Acknowledgements

I would like to express my profound gratitude to my supervisor, Lu Xuan, whose expertise, understanding, and patience, added considerably to my experience, and whose guidance supported me throughout my research. I appreciate her willingness to give her time so generously.

Special thanks are due to my family. I must acknowledge my parents for their unwavering support and encouragement throughout my education and study, from the very beginning to this point. Their belief in my abilities and their sacrifices have been the cornerstone of my achievements.

I am also grateful to my husband, who has been a constant source of love, encouragement, and support. His understanding and personal sacrifices have played a significant role in my success.

Additionally, I would like to thank my friends, who have supported me throughout this journey. Their camaraderie and constant morale boosts have been invaluable.

This research would not have been possible without the collective support and encouragement of all mentioned, and for that, I am eternally grateful.

Table of Contents

Abstract....3
Introduction and Problem Statement....4
Review of Literature....5
Methodology and Data....7
    Data....7
    Descriptive Statistics....7
    Bivariate Analysis....8
    Preprocessing....8
    Model Development and Evaluation....9
Results....10
    Descriptive Statistics....10
    Bivariate Analysis....12
    Machine Learning Algorithms....13
Limitations and Future Work....17
References....19

The Analysis of Child Nutrition Status:
For the Case of Uzbekistan

Abstract

Background: Stunting in children under five years of age remains a significant public health issue globally, associated with long-term developmental impairments. Early identification of risk factors for stunting can facilitate targeted interventions, potentially reducing its prevalence and improving child health outcomes.

Objective: This study aimed to utilize machine learning algorithms to identify and predict the risk factors associated with stunting in children, leveraging a dataset that includes demographic, nutritional, and health-related variables.

Methods: The project employed several preprocessing techniques to prepare the data, including scaling of numerical features, one-hot encoding of categorical variables, and outlier removal. Four machine learning models—Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbors—were evaluated for their ability to predict stunting. The models were assessed based on accuracy, precision, recall, and F1 scores, with a particular focus on their performance in predicting the stunting class.

Results: Logistic Regression and Random Forest demonstrated higher overall accuracy and effectiveness in predicting non-stunting cases. However, challenges were observed in enhancing the recall for stunting predictions across all models, indicating difficulties in identifying all positive cases. Logistic Regression provided the most balanced performance with respect to precision and recall, suggesting its potential utility in practical screening applications.

Conclusion: The findings underscore the potential of machine learning in enhancing the understanding and prediction of stunting in children. Future research should focus on optimizing model parameters and incorporating more diverse datasets to improve predictive accuracy, particularly for stunting cases. The study highlights the importance of advanced predictive analytics in public health strategies aimed at combating child stunting.

Key Words: Uzbekistan, child stunting, maternal education machine learning algorithms, Logistic Regression, SVM, KNN, Random Forest

Introduction and Problem Statement

Child nutrition has emerged as a significant focal point within health-related research, particularly due to its complex interactions with socio-economic factors. Despite numerous studies examining potential causes and effects of child malnutrition, the specific influence of maternal education on child nutrition has not been adequately highlighted, especially in the context of Central Asia. This region is notably challenged by restricted access to higher education for women, stemming largely from its patriarchal social structures. In such settings, educational opportunities are preferentially allocated to males, a practice underpinned by the prevailing belief that investments in the education of sons yield superior economic returns compared to equivalent investments in daughters. This imbalance not only undermines female empowerment but also precipitates a cascade of socio-economic disadvantages, of which child malnutrition is a significant manifestation.

This paper focuses on unraveling the correlation between maternal higher education and the incidence of child stunting in Uzbekistan. The selection of Uzbekistan as the study locale is strategic, based on the availability of recent data which is presumed to be representative of stunting patterns across the entire Central Asian region. The investigation addresses the primary factor of maternal education while integrating additional demographic and household variables that literature suggests are relevant. Employing machine learning algorithms, this study aims to predict and analyze the key drivers of child stunting, thereby providing insightful data that could inform targeted interventions. By enhancing our understanding of how maternal educational attainment influences child nutrition outcomes, this research is poised to make a substantial contribution towards improving the patterns of child malnutrition in Uzbekistan and potentially across similar contexts globally. Through this exploration, the paper seeks to underscore the broader implications of educational equity and its critical role in addressing public health challenges.

