage across regions, offering important geographical insights that can guide focused interventions. This research has certain limitations that should be taken into account when analyzing our data. Firstly, the data utilized in this research is cross-sectional, which restricts the capacity to draw causal inferences between the identified factors and early marriage, as the observed associations do not confirm temporal relationships. Secondly, depending on self-reported data regarding the age at which individuals first marry may lead to recall or social desirability bias. Additionally, several significant factors, including cultural practices, decisions related to the first marriage, and specific parental influences, were excluded because of data constraints. While the sample was prepared for analysis, some missing data in the explanatory variables may affect our estimates. Furthermore, the study is deficient in qualitative data that could offer a more profound insight into the socio-cultural context, such as the influence of religious leaders and personal experiences associated with early marriage.

5. Conclusion

This research emphasizes the significant prevalence of early marriage among women in Bangladesh, with three-quarters of them experiencing it, especially in certain regions like Dhaka, Rajshahi, and Khulna division. The incidence of early marriage is particularly elevated in rural areas and among women from lower socio-economic backgrounds. In contrast, media exposure like as newspaper, Muslim women, and age for both women and their husbands were strongly linked to a decreased probability of early marriage. Our findings suggest that culturally appropriate and effective interventions should focus on empowering and enhancing education, particularly in rural and impoverished households. These efforts could potentially alter socio-cultural practices and contribute to reduce early marriage among women in Bangladesh. Our research offers valuable insights into the individual, household, and community factors that contribute to early marriage.

This information can guide decision-making and play a crucial role in reaching the Sustainable Development Goal of eradicating early marriage by 2030. Future research should include comparative study and identify the present scenario of early marriage among neighboring countries.

Acknowledgement: The author sincerely thanks the reviewer for his encouraging and detailed feedback, which was both inspiring and useful.

References

[1] M. Lokot, M. Sulaiman, A. Bhatia, N. Horanieh, and B. Cislaghi (2021). Conceptualizing 'agency' within child marriage: Implications for research and practice, Child Abus. Negl., vol. 117, pp. 1–29, 2021, doi: 10.1016/j.chiabu.2021.105086.

[2] A. K. Sinha et al., (2013). Child marriage, Econ. Polit. Wkly., vol. 48, no. 52, p. 5, 2013, https://data.unicef.org/Child-Marriage-Data-Brief.

[3] C. Profile, U. G. Programme, T. O. End, and C. Marriage, Child marriage context, 2021, [Online]. Available: https://data.unicef.org/resources/covid-19-a-threat-to-progress-against-child-marriage/

[4] M. M. Rashid et al. (2024). Exploring determinants of early marriage among women in Bangladesh: A multilevel analysis, PLoS One, vol. 19, no. 10 October, pp. 1–14, 2024, doi: 10.1371/journal.pone.0312755.

[5] S. Scott et al., (2005). Early marriage and early childbearing in South Asia: trends, inequalities, and drivers from 2005 to 2018, Ann. N. Y. Acad. Sci., vol. 1491, no. 1, pp. 60–73, 2021, doi: 10.1111/nyas.14531.

[6] S. Wahyuningsih, S. Widati, N. Puspitasari, and L. A. Salim (2025). narra j, pp. 1–15, 2025.

[7] S. Tauseef and F. D. Sufian (2024). The Causal Effect of Early Marriage on Women's Bargaining Power: Evidence from Bangladesh, World Bank Econ. Rev., vol. 38, no. 3, pp. 598–624, 2024, doi: 10.1093/wber/lhad046.

[8] A. Talukder, M. M. Hasan, S. R. Razu, and M. Z. Hossain (2020). Early marriage in Bangladesh: A cross-sectional study exploring the associated factors, J. Int. Womens. Stud., vol. 21, no. 1, pp. 68–78, 2020.

