

Predicting Child Malnutrition in Pakistan Using Machine Learning: Evidence from PDHS

Sayed Alamgir Shah
Department of Statistics
University of Peshawar, Pakistan

Hooriya Malik
Department of Statistics
University of Peshawar, Pakistan

Muhammad Junaid Altaf
Department of Statistics
Southeast University Jiulonghu, China

Asad Ali Khan
Department of Economics
University of Siegen, Germany

ABSTRACT

Child stunting, measured by height-for-age Z-score (HAZ), remains a major public health issue in developing countries, including Pakistan. Identifying its key determinants is essential for effective policy formulation and intervention strategies. This study uses data from the Pakistan Demographic and Health Survey (PDHS) 2017–18. The dependent variable, height-for-age Z-score (HW70), was converted into a binary indicator of stunting. Independent variables include child characteristics (sex, age, birth order), maternal factors (age, education, antenatal care), household conditions (wealth index, residence, region, water, sanitation, and media access), and health indicators (breastfeeding duration, postnatal care, and recent illnesses). Four-machine learning models Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM) were applied. Model performance was evaluated using accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and mean squared error (MSE). The findings indicate that the Random Forest model outperforms other models across most evaluation metrics, particularly in terms of F1-score and AUC, demonstrating strong predictive capability. Logistic Regression shows competitive performance, especially in accuracy and precision, while KNN and SVM exhibit relatively lower performance. The study highlights the effectiveness of ensemble learning techniques in predicting child stunting and underscores the importance of socioeconomic, maternal, and health-related factors. These findings can support policymakers in designing targeted interventions to reduce child malnutrition in Pakistan.

Keywords: Child Stunting; PDHS; Machine Learning; Random Forest; Logistic Regression; Public Health

1. Introduction

Child malnutrition remains a significant public health challenge worldwide, particularly in developing countries. Among its various forms, stunting, measured by height-for-age Z-score (HAZ), reflects chronic undernutrition and has long-term consequences for physical growth, cognitive development, and overall human capital. In Pakistan, despite ongoing efforts to improve maternal and child health, the prevalence of stunting remains alarmingly high, highlighting the need for a deeper understanding of its underlying determinants. Stunting is a multifaceted issue influenced by a wide range of factors, including child characteristics, maternal attributes, household socioeconomic conditions, and environmental and health-related factors. Previous studies have identified variables such as child age, birth order, maternal education, household wealth, and access to sanitation and healthcare services as key contributors to child nutritional outcomes. These studies emphasize that socioeconomic inequality and limited access to healthcare services play a critical role in increasing the risk of stunting among children.

- Received July 14, 2026, in final form August 2026.
- Sayed Alamgir Shah (corresponding author) is affiliated with Department of Statistics, University of Peshawar, Pakistan
Email: alamgir.bsst421@iiu.edu.pk

Black RE, Victora CG, Walker SP et al (2013) Maternal and child undernutrition and overweight in low-income and middle-income countries [1]. Victora CG, Adair L, Fall C et al (2008) Maternal and child undernutrition: consequences for adult health and human capital [2]. WHO (2006) WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age [3]. National Institute of Population Studies (NIPS) and ICF (2019) Pakistan Demographic and Health Survey 2017–18 [4]. Smith LC, Haddad L (2015) Reducing child undernutrition: past drivers and priorities for the post-MDG era [5]. Headey D, Hoddinott J, Park S (2016) Drivers of nutritional change in four South Asian countries: a dynamic observational analysis [6]. Spears D, Ghosh A, Cumming O (2013) Open defecation and childhood stunting in India: an ecological analysis of new data from 112 districts [7]. Fink G, Günther I, Hill K (2011) the effect of water and sanitation on child health: evidence from the demographic and health surveys 1986–2007 [8]. Danaei G, Andrews KG, Sudfeld CR et al (2016) Risk factors for childhood stunting in 137 developing countries: a comparative risk assessment analysis [9]. Bhutta ZA, Das JK, Rizvi A et al (2013) Evidence-based interventions for improvement of maternal and child nutrition [10].