This project is among the first to leverage newly available data to explore the analysis of child nutrition status through a number of key determinants including maternal education level in Uzbekistan. This novelty provides fresh insights into a previously under-researched area. The findings that the research reveal have the potential to empirically support, or challenge existing assumptions and policies related to gender, education, and health in Uzbekistan. This can be instrumental in shaping future policies and interventions aimed at improving child nutrition and reducing gender disparities.

Indeed, the application of various statistical and machine learning algorithms, such as logistic regression, random forest, SVM, and KNN, to analyze the data is a novel approach in this context. It allows for a comprehensive and multifaceted analysis of the impact of maternal education on child nutrition. By highlighting the role of maternal education in child health, this research could contribute to a greater understanding and appreciation of women education and its broader societal benefits in Uzbekistan.

Review of Literature

The body of research investigating the interplay between maternal education and child nutrition has predominantly focused on delineating the various mechanisms through which a mother's literacy level affects her children's health outcomes. A seminal study by Caldwell (1979), which utilized data from Nigeria, was instrumental in identifying maternal education as a crucial determinant of child survival rates, surpassing even the influential socio-economic conditions of households. Caldwell's research emphasized the enhanced comprehension of health-related information among educated mothers and their increased authority over health-related decisions for their children. Since Caldwell's groundbreaking study, a succession of research efforts has continued to validate the positive correlation between maternal education and enhanced child nutrition outcomes. Notably, Gwatkin et al. (2000) observed that instances of child malnutrition were considerably less frequent among children of educated mothers. On the other hand, Desai and Alva (1998) reported more varied results; they found the positive impact of maternal education on children's height-for-age was consistent in only 5 out of 22 countries examined. Similarly, research by Bairagi (1980) and Solon et al. (1985) showed a significant relationship between maternal education and child nutritional status predominantly within wealthier segments of populations, with no significant correlation found within poorer groups.

Subsequent reviews of related literature have outlined that the influence of maternal education on child nutritional outcomes is mediated through four distinct channels: socio-economic status, women's empowerment, health-related knowledge and practices, and reproductive behavior. Various studies have used maternal education as a proxy for socio-economic status at both personal and household levels. These studies suggest that women with higher educational attainments are more likely to secure better job opportunities and are inclined to marry partners who are also well-educated (Cleland and Van Ginneken, 1988). Moreover, educated women frequently reside in urban settings where they benefit from superior health facilities and hygienic conditions.

In a significant study utilizing the 2005 Demographic and Health Survey (DHS) data from Cambodia, Miller and Rodgers (2009) probed how maternal education influences the nutrition of children under five years old, taking into account variables such as birth size, and indicators of malnutrition such as wasting and stunting. Their findings underscored a negative correlation between maternal education and child malnutrition, suggesting that higher educational levels in mothers are associated with better nutritional outcomes in children. This inverse relationship was similarly highlighted in reports by Mukuria et al. (2005) and Frost et al. (2005), which further reinforced the beneficial impact of maternal education on child health.

Moreover, research by Emina et al. (2011) has shown that women's education not only promotes empowerment but also significantly enhances their involvement in decision-making processes that directly impact child nutrition and access to healthcare services. Similarly, Hobcraft (1993) pointed out that educated mothers are better positioned to make informed health choices for their children. Comprehensive reviews by Frost et al. (2005) and others have argued that maternal education affects child nutrition primarily through the transmission of health knowledge and fostering positive health attitudes. Educated mothers are more aware of the causes of diseases, their prevention, and treatment options. They also exhibit favorable attitudes toward healthcare-seeking behavior for their children, including recognizing the critical importance of immunizations (Ruel et al., 1992). Children of mothers with higher literacy levels are likely to live in areas where sanitation is prioritized, and due to regular vaccinations, they experience better nutritional outcomes.