[9] M. S. Alam, M. I. Tareque, E. D. Peet, M. M. Rahman, and T. Mahmud (2021). Female Participation in Household Decision Making and the Justification of Wife Beating in Bangladesh, J. Interpers. Violence, vol. 36, no. 7–8, pp. 2986–3005, 2021, doi: 10.1177/0886260518772111.

[10] S. C. Biswas, S. Karim, and S. F. Rashid (2020). Should we care: A qualitative exploration of the factors that influence the decision of early marriage among young men in urban slums of Bangladesh, BMJ Open, vol. 10, no. 10, pp. 1–12, 2020, doi: 10.1136/bmjopen-2020-039195.

[11] NIPORT and ICF (2023). Bangladesh Demographic Health Survey 2022: Key Indicator Report, p. 84, 2023.

[12] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[13] H. Ahmadpour, O. Bazrafshan, E. Rafiei-Sardooi, H. Zamani, and T. Panagopoulos (2021). Gully erosion susceptibility assessment in the kondoran watershed using machine

learning algorithms and the boruta feature selection, Sustain., vol. 13, no. 18, 2021, doi: 10.3390/su131810110.
[14] M. B. Kursa and W. R. Rudnicki (2010). Feature selection with the boruta package, J. Stat. Softw., vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.
[15] V. Fonti and E. Belitser (2017). Feature selection using lasso, VU Amsterdam Res. Pap. Bus. Anal., vol. 30, pp. 1–25, 2017.
[16] R. Tibshirani (1996). Lasso Tibshirani.pdf, Journal of the Royal Statistical Society, Series B (Methodogical), vol. 58, no. 1. pp. 267–288, 1996.
[17] T. Zin Win and N. Saing Moon Kham (2019). Information Gain Measured Feature Selection to Reduce High Dimensional Data, Int. Conf. Comput. Appl., vol. 17, no. 1, pp. 68–73, 2019.
[18] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj (2016). Feature selection using information gain for improved structural-based alert correlation, PLoS One, vol. 11, no. 11, pp. 1–18, 2016, doi: 10.1371/journal.pone.0166017.
[19] B. Azhagusundari and A. S. Thanamani (2013). Feature Selection based on Information Gain, Int. J. Innov. Technol. Explor. Eng., vol. 2, no. 2, pp. 18–21, 2013.
[20] "Logistic Regression in Machine Learning | GeeksforGeeks." Accessed: Apr. 29, 2025. [Online]. Available: https://www.geeksforgeeks.org/understanding-logistic-regression/
[21] E. Bisong (2019). Logistic Regression, Build. Mach. Learn. Deep Learn. Model. Google Cloud Platf., pp. 243–250, 2019, doi: 10.1007/978-1-4842-4470-8_20.
[22] A. Arista (2022). Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19, Sinkron, vol. 7, no. 1, pp. 59–65, 2022, doi: 10.33395/sinkron.v7i1.11243.
[23] "Decision Tree in Machine Learning - GeeksforGeeks." Accessed: Jun. 28, 2025. [Online]. Available: https://www.geeksforgeeks.org/decision-tree-introduction-example/
[24] "Decision Tree - GeeksforGeeks." Accessed: Dec. 13, 2024. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/
[25] "Random Forest Algorithm in Machine Learning." Accessed: Apr. 29, 2025. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm
[26] "Random Forest Algorithm in Machine Learning." Accessed: Apr. 29, 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
[27] P. Probst, M. N. Wright, and A. L. Boulesteix (2019). Hyperparameters and tuning strategies for random forest, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 3, pp. 1–19, 2019, doi: 10.1002/widm.1301.
[28] "Gradient Boosting in ML | GeeksforGeeks." Accessed: Apr. 29, 2025. [Online]. Available: https://www.geeksforgeeks.org/ml-gradient-boosting/
[29] R. Sibindi, R. W. Mwangi, and A. G. Waititu (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices, Eng. Reports, vol. 5, no. 4, pp. 1–19, 2023, doi: 10.1002/eng2.12599.
[30] A. Natekin and A. Knoll (2013). Gradient boosting machines, a tutorial, Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
[31] "What is the XGBoost algorithm and how does it work?" Accessed: Apr. 29, 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
[32] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz (2021). A comparative analysis of gradient

boosting algorithms, vol. 54, no. 3. Springer Netherlands, 2021. doi: 10.1007/s10462-020-09896-5.

