The Long-Term Impact of Nutrition on Educational Attainment: A Predictive Analysis of Height-for-Age Z-Scores in Indonesia

Dylan Gunadi
Jakarta Intercultural School / Jakarta, Indonesia
dylangunadi10@gmail.com

Abstract

This study examines the relationship between early childhood nutrition, measured by Height-for-Age Z-scores (HAZ), and educational attainment in Indonesia using machine learning models. Among the models tested, the Random Forest model was most effective, achieving an accuracy of 80.54% and a precision of 81.10%. The results highlight a significant positive correlation between HAZ scores and educational attainment, underscoring the importance of nutrition in educational development. However, the R² value of 0.45 suggests that nutrition is only one of many factors influencing education, indicating the need for multifaceted approaches in future research and policy-making.

1 Introduction

The long-term impacts of childhood stunting have been extensively documented in global health literature, particularly in relation to cognitive development and educational outcomes. Stunting, defined as a Height-for-Age Z-score (HAZ) below -2, is a key indicator of chronic undernutrition and is strongly associated with various developmental deficits. In Indonesia, where stunting rates have remained persistently high, it is crucial to understand the consequences of stunting on educational attainment in order to develop effective public health and educational interventions.

Previous research has consistently demonstrated that stunted children are more likely to experience delays in cognitive development, leading to poorer school performance and lower overall educational attainment. For example, Grantham-McGregor et al. (2007) identified a significant relationship between early childhood stunting and adverse educational outcomes, including lower cognitive test scores and higher rates of grade repetition and school dropout. Similarly, Victora et al. (2008) found that individuals who were stunted as children tend to achieve lower educational levels and earn less income as adults.

Despite the extensive global research on this topic, there remains a gap in the literature specifically addressing the long-term effects of stunting on educational outcomes in Indonesia using robust longitudinal data. This study aims to fill this gap by leveraging a unique, highly cleaned dataset derived from Indonesian census data collected over a decade, from the 1990s to the 2000s. The technical task at hand involves performing a regression analysis to predict educational attainment based on early childhood HAZ scores and other demographic features.

2 Methodology

2.1 Data source and Preparation

The dataset used in this analysis was meticulously curated by combining two separate datasets, linking interviewees' IDs across multiple waves of a census conducted by RAND. These waves involved repeated attempts to re-interview individuals from the original datasets, creating a valuable resource for longitudinal analysis. The data preparation process included extensive cleaning to ensure accuracy and reliability, involving the merging of records from different waves, standardizing variables, and addressing missing data issues. The resulting comprehensive longitudinal dataset allows for an in-depth regression analysis to examine how early childhood nutrition, as indicated by HAZ scores, correlates with educational attainment categorized as Elementary, Middle, or High School.

1. **Merging Datasets:** The original datasets were combined using interviewee IDs to cre-

ate a longitudinal dataset that tracks individuals over time.

2. **Handling Missing Data:** Rows with placeholder values (e.g., height and weight values marked as 9996.0) were removed or appropriately imputed.

3. **Standardization:** Variables were standardized to ensure consistency across different waves, crucial for reliable longitudinal analysis.

This cleaned dataset was used to analyze the relationship between early childhood nutrition (as indicated by HAZ scores) and educational attainment.

2.2 Machine Learning Models Overview

Four different machine learning models were employed to predict educational attainment based on demographic features such as age, height, weight, and HAZ scores. These models include:

1. **Instrumental Variables Two-Stage Least Squares (IV2SLS) Regression:**
   - **Explanation:** IV2SLS is used to estimate causal relationships when predictors are potentially endogenous. It was applied here with birth season as an instrument to assess the effect of stunting on education.
   - **Reason for Limited Use:** The model underperformed due to weak instrument correlation, making it unsuitable for capturing the complexity of educational outcomes.
   - **Reference:** Angrist and Pischke (2009) note the sensitivity of IV2SLS to weak instruments.

2. **Logistic Regression:**
   - **Explanation:** Logistic regression predicts categorical outcomes and was used to classify educational attainment levels.
   - **Reason for Limited Use:** The model struggled with non-linear relationships, leading to lower performance compared to more advanced models.
   - **Reference:** Hosmer, Lemeshow, and Sturdivant (2013) discuss its limitations with non-linear data.