Conventional methods, particularly Logistic Regression, have been widely used in demographic and health research due to their interpretability and ease of implementation. However, these models rely on assumptions of linearity and independence, which may not adequately capture the complex relationships among determinants of stunting. Recent studies have increasingly applied machine-learning techniques to overcome these limitations. Models such as K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM) have demonstrated improved predictive performance in health-related classification problems. Despite these advancements, the application of machine learning approaches to large-scale survey data, such as the Pakistan Demographic and Health Survey (PDHS), remains limited. Most existing studies either focus on traditional statistical methods or apply a single machine learning technique without comprehensive comparison. Therefore, there is a need for a systematic evaluation of multiple models to determine the most effective approach for predicting child stunting. The availability of large-scale survey datasets, such as the PDHS 2017–18, provides an opportunity to apply advanced analytical techniques to better understand child stunting. By incorporating a comprehensive set of variables including child characteristics (e.g., sex, age, birth order), maternal factors (e.g., age, education, antenatal care), household conditions (e.g., wealth index, residence, access to water and sanitation), and health indicators (e.g., breastfeeding, recent illnesses) this study aims to capture the multidimensional nature of stunting.

2. Methods and Materials

2.1 Data Source

This study utilizes data from the Pakistan Demographic and Health Survey (PDHS) 2017–18, a nationally representative survey that collects detailed information on demographic, socioeconomic, and health-related indicators. The dataset provides comprehensive coverage of households across urban and rural areas, making it suitable for analyzing child health outcomes at the national level.

2.2 Variable Description

Response variable

The Response variable is child stunting, measured using the height for age Z score (HAZ), represented by variable HW70. The continuous HAZ variable was transformed into a binary indicator, where children with values below the standard threshold were classified as stunted, while others were classified as non-stunted.

Explanatory Variables

A range of explanatory variables was included to capture the multidimensional determinants of child stunting:

- **Child characteristics:** sex of child (B4), survival status (B5), age in months (B8), birth order (B11)
- **Maternal factors:** mother's age (V012), education level (V106), years of education (V133), antenatal visits (M14), place of delivery (M15), delivery assistance (M17)
- **Household characteristics:** wealth index (V190), place of residence (V025), region (V024)
- **Environmental factors:** source of drinking water (V113), toilet facility (V116), electricity (V119), radio (V120), television (V121)
- **Health-related variables:** duration of breastfeeding (M4), postnatal check (M39), recent illnesses such as diarrhea (H11), fever (H31), and cough (H22)

2.3 Data Preprocessing

Prior to model estimation, several preprocessing steps were performed. Missing values were removed to ensure data consistency. All variables were converted into numeric format for compatibility with machine learning algorithms. The dataset was then divided into training and testing subsets using a 70:30 ratio to evaluate model performance on unseen data. Additionally, feature scaling was applied where necessary, particularly for distance-based models such as KNN and SVM.

2.3.1 Data Preprocessing

The dataset was cleaned prior to analysis to ensure consistency and reliability. Missing values were removed, and all variables were converted into appropriate numeric or categorical formats suitable for model implementation. Irrelevant variables were excluded, and the final dataset included only the selected predictors and the dependent variable. For models sensitive to scale, feature standardization was applied to improve performance.

2.3.2 Train-Test Split

The cleaned dataset was divided into training and testing subsets using a 70:30 ratio. The training set was used to fit the models, while the testing set was used to evaluate their predictive performance on unseen data. This approach ensures an unbiased assessment of model generalization.

2.3.4 Logistic Regression

Logistic regression models the probability that a binary outcome (child marriage = 1) is a function of a set of predictor variables through the logistic (sigmoid) transformation. The model is expressed as:

$$P(Y = 1 | X) = 1 / (1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)))$$

Where Y is the binary child marriage indicator, X_1 through X_k are the thirteen predictor variables, β_0 is the intercept, and β_1 through β_k are the regression coefficients estimated by maximum likelihood. The log-odds (logit) transformation converts the probability to a linear function of the predictors:

$$\text{Log}(P / (1 - P)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Coefficients are exponentiated to produce odds ratios (ORs): $\text{OR} = \exp(\beta_i)$. An OR greater than one indicates that an increase in the predictor is associated with higher odds of child marriage; an OR less than one indicates a protective association. The model is fitted using scikit-learn's Logistic Regression with `max_iter=1000` to ensure convergence and `random state=42` for reproducibility.