[33] "AdaBoost Classifier | TDS Archive." Accessed: Apr. 29, 2025. [Online]. Available: https://medium.com/data-science/adaboost-classifier-explained-a-visual-guide-with-code-examples-fc0f25326d7b

[34] W. Hu, S. Member, W. Hu, and S. Maybank (2008). AdaBoost-Based Algorithm for Network, IEEE Trans. Syst. Man. Cybern., vol. 38, no. 2, pp. 577–583, 2008.

[35] X. Li, L. Wang, and E. Sung (2008). AdaBoost with SVM-based component classifiers, Eng. Appl. Artif. Intell., vol. 21, no. 5, pp. 785–795, 2008, doi: 10.1016/j.engappai.2007.07.001.

[36] "LogitBoost: An Overview and Implications in Modern Machine Learning | by Everton Gomede, PhD | The Modern Scientist | Medium." Accessed: Apr. 29, 2025. [Online]. Available: https://medium.com/the-modern-scientist/logitboost-an-overview-and-implications-in-modern-machine-learning-b94f4ac0dc23

[37] H. R. Pourghasemi, A. Gayen, S. Park, C. W. Lee, and S. Lee (2018).Assessment of landslide-prone areas and their zonation using logistic regression, LogitBoost, and naïvebayes machine-learning algorithms, Sustain., vol. 10, no. 10, 2018, doi: 10.3390/su10103697.

[38] "Activation Functions: Sigmoid, Tanh, ReLU, Leaky ReLU, Softmax | by Mukesh Chaudhary |Medium." Accessed: Apr. 29, 2025. [Online]. Available: https://medium.com/@cmukesh8688/activation-functions-sigmoid-tanh-relu-leaky-relu-softmax-50d3778dcea5

[39] A. Koutsoukas, K. J. Monaghan, X. Li, and J. Huan (2017). Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data, J. Cheminform., vol. 9, no. 1, pp. 1–13, 2017, doi: 10.1186/s13321-017-0226-y.

[40] Z. S. Kadhim, H. S. Abdullah, and K. I. Ghathwan (2022). Artificial Neural Network Hyperparameters Optimization: A Survey," Int. J. online Biomed. Eng., vol. 18, no. 15, pp. 59–87, 2022, doi: 10.3991/ijoe.v18i15.34399.

[41] "Confusion matrix - Wikipedia." Accessed: Apr. 29, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix

[42] R. Huang et al., (2022). Hossin, M. 1 and Sulaiman, M.N. 2 1, Med. Image Anal., vol. 80, no. 2, p. 102478, 2022.

[43] M. Singh, C. Shekhar, and J. Gupta (2024). Distribution and determinants of early marriage and motherhood: a multilevel and geospatial analysis of 707 districts in India., BMC Public Health, vol. 24, no. 1, p. 2844, 2024, doi: 10.1186/s12889-024-20038-2.

[44] R. Kabir et al., (2019). Domestic violence and decision-making power of married women in Myanmar: Analysis of a nationally representative sample, Ann. Saudi Med., vol. 39, no. 6, pp. 395–402, 2019, doi: 10.5144/0256-4947.2019.395.

[45] O. I. Asghr, L. Anwar, S. M. Hina, N. Younus, et al., (2023). Determinants of Early Marriage and Its Impact on Women Empowerment in AJK Pakistan, J. Asian …, vol. 12, no. 3, pp. 187–204, 2023, [Online]. Available: https://poverty.com.pk/index.php/Journal/article/view/107%0Ahttps://poverty.com.pk/index.php/Journal/article/download/107/80