Research also explores the influence of maternal education through health and reproductive behavior. Forste (1998) utilized data from Bolivia to show that shorter birth intervals correlate with higher rates of stunting, indicating poorer nutritional outcomes. Other studies by Cleland, Van Ginneken (1998), and Mukuria et al. (2005) have demonstrated that in cases of educated women, longer birth intervals are associated with low-risk birth ages, which are conducive to better nutritional results for children. These studies highlight a positive effect of maternal education on child nutrition within affluent households, yet note that this effect is non-existent in poorer families.

Kabubo-Mariara et al. (2009) conducted an in-depth analysis of the determinants of nutritional status of Kenyan children using data from the 1998 and 2003 Kenya DHS datasets. Their research focused on the effects of characteristics related to the child, parents, household, and community on the height of children and their likelihood of suffering from stunting. Intriguingly, their results indicated that male children were more susceptible to undernutrition than female children, and those from multiple births faced greater nutritional challenges than their single-birth counterparts. This study notably emphasized maternal education as a more significant determinant of children’s nutritional status than paternal education, underlining the profound and broad-reaching impact of maternal literacy on child health across various socio-economic backgrounds.

Methodology and Data

Data

The cross-sectional survey data is taken from Multiple Indicator Cluster Survey (MICS) carried by UNICEF for Uzbekistan. The selection of the survey for this research is based on the availability of the most recent conducted survey for selected country and the survey round of MICS-2021 was chosen. From seven different recodes available within this particular survey, three of them, individual women, under – five children of interviewed women and corresponding birth history recodes were employed. The women and birth history recodes were merged to children under five recode in accordance with the specific IDs of the observations (cluster number, household number, and line number), and post-merge sample size of 1,985 children between age of 0-59 months was used for empirical research. Merging of three recodes, namely women, birth history and children, was implemented in R programming language and the final data with 1,985 observations was downloaded as an excel file for research analysis.

To measure the nutrition status of children, stunting was used as proxy, where children (0-59 months) with height-to-age Z score of below minus two standard deviations (-2 SD) from the median of the WHO reference population are considered as stunted/malnourished, while those with above minus two standard deviations are considered non-stunted/nourished1. Dichotomous dependent variable for this study is stunting, including only 0 and 1 values, where 1 stands for the stunting and 0 otherwise for selected sample of children.

Descriptive Statistics

Initial data exploration involved generating descriptive statistics to understand the distribution and characteristics of both continuous and categorical variables within the dataset.

Continuous Variables: Variables such as age of child (in months), weight of child at birth, number of children in the family, age of mother were analyzed to provide insights into their mean, median, standard deviation, and range. This helped in understanding the central tendencies and dispersions of these measurements.

Categorical Variables: For categorical variables like stunting, birth order, birth interval, gender of child, education level of mother, wealth index, and place of residence frequency distributions were computed. This analysis helped identify the proportion of each category, facilitating an understanding of how these variables might influence stunting.

These statistics were computed using Python's Pandas and NumPy libraries, and the results were saved and printed for further reference (Look at Table 1).

Bivariate Analysis

As a further step in data exploration, bivariate analysis was conducted to examine the relationship between stunting and other categorical predictors. This involved calculating the percentage of children experiencing stunting across different categories of each predictor variable. Crosstabs, facilitated by Pandas, provided a normalized index to better understand the potential influence of these predictors on stunting. This step was crucial for identifying variables for inclusion in the machine learning models. The results were saved and printed for further reference (Look at Table 2).

Preprocessing

Date preprocessing is a critical step in preparing the dataset for effective machine learning. This process included several techniques to ensure that the data fed into the models would allow for the most accurate predictions possible.

Scaling of Numeric Features: The numeric variables such as 'month', 'weight', 'number of children', and 'age of mother' were standardized. This was done using the StandardScaler from Scikit-learn, which removes the mean and scales the data to unit variance. This step is crucial because many machine learning algorithms assume that all features are centered around zero and have variance in the same order. Without this step, models like SVM and KNN could behave unpredictably as they are sensitive to the scale of input features.