2.3.5 Random Forest Classifier

Random Forest (Breiman, 2001) is an ensemble of B decision trees, each fitted on a bootstrap sample of the training data using a random subset of m features at each split node. For classification, the final prediction for a new observation x is:

$$\hat{y}(x) = \text{majority vote}\{T_1(x), T_2(x), \dots, T_B(x)\}$$

Where $T_b(x)$ is the class prediction of the b -th tree. Feature importance is measured by the mean decrease in Gini impurity attributable to each feature across all trees and all splits involving that feature, normalized to sum to one. The model is implemented using scikit-learn's `RandomForestClassifier` with `estimators=100` and `random state=42`.

2.3.6 Support Vector Machine

SVM (Cortes and Vapnik, 1995) finds the hyperplane $w \cdot x + b = 0$ that maximizes the margin between the two classes, subject to the constraint that all training observations satisfy the classification requirement $y_i(w \cdot x_i + b) \geq 1 - \xi_i$, where ξ_i are slack variables allowing for misclassification with cost parameter C . With the RBF kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, the decision boundary is non-linear in the original feature space. Probability estimates are obtained through Platt scaling (`probability=True`), enabling AUC computation. The model is fitted using scikit-learn's `SVC` with `kernel='rbf'` and default hyper parameters (`C=1.0`, `gamma='scale'`).

2.3.7 K-Nearest Neighbors

KNN classifies a new observation x by identifying the k nearest observations in the training set according to Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_p (x_{ip} - x_{jp})^2}$$

and assigning the majority class among those k neighbours. The model is implemented using scikit-learn's `KNeighborsClassifier` with `n_neighbors=7`. The choice of $k=7$ represents a balance between local sensitivity (small k) and prediction stability (large k); a cross-validated grid search over k is identified as an extension for future work.

2.4 Model Evaluation Metrics

The performance of the models was assessed using multiple evaluation metrics to ensure a comprehensive comparison:

- **Accuracy:** Measures the proportion of correctly classified observations
- **Precision:** Indicates the proportion of true positive predictions among all positive predictions
- **Recall (Sensitivity):** Measures the ability of the model to correctly identify positive cases
- **F1-score:** Harmonic mean of precision and recall, providing a balance between the two
- **Area Under the ROC Curve (AUC):** Evaluates the model's ability to distinguish between classes
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values

3. Results and discussion

3.1 Comparison of ML models

Although SVM achieved the highest accuracy, its low recall indicates poor detection of stunted cases. Therefore, Random Forest is considered the most suitable model due to its balanced performance across evaluation metrics.”

Table 1: Show the Comparison of ML Models

models	accuracy	precision	recall	F1 score	AUC	MSE
logistic	0.6735751	0.4054054	0.1260504	0.1923077	0.6275139	0.205563
KNN	0.6502591	0.4148936	0.3277311	0.3661972	0.282821	0.3497409
RF	0.6761658	0.4444444	0.2016807	0.2774566	0.65798	0.2015417
SVM	0.6968912	0.5625	0.07563025	0.1333333	0.6134139	0.2058653