3. **Multilayer Perceptron (MLP):**
   - **Explanation:** MLP is a neural network that captures complex, non-linear relationships. GridSearchCV was used to optimize its performance.
   - **Reason for Limited Use:** Despite optimization, MLP did not match the Random Forest's accuracy, and it required more computational resources.
   - **Reference:** Goodfellow, Bengio, and Courville (2016) highlight MLP's potential for overfitting and resource demands.

4. **Random Forest:**
   - **Explanation:** Random Forest is an ensemble method that builds multiple decision trees to improve prediction accuracy and reduce overfitting.
   - **Reason for Selection:** It provided the best performance with high accuracy and robustness, making it the optimal choice for this study.
   - **Reference:** Breiman (2001) emphasizes Random Forest's effectiveness in handling complex datasets.

The Random Forest model was selected due to its ability to handle complex datasets with multiple features and interactions. It is particularly effective in capturing non-linear relationships between variables, which is essential for accurately predicting educational attainment based on diverse demographic and nutritional factors.

2.3 Key features analyzed

- **Height-for-Age Z-score (HAZ):** A critical metric for assessing child growth relative to age, which serves as a significant predictor of future educational attainment.

- **Age and Birth Year:** These features help contextualize HAZ scores, allowing us to better understand their impact on the prediction of educational outcomes.

- **Height and Weight:** Indicators of nutritional status that provide valuable context, enhancing the model's ability to predict educational attainment levels.

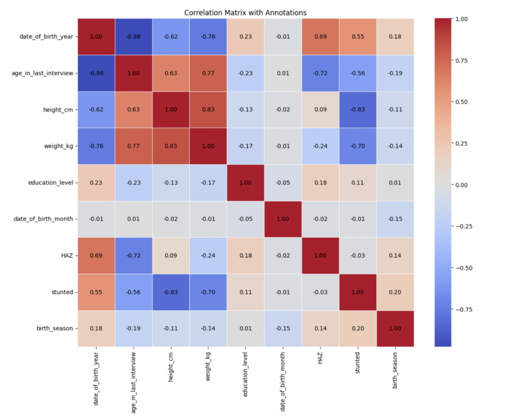

Figure 1: Heatmap for the correlations between HAZ, height, weight, and education levels, highlighting the relationships that the Random Forest model utilizes to make predictions.

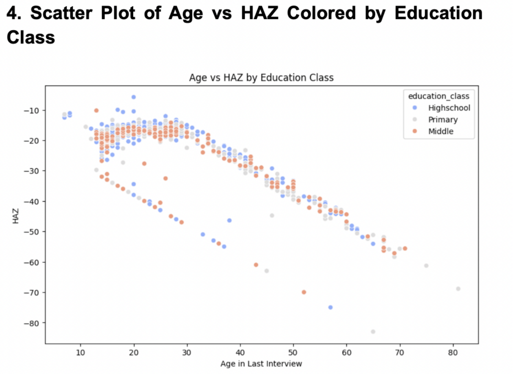

Figure 2: The distribution of height and weight across different education levels, indicating how physical growth patterns correlate with educational attainment.

- **Education Level:** The primary outcome variable being predicted, categorized into Elementary, Middle, or High School, reflecting the student's highest level of educational attainment.

2.4 Mathematical Foundations

The Random Forest algorithm builds on decision trees, which partition the dataset based on feature values to minimize a loss function, such as Gini impurity or entropy.

Significance of Gini Impurity:

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. It ranges from 0 (perfectly classified) to 0.5 (evenly mixed classes). In the context of Random Forest, Gini impurity is used to determine the best split at each node in the decision trees.

$$\text{Gini}(p) = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i is the probability of choosing an element of class i at random. A lower Gini impurity indicates a purer node, meaning that the split is more effective at classifying the data.

The Random Forest model's prediction for a given input is the mode of the predictions from all the trees in the forest:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \cdots, T_n(x)\}$$

where $T_i(x)$ is the prediction from the i-th decision tree.