Encoding of Categorical Features: Categorical variables such as 'gender', 'birth order', 'birth interval', 'educational level', 'wealth', and 'area' were transformed using OneHotEncoder. This encoder converts categorical data into dummy/indicator variables allowing models to better understand the patterns in the data. The drop='first' parameter was used to avoid multicollinearity, which could otherwise lead to inflated variances and covariances in models like logistic regression.

Outlier Detection and Removal: The preprocessing phase also involved cleaning the data by removing outliers that could skew the results. This was performed by calculating the interquartile range (IQR) for each numeric variable. Data points that fell outside of 1.5 times the IQR below the first quartile and above the third quartile were considered outliers and removed. This method helped in mitigating the influence of extreme values which might distort the true pattern and lead to less accurate models.

Correlation Analysis: Before finalizing the dataset for modeling, a correlation matrix was generated to identify any potential multicollinearity among predictors. High correlation between predictors can affect the performance of some machine learning algorithms by making the model unstable and difficult to interpret. Variables found to be highly correlated could either be transformed or one of the variables removed to ensure better model performance.

Model Development and Evaluation

The choice of algorithms was influenced by the nature of the data and the objective of the project. Each algorithm was chosen for its strengths and appropriateness in handling binary classification tasks, especially in the context of imbalanced datasets.

Logistic Regression: This model was chosen as it provides a probabilistic framework and is interpretable, which is valuable in a medical context where understanding the influence of various factors is important. Logistic regression estimates probabilities using a logistic function, which is particularly useful for binary classification.

Random Forest: Known for its robustness and effectiveness in classification tasks, Random Forest constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) of the individual trees. It is less likely to overfit than a decision tree and provides a very good indicator of feature importance.

Support Vector Machine (SVM): SVM was selected for its effectiveness in high-dimensional spaces and its ability to model non-linear decision boundaries thanks to the kernel trick. In this project, the RBF kernel was used to allow the SVM to handle an even broader range of data structures.

K-Nearest Neighbors (KNN): This algorithm was included for its simplicity and effectiveness, making no assumptions about the functional form of the problem being solved. It classifies new cases based on a similarity measure (e.g., distance functions).

Each model was integrated into a pipeline that included scaling, SMOTE for addressing class imbalance, and the estimator itself. The SMOTE technique was particularly crucial for balancing the dataset, as stunting is typically less common than not stunting, and imbalance can bias a model towards the majority class.

Model Evaluation: The primary metric for model evaluation was the F1 score, which balances the precision and recall of a model. Precision is the number of true positives divided by the number of all positive calls (true and false positives), and recall is the number of true positives divided by the number of positives in the actual data. The F1 score is particularly useful when dealing with imbalanced classes. Cross-validation was used to assess the robustness of the models, ensuring they perform well across different subsets of the dataset.

By implementing these detailed and meticulous preprocessing and modeling steps, the project ensures rigorous analysis and robustness, thereby enhancing the reliability of the findings and recommendations.

Results

Descriptive Statistics

Table 1 offers a comprehensive descriptive analysis of the study sample, essential for ensuring robustness before conducting regression analyses. The data highlights that stunting is a significant issue in Uzbekistan, affecting approximately 20 percent of the children in the sample, indicating its prevalence as a substantial health concern within the population studied.

The demographic profile of the sampled children shows an average age of 21.13 months, with a nearly equal distribution across genders. The average birth weight recorded was 3.35 kilograms. In terms of birth order, nearly half of the sample consisted of children who were either the second or third child within their family. The distribution of birth order was as follows: 30.52 percent were firstborns, 20.45 percent were fourth to sixth children, and a negligible 0.64 percent accounted for seventh or higher order births. The intervals between the births varied, with 30 percent of the intervals being less than two years, and the smallest category, three-year intervals, comprising 11.2 percent of the sample.