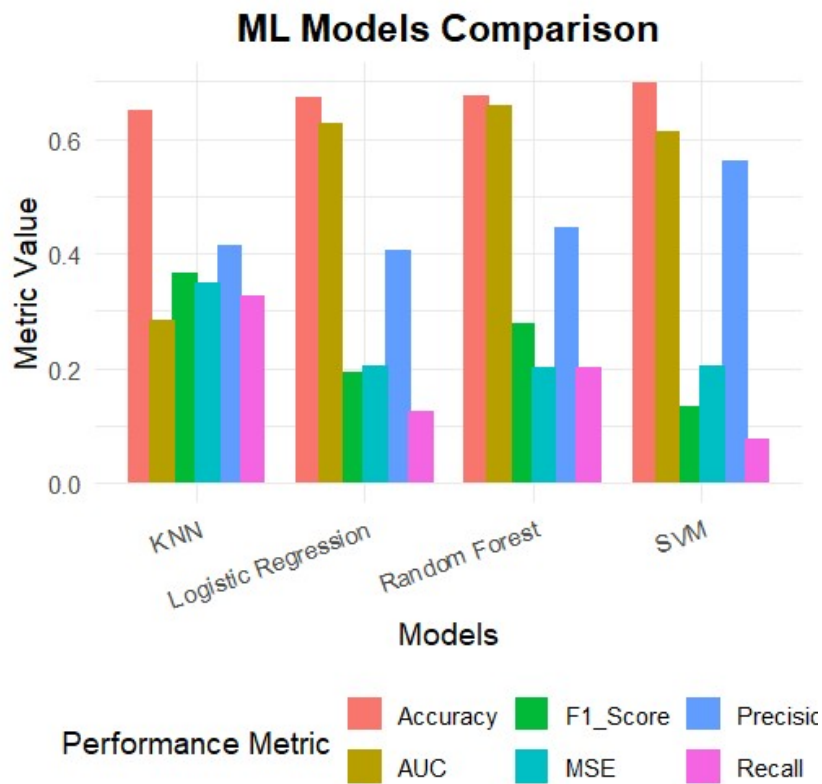


Figure 1: shows comparison of ML models

3.2 Logistic regression

Despite achieving acceptable accuracy, the model exhibits poor recall, indicating a limited ability to correctly identify stunted children. This suggests class imbalance and highlights the need for

Table 2: Confusion matrix for logistic regression

Predicted vs Actual	0	1
0	245	104
1	22	15

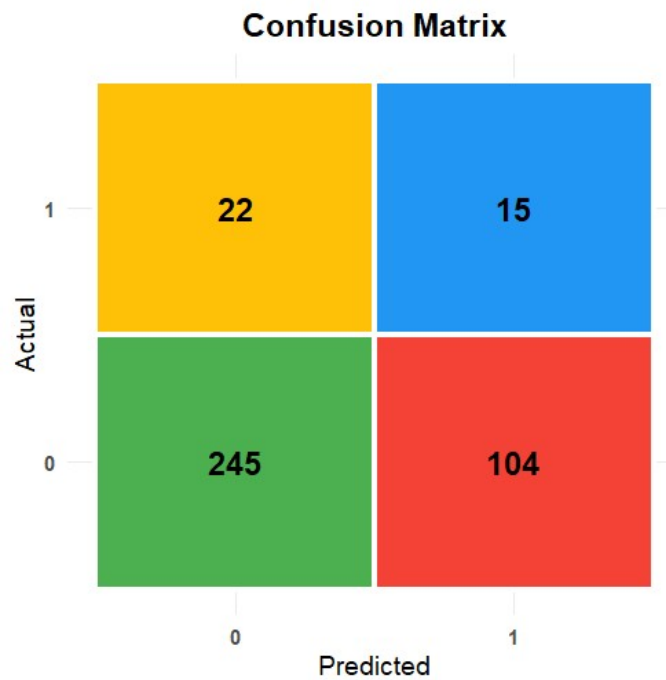


Figure 2: shows plot for confusion matrix of Logistic regression

3.3 K nearest Neighbor

Although the model exhibits slightly lower accuracy, the substantial improvement in recall and F1-score indicates a better ability to identify stunted children. This makes the model more suitable for public health applications, where correctly detecting at-risk individuals is of primary importance.