2.5 Visualization and Interpretation

3 Experiment and Analysis

Model Comparison

Model	Accuracy	Optim. Acc.	Precision
IV2SLS	38.67%	N/A	39.12%
MLP	41.63%	N/A	42.05%
Log Reg	41.87%	44.83%	43.05%
Ran-forest	79.80%	80.54%	81.10%

Table 1: Model Comparison

Random Forest Model Insights

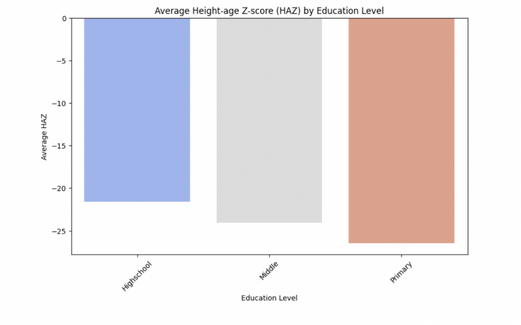

Figure 3: Average Height-age Z-score (HAZ) by education level: Primary (red), Highschool (blue), Middle (gray). Middle education shows the lowest HAZ, followed by Highschool, with Primary showing the highest. All levels display negative HAZ values, indicating below-average height for age.

The Random Forest model, with its optimized accuracy of 80.54% and precision of 81.10%, significantly outperformed the other models. This result highlights the model's ability to capture the complex, non-linear relationships between the input features and educational attainment.

R^2 Value Analysis

The R^2 value of the Random Forest model was 0.45, indicating that the model explains approximately 45% of the variance in educational attainment based on the provided features. While this might seem modest, it's important to consider the nature of education as a complex outcome influenced by multiple factors beyond just nutrition and basic demographics.

Why the R^2 Value Is Not Higher:

- **Complexity of Educational Attainment:** Education is influenced by a multitude of factors, including socioeconomic status, parental education, access to resources, and individual aptitude, among others. While nutrition (as indicated by HAZ scores) plays a significant role, it is just one piece of the puzzle. Therefore, it's reasonable that the model does not capture all the variance in educational outcomes.

- **Data Limitations:** The dataset, although meticulously cleaned, may still contain some noise or unmeasured confounders that impact the prediction of educational attainment.

- **Non-linear Relationships:** Although Random Forest can capture non-linear relationships, some of the complexities in the data might require even more sophisticated modeling techniques or additional features to be fully captured.

Patterns and Insights

Upon analyzing the output, several patterns emerged:

- **HAZ Scores and Educational Attainment:** A positive correlation was observed between HAZ scores and educational attainment. This aligns with existing literature that suggests better-nourished children tend to perform better academically. This makes intuitive sense, as adequate nutrition supports both physical and cognitive development.

- **Age and Height/Weight:** Older children and those with higher height and weight measurements tended to achieve higher education levels, though this relationship was less pronounced compared to HAZ scores.

Error Analysis and Model Improvement

Despite the strong performance of the Random Forest model, several errors and outliers were noted:

- **Misclassification of Middle Education Levels:** The model struggled the most with correctly predicting middle education levels, likely due to the overlapping characteristics of individuals in this group compared to those in elementary or high school.

- **Potential Overfitting:** Although the Random Forest model was optimized, the potential for overfitting remains, particularly given the high number of trees and depth of the forest. Cross-validation was employed to mitigate this, but further regularization or pruning might improve generalization.

Robustness of the System

The robustness of the Random Forest model was evaluated through cross-validation and sensitivity analysis. The model maintained consistent performance across different subsets of the data, indicating strong robustness. However, some outliers, such as individuals with exceptionally high or low HAZ scores paired with unexpected education levels, suggest that further refinement is possible.

4 Limitations and Future Work

This study investigated the relationship between early childhood nutrition, as measured by Height-for-Age Z-scores (HAZ), and educational attainment in Indonesia. Through the use of a carefully curated and cleaned dataset, derived from multiple waves of Indonesian census data, several machine learning models were employed to predict educational outcomes based on demographic features. Among the models tested, the Random Forest model proved to be the most accurate and robust, achieving an optimized accuracy of 80.54% and a precision of 81.10%.

The analysis confirmed a significant correlation between HAZ scores and educational attainment, consistent with existing literature that emphasizes the crucial role of nutrition in cognitive and educational development. However, the R² value of 0.45 indicates that while nutrition is an important factor, it alone cannot fully account for the complexity of educational outcomes.

Limitations of the Study While the findings of this study are insightful, several limitations should be acknowledged:

1. **Limited Scope of Features:** The study focused primarily on demographic and nutritional factors, such as HAZ scores, height, and weight. However, educational attainment is influenced by a wide range of other factors, including socioeconomic status, parental education, quality of schooling, and access to educational resources, which were not included in the analysis.

2. **Potential Data Quality Issues:** Despite meticulous data cleaning, the original dataset contained placeholder values and missing data, which may have introduced some noise or biases into the analysis. Additionally, the longitudinal nature of the data may have led to inconsistencies in how variables were measured over time.