The study particularly focuses on maternal education, revealing that a significant 70 percent of the women in the sample had only primary or secondary education, underscoring the limited access to higher education and subsequent reduced empowerment for women in Uzbekistan. Other maternal demographic data include an average maternal age of 27.96 years and an average number of children per woman at 2.85, suggesting most women have two to three children. Economic stratification within the sample shows that approximately 40 percent of families were classified as either poor or rich, while only 19.45 percent fell into the middle wealth category at the time of the survey. Additionally, the residential distribution of the children indicated that 60.71 percent lived in rural areas, highlighting a significant rural majority in the sample, compared to those residing in urban settings.

This demographic and maternal educational background lays the groundwork for understanding the context within which child nutritional outcomes are being studied, emphasizing the intersections of health, education, and socio-economic factors in shaping these outcomes in Uzbekistan.

Table 1

Summary Characteristics of Variables Used in the Research: Analyzing Nutritional Status of Children in Uzbekistan 2021 (N=1,985)

	Mean	**%**	**Number of Observations**
Child Nutritional Status			
Stunted		80.20	1,592
Not Stunted		19.80	393
Child Demographic Factors			
Age of Child (in months)	21.13		1,985
Weight at Birth (in kilos)	3.348		1,985
Birth Order			
First Child		30.52	606
2-3		48.36	960
4-6		20.45	406
7+		0.64	13
Birth Interval			
First Birth		30.68	609
<2 years		17.83	354
2 years		20.20	401
3 years		11.18	222
4+ years		20.10	399
Sex of Child			
Boy		51.18	1,016
Girl		48.82	969
Maternal Demographic Factors			
Higher Educational Level of Mother			
Primary		10.23	203
Secondary		59.80	1,187
Tertiary		29.97	595
Age of Mother	27.96		1,985
Number of Children	2.85		1,985
Household Characteristics			
Wealth Status			
Poor		40.55	805
Middle		19.45	386
Rich		40.00	794
Place of Residence			
Urban		39.29	780
Rural		60.71	1,205
Total			**1,985**

Bivariate Analysis

Table 2 provides a detailed exploration of the stunting patterns among children within the study, segmented by various categorical measures. The data elucidates significant trends in child stunting relative to birth order, revealing an escalating pattern of stunting as the birth order increases. Specifically, the incidence of stunting was recorded at 19.80% for first-born children, rising dramatically to 61.54% for children of the seventh or higher birth order. This trend suggests a potential resource dilution effect, where the resources available per child diminish as family size increases, potentially impacting child health negatively.

Furthermore, the analysis of birth intervals presents compelling evidence regarding their influence on stunting rates. Children born with less than two years between births exhibited the highest stunting rate at 30.23%. Conversely, as the interval between births extended, the incidence of stunting decreased, with children born after an interval of four or more years exhibiting the lowest stunting rate at 10.03%. This pattern supports the hypothesis that longer intervals between births may provide mothers with sufficient time to recover physically and nutritionally, thereby enhancing the health outcomes of subsequent children.

Regarding gender, the stunting rates between male and female children were similar, indicating that gender does not significantly influence stunting patterns within the Uzbek context. This observation suggests that the determinants of stunting may be more strongly influenced by other socio-economic and environmental factors than by gender alone. In analyzing the impact of maternal education on stunting, a clear inverse relationship is observed. Children of mothers with only primary education experienced a stunting rate of 30%, which decreased to 21% among those whose mothers had secondary education, and further to 11% for children whose mothers attained tertiary education. This trend underscores the critical role of maternal education in enhancing child health outcomes, possibly due to improved maternal knowledge and practices concerning nutrition and childcare.

Similarly, the socio-economic status of the household, as indicated by wealth, also showed a pronounced effect on stunting rates. Children from poor households exhibited a stunting rate of approximately 25%, which decreased to 19% among middle-income families, and further reduced to 15% in rich households. This gradient reflects the influence of economic resources on access to nutritional foods and health care services, which are crucial for child development. The place of residence, comparing rural versus urban settings, did not show a significant variance in the impact on child malnutrition, though a slightly higher stunting rate was observed in rural areas.