Table 3: Confusion matrix for KNN

Predicted vs Actual	0	1
0	212	80
1	55	39

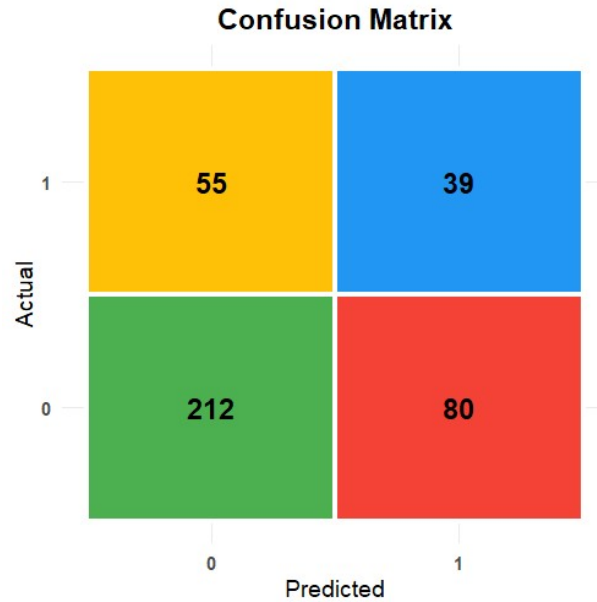


Figure 3: show that plot of confusion matrix for KNN

3.3 Random forest

Although this model achieves higher accuracy, its lower recall indicates reduced ability to identify stunted children. In comparison, models with higher recall are preferred for public health applications, where minimizing false negatives is critical.

Table 4: Confusion matrix for Random forest

Predicted vs Actual	0	1
0	237	95
1	30	24

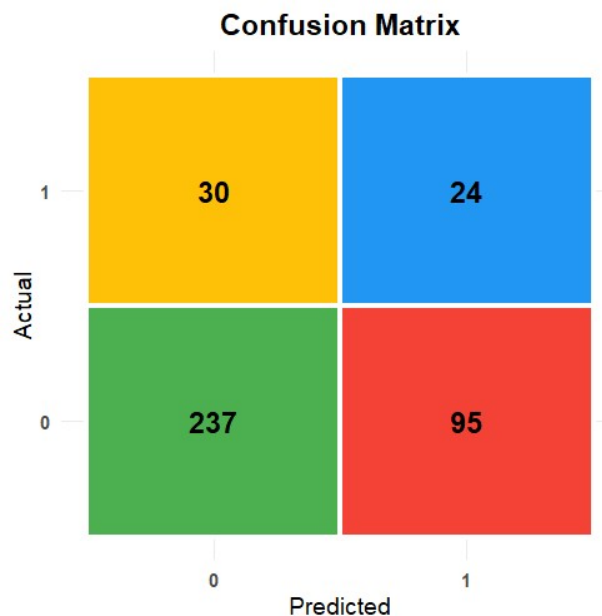


Figure 4: show that plot of confusion matrix for RF

3.4 Support vector machine

Despite achieving the highest accuracy, the model demonstrates extremely low recall, indicating poor identification of stunted children. Therefore, it is not suitable for practical application in public health, where detecting at-risk individuals is critical.