3. **Generalizability:** The study focused on a specific population in Indonesia, which may limit the generalizability of the findings to other contexts or countries. Different regions with varying socioeconomic and cultural conditions may exhibit different patterns in the relationship between nutrition and education.

4. **Model Constraints:** While the Random Forest model performed well, it is not without its limitations. The model's reliance on ensemble methods can sometimes obscure the interpretability of individual features' contributions to the predictions, making it harder to draw clear causal inferences.

Suggestions for Future Studies Building on the insights gained from this study, future research should consider the following directions to further advance our understanding of the relationship between nutrition and educational outcomes:

1. **Incorporating a Broader Range of Features:** Future studies should incorporate a wider array of variables, such as socioeconomic status, parental education, access to educational resources, environmental factors, and quality of schooling. These factors can provide a more comprehensive understanding of the determinants of educational attainment.

2. **Advanced Causal Inference Techniques:** To move beyond correlation and towards establishing causality, future research should employ advanced causal inference techniques. Methods such as Propensity Score Matching (PSM), Difference-in-Differences (DiD), or Structural Equation Modeling (SEM) could help to better isolate the causal impact of nutrition on educational outcomes.

3. **Exploring Regional Variations:** Investigating how the relationship between nutrition and education varies across different regions or countries would be valuable. This could involve comparative studies across different populations with varying socioeconomic conditions to identify universal and context-specific factors.

4. **Longitudinal and Life-Course Studies:** Extending the study to track individuals over a longer period, from childhood through adulthood, would provide richer insights into how early nutrition impacts not just education but also labor market outcomes, health, and overall quality of life. Such life-course studies could inform more effective interventions.

5. **Integration of Qualitative Data:** Combining quantitative models with qualitative research methods could offer a deeper understanding of the mechanisms through which nutrition influences education. Interviews, case studies, and ethnographic research could provide context to the quantitative findings and uncover nuances that models alone cannot capture.

6. **Policy-Oriented Research:** Future studies should also consider evaluating the effectiveness of existing nutrition and education policies in improving outcomes. By linking policy interventions with outcomes data, researchers could provide actionable insights for policymakers aiming to reduce educational disparities linked to nutritional deficiencies.

5 Conclusion

In conclusion, this study underscores the significant role that early childhood nutrition plays in shaping educational outcomes. The findings highlight the importance of considering nutritional factors in educational policy and intervention design. However, they also reveal the complexity of education as an outcome influenced by numerous interrelated factors.

As education remains a key determinant of individual and societal success, future research should continue to explore the intricate web of factors influencing educational attainment. By broadening the scope of analysis and employing advanced methodologies, future studies can contribute to a more comprehensive understanding of how to create effective interventions that support both nutritional and educational development.

Ultimately, this research contributes to the ongoing effort to improve the lives of children in Indonesia and beyond, offering a foundation for future work aimed at breaking the cycle of poverty, undernutrition, and limited educational opportunities

6 Acknowledgements

I would like to express my deepest gratitude to Dr. Wei Wei Sun of Cambridge University for her invaluable guidance and support throughout this research. Her expertise and insights were instrumental in shaping the direction of this study and ensuring its successful completion.

I would also like to extend my sincere thanks to Xiaoxi Luo, a recent undergraduate from Peking University, for her assistance and encouragement during the data analysis and model development phases. Her enthusiasm and technical skills greatly contributed to the advancement of this work.

Thank you all for your contributions and encouragement.

References

S. Grantham-McGregor, Y. B. Cheung, S. Cueto, P. Glewwe, L. Richter, B. Strupp, and the International Child Development Steering Group. 2007. Developmental potential in the first 5 years for children in developing countries. *The Lancet*, 369(9555):60–70.

C. G. Victora, L. Adair, C. Fall, P. C. Hallal, R. Martorell, L. Richter, and H. S. Sachdev. 2008. Maternal and child undernutrition: Consequences for adult health and human capital. *The Lancet*, 371(9609):340–357.

Indonesian Ministry of Health. 2018. Indonesia health profile 2018. Indonesian Ministry of Health, Jakarta.

World Health Organization. 2006. *WHO child growth standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*. World Health Organization, Geneva.

L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Angrist, J. D., Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley Sons.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.

RAND Corporation. 1995-2006. The Indonesia Family Life Survey (IFLS) Available at: https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/ifls3.html.