Table 2

Percentage of Children Stunting Classified by the Categorical Predictors in the Study

	Children Stunting		Number of Observations
	No	**Yes**	
Birth Order			
First Child	80.20	19.80	606
2-3	81.46	18.54	960
4-6	78.57	21.43	406
7+	38.46	61.54	13
Birth Interval			
First Birth	79.80	20.20	609
<2 years	69.77	30.23	354
2 years	76.56	23.44	401
3 years	86.96	13.06	222
4+ years	89.97	10.03	399
Sex of Child			
Boy	79.43	20.57	1,016
Girl	81.01	18.99	969
Higher Educational Level of Mother			
Primary	70.44	29.56	203
Secondary	78.52	21.48	1,187
Tertiary	86.89	13.11	595
Wealth Status			
Poor	75.28	24.72	805
Middle	81.09	18.91	386
Rich	84.76	15.24	794
Place of Residence			
Urban	84.49	15.51	780
Rural	77.43	22.57	1,205
Total			1,985

Machine Learning Algorithms

This project utilized several machine learning algorithms to predict the likelihood of stunting in children based on various predictors. The performance of each model was rigorously evaluated using classification reports that detail precision, recall, and F1 scores for each class, alongside overall accuracy. The analysis focused on enhancing our understanding of which factors most significantly predict stunting, aiming to aid interventions that could prevent or mitigate this condition.

The results (Table 3) from the machine learning models provided insightful findings into their capability to handle the binary classification task of predicting child stunting. Each model's performance was assessed based on its accuracy, recall, precision, and F1 scores, as shown in the classification reports. A detailed examination of each model's efficacy is provided below:

- Logistic Regression (LR): This model achieved an overall accuracy of 74%. It showed a precision of 41% for predicting stunting (class 1) and 93% for not stunting (class 0). The recall rates were 75% and 74% for stunting and not stunting, respectively, with corresponding F1 scores of 53% and 82%. The LR model, despite its simplicity, demonstrated a robust performance, especially in identifying non-stunting cases, but struggled slightly with stunting cases, suggesting a need for further tuning to balance recall and precision.
- Random Forest (RF): The Random Forest classifier achieved a higher accuracy of 82% with a precision of 55% for stunting and 86% for not stunting. However, its recall for stunting at 40% was lower compared to LR, indicating some difficulties in capturing all positive instances of stunting. The F1 scores were 46% for stunting and 89% for not stunting, highlighting its strength in identifying non-stunting instances but pointing to potential overfitting issues or the need for parameter optimization.
- Support Vector Machine (SVC): The SVC model reached an accuracy of 77%, with a stunting precision of 44% and non-stunting precision of 89%. It exhibited a recall of 56% for stunting, showing a modest ability to identify positive cases. The F1 scores of 49% for stunting and 86% for not stunting suggest that while SVC is fairly effective overall, the balance between sensitivity and specificity could be improved.
- K-Nearest Neighbors (KNN): This model showed the lowest overall accuracy of 73%. KNN's precision for stunting was notably lower at 38%, and the recall was 62%, resulting in an F1 score of 47%. The performance of KNN underscores some challenges typical of distance-based classifiers in this application, possibly due to the inherent sparsity of the dataset after encoding categorical variables.

Table 3

Performance Measurements of Machine Learning Models

Algorithm		Precision	Recall	F1 Score	Accuracy
Logistic Regression	Non-Stunt	0.93	0.74	0.82	0.74
	Stunt	0.41	0.75	0.53	0.74
Random Forest	Non-Stunt	0.86	0.92	0.89	0.82
	Stunt	0.55	0.40	0.46	0.82
Support Vector Machine	Non-Stunt	0.89	0.83	0.86	0.77
	Stunt	0.44	0.56	0.49	0.77
K-Nearest Neighbors	Non-Stunt	0.89	0.76	0.82	0.73
	Stunt	0.38	0.62	0.47	0.73

The evaluation indicates varied strengths and weaknesses across models. Logistic Regression, while simpler, provided a balanced performance, making it a potentially valuable tool for initial screenings in epidemiological studies of stunting. Random Forest excelled in overall accuracy and precision for non-stunting predictions, suggesting its usefulness in applications where reducing false positives is critical. SVC and KNN, though slightly less accurate, still hold promise for specific scenarios after further parameter adjustments and feature selection.