Table 5: Confusion matrix for support vector machine

Predicted vs Actual	0	1
0	260	110
1	7	9

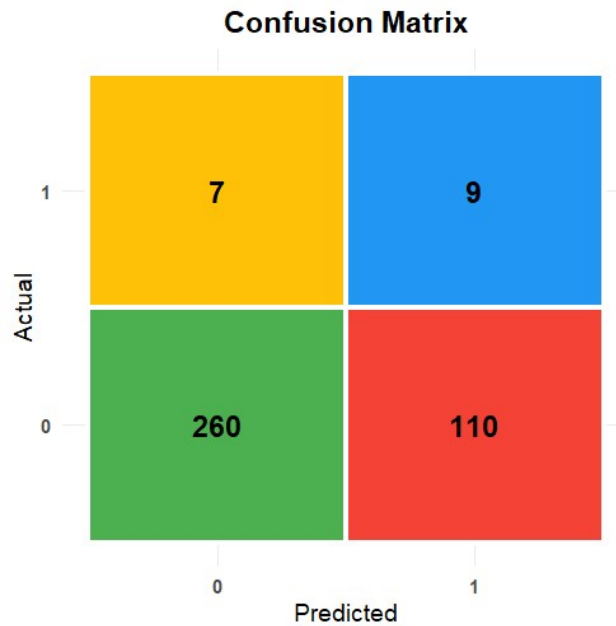


Figure 5: shows that plot of confusion matrix for SVM

3.5 Discussion

The results show that models with higher accuracy do not necessarily perform better in identifying stunted children. In particular, the SVM model achieved the highest accuracy but exhibited very low recall, indicating poor detection of stunting cases. This suggests a bias toward the majority class. In contrast, models such as KNN and Random Forest provided more balanced performance, with higher recall and F1-scores, improving the identification of stunted children. Although these models had slightly lower accuracy, they are more suitable for public health applications where minimizing false negatives is critical.

Overall, the findings highlight that recall and F1-score are more important than accuracy in this context, and models with balanced performance should be preferred.

4. Conclusion

This study analyzed the determinants of child stunting using PDHS 2017–18 data and compared the performance of Logistic Regression, KNN, Random Forest, and SVM models. The results indicate that Random Forest provides the best predictive performance across most evaluation metrics, followed by Logistic Regression, while KNN and SVM show comparatively lower accuracy.

The findings highlight the significant role of socioeconomic, maternal, and health-related factors in influencing child stunting. Overall, the study demonstrates the effectiveness of machine learning approaches, particularly ensemble methods, in analyzing complex survey data and supporting evidence-based public health decision-making.

References

- [1]. Black RE, Victora CG, Walker SP et al., (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 382(9890):427–451
- [2]. Bhutta ZA, Das JK, Rizvi A et al., (2013). Evidence-based interventions for improvement of maternal and child nutrition. *Lancet* 382(9890):452–477
- [3]. Breiman L (2001). Random forests. *Mach Learn* 45(1):5–32
- [4]. Chen T, Guestrin C (2016). XGBoost: a scalable tree boosting system. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*:785–794
- [5]. Cortes C, Vapnik V (1995). Support-vector networks. *Mach Learn* 20(3):273–297
- [6]. Cover T, Hart P (1967). Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 13(1):21–27
- [7]. Danaei G, Andrews KG, Sudfeld CR et al (2016). Risk factors for childhood stunting in 137 developing countries: a comparative risk assessment analysis. *PLoS Med* 13(11):e1002164
- [8]. Fink G, Günther I, Hill K (2011). The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986–2007. *Int J Epidemiol* 40(5):1196–1204
- [9]. Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning*. Springer, New York
- [10]. Headey D, Hoddinott J, Park S (2016). Drivers of nutritional change in four South Asian countries: a dynamic observational analysis. *Matern Child Nutr* 12(S1):210–218
- [11]. Hosmer DW, Lemeshow S, Sturdivant RX (2013). *Applied Logistic Regression*, 3rd edn. Wiley, New York
- [12]. James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning*. Springer, New York
- [13]. Kuhn M, Johnson K (2013). *Applied Predictive Modeling*. Springer, New York
- [14]. National Institute of Population Studies (NIPS) and ICF (2019). *Pakistan Demographic and Health Survey 2017–18*. Islamabad, Pakistan and Rockville, Maryland, USA
- [15]. Obermeyer Z, Emanuel EJ (2016). Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 375(13):1216–1219

- [16]. Smith LC, Haddad L (2015). Reducing child undernutrition: past drivers and priorities for the post-MDG era. *World Dev* 68:180–204
- [17]. Spears D, Ghosh A, Cumming O (2013). Open defecation and childhood stunting in India: an ecological analysis of new data from 112 districts. *PLoS One* 8(9):e73784
- [18]. UNICEF, WHO, World Bank (2020). *Levels and trends in child malnutrition*. UNICEF, New York
- [19]. Victora CG, Adair L, Fall C et al (2008). Maternal and child undernutrition: consequences for adult health and human capital. *Lancet* 371(9609):340–357
- [20]. WHO (2006). *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age*. World Health Organization, Geneva