Table 4

Coefficients and Feature Importances

	Logistic Regression Coefficients	**Random Forest Feature Importances**
Age (in months)	1.047	0.243
Weight at birth (in kilos)	–0.997	0.249
Birth Order		
First Child		
2-3	0.161	0.018
4-6	0.686	0.010
7+	0.227	0.001
Birth Interval		
First Birth		
<2 years	0.284	0.022
2 years	0.082	0.016
3 years	–0.239	0.012
4+ years	–0.249	0.019
Sex of Child		
Boy		
Girl	–0.221	0.046
Higher Educational Level of Mother		
Primary		
Secondary	–0.385	0.028
Tertiary	–0.620	0.031
Age of Mother	0.303	0.119
Number of Children	–0.566	0.077
Wealth Status		
Poor		
Middle	–0.232	0.025
Rich	0.050	0.039
Place of Residence		
Urban		
Rural	0.255	0.045

Since the Logistic Regression and Random Forest performed relatively better in predicting stunting cases according to the performance metrics of algorithms used for analysis, I decided to go deep with these two models to analyze which predictors affects stunting the most in Uzbekistan. In this regard, I calculated the logistic regression coefficients and random forest feature importances for independent variables used in the model (Look at table 4).

The results derived from Logistic Regression and Random Forest models offer significant insights into the factors influencing stunting in children, with each model highlighting different aspects of influence due to their distinct methodological approaches.

From the logistic regression coefficients, it is apparent that age and weight at birth play critical roles in the likelihood of stunting. The positive coefficient for age suggests that as children grow older, the likelihood of stunting increases, possibly indicating cumulative environmental or nutritional deficiencies over time. Conversely, weight at birth shows a strong negative coefficient, implying that higher birth weights significantly decrease the risk of stunting. This underscores the importance of maternal health and nutrition during pregnancy, which directly affects birth outcomes.

The gender of the child also emerges as a noteworthy factor; female children are less likely to be stunted compared to males, possibly due to biological differences in growth patterns or variations in caregiving practices across genders. Additionally, the educational level of the mother is a potent determinant; mothers with higher education levels, particularly those with tertiary education, dramatically reduce the risk of stunting in their children. This effect likely stems from better health practices, improved nutrition, and greater access to healthcare services, which are more prevalent among better-educated individuals.

Wealth status further influences stunting, with children from poorer families facing a higher risk, although the coefficients for wealth are not as pronounced as those for other variables. The rural or urban setting also plays a role, with children in rural areas having a higher likelihood of stunting, potentially due to poorer access to healthcare and nutritional services compared to urban settings.

The feature importance scores from the Random Forest model reinforce some of the findings from the logistic regression analysis. Weight at birth is highlighted as the most significant predictor of stunting, confirming its critical role as observed in the regression model. Age also ranks highly, supporting the notion that older children are more susceptible to stunting if early-life interventions are not effective.

The model also gives importance to variables like the mother's age and the number of children, which were less emphasized in the logistic regression results. This indicates that younger maternal age and higher numbers of siblings might contribute to resource constraints or diminished care, thereby increasing stunting risk.

Combining insights from both models, it is evident that factors like age, weight at birth, maternal education, and socioeconomic status are integral to understanding and preventing stunting. The consistency across both analytical approaches in highlighting these factors attests to their robustness and the need for targeted interventions that address these areas. Efforts to improve maternal education and prenatal care, along with socioeconomic support for families, could significantly reduce the prevalence of stunting. Additionally, tailored strategies to support male children and those in rural settings may also be necessary to address the specific risks faced by these groups.

This analysis not only aids in pinpointing critical factors but also underscores the complexity of stunting, which is influenced by an interplay of genetic, environmental, and social determinants. Effective prevention strategies must therefore be multifaceted and context-specific, reflecting the nuanced nature of this global health challenge.

Limitations and Future Work

The current study, despite its insightful findings, is subject to several limitations that must be acknowledged. Firstly, the reliance on the dataset available may limit the generalizability of the results. While the dataset includes a variety of predictors, it may not capture all relevant factors that contribute to stunting, such as detailed dietary information, parental genetic factors, or more granular socioeconomic data. Moreover, the static nature of the data does not allow for the examination of changes over time in individual children, which could provide insights into dynamic risk factors for stunting.

Another critical limitation is the imbalance in the class distribution within the dataset, with far fewer instances of stunting compared to non-stunting. Despite the application of SMOTE to address this imbalance, the synthetic oversampling technique itself may introduce artificial patterns that do not exist in the real population, potentially leading to model overfitting or bias in predictions.

To build on the findings of this study and overcome its limitations, several recommendations are proposed for future research:

- Incorporation of Longitudinal Data: Future studies could benefit from using longitudinal datasets that track children over time. This would allow for the analysis of temporal patterns and the effects of interventions on stunting outcomes, providing a deeper understanding of when and how stunting can be most effectively prevented.

- Expansion of Predictor Variables: Including additional variables that capture more detailed nutritional data, genetic information, and micro-level socioeconomic factors could enhance the predictive accuracy of the models. Exploring environmental factors such as access to clean water and sanitation might also provide new insights into community-level determinants of stunting.
- Advanced Machine Learning Techniques: Exploring more sophisticated machine learning techniques that can handle complex interactions and nonlinear relationships without strong parametric assumptions could be valuable. Techniques such as deep learning or ensemble methods that combine multiple models to improve prediction accuracy should be considered.

By addressing these limitations and suggestions for future work, subsequent research can further enhance our understanding of the complex phenomenon of stunting and contribute to the development of more effective preventive and intervention strategies. This would not only extend the scientific understanding of child growth and development but also provide practical benefits in terms of public health policy and practice.

References

Caldwell, J., (1979). Education as a Factor in Mortality Decline: An Examination of Nigerian Data. *Population Studies* **33**(3): 395-413

Gwatkin, D. et al., (2000). Socio-economic differences in health, nutrition, and population in Burkina Faso. Washington DC., USA: NP/Poverty Thematic Group, the *World Bank*

Desai, S. and Alva, A., (1998). Maternal Education and Child Health: Is There a Strong Causal Relationship? *Demography* **35**(1): 71-81

Bairagi, R., (1980). Is Income the Only Constraint on Child Nutrition in Rural Bangladesh? *Bull World Health Organ* **58**(5): 767-72

Solon, F. et al., (1985). The Bulacan Nutrition and Health Study: Part 1. Baseline Socioeconomic and Related Characteristics of Subject Families and Their Impact on the Nutritional Health of Infants. *Ecology Food and Nutrition* **16**: 299-315

Mukuria, A. et al., (2005*). Nutritional Status of Children: Results from the Demographic and Health Surveys 1994-2001*. DHS Comparative Report No. 10. Calverton, MD, USA: ORC Macro

Frost, M. et al., (2005). Maternal Education and Child Nutritional Status in Bolivia: Finding the Links. *Social Sciences and Medicine* **60**(2): 395-407

Miller, J. and Rodgers, Y., (2009). Mothers' Education and Children's Nutritional Status: New Evidence from Cambodia. *Asian Development Review* **26**(1): 131-65.

Emina, J. et al., (2011). The Effect of Maternal Education on Child Nutritional Status in the Democratic Republic of Congo. *Journal of Public Health and Epidemiology* **3**(12): 576-592

Kabubo-Mariara, J. et al., (2009). Determinants of Children's Nutritional Status in Kenya: Evidence from Demographic and Health Surveys. *Journal of African Economies* **18**(3): 363–87