EARLY CHILDHOOD CHARACTERISTICS AS PREDICTORS OF NON-COGNITIVE SKILLS OF VIETNAMESE: A MACHINE LEARNING APPROACH

Nguyen Ngoc Nhat* - Emmanuel Lance Christopher VI M. Plan*

Abstract: ***Grasping how childhood experiences affect the development of non-cognitive skills allows policymakers to create supportive structures to ensure positive progression. This study explores whether or not non-cognitive skills are associated with various early childhood characteristics. This research focuses on exploring the data from The Young Lives Study and utilizes models using Artificial Neural Networks (ANNs) to predict a subject's non-cognitive skills. Shapley Values are employed to identify the best predictors. Results highlight that health indicators, especially Body Mass Index (BMI), have the strongest and most widespread impact. This highlights that physical health is imperative for socio-emotional health development. Parental/Caregiver Characteristics and Household Characteristics follow in order of significance, proving the importance of family and home background in developing these skills.***

• Keywords: *non-cognitive skills, early childhood development, artificial neural networks, shapley values, young lives study.*

JEL codes: I18, I28, I31, O15

Date of receipt: 25th Feb., 2025
Date of delivery revision: 04th Mar., 2025
Date of receipt revision: 30th May, 2025
Date of approval: 30th Jun., 2025
DOI: https://doi.org/10.71374/jfar.v25.i4.28

1. Introduction

Non-cognitive and cognitive skills are crucial for individual development. Empirical studies increasingly show that non-cognitive skills possess predictive power for life outcomes that are equal to or even surpass that of cognitive skills. For example, Heckman and Rubinstein's (2001) analysis of General Educational Development testing program recipients demonstrated that while cognitive test scores were similar to high school graduates, poorer life outcomes highlighted the critical influence of non-cognitive skills like perseverance and motivation, which cognitive tests did not capture.

Cognitive skills encompass abilities related to handling abstract problems and approaching measures in various ways, depending on their type and content. Non-cognitive skills, in contrast, have a broader definition and are defined using various terms, including generic competencies, life skills, and socio-emotional skills, and encompass a collection of traits, behaviors, mindsets, and attitudes. Studies emphasize that these skills are valuable assets for both academic outcomes and broader personal development (Anghel & Balart, 2017). These skills cover a spectrum of traits and personality characteristics that significantly influence educational attainment, occupational success, and personal development throughout the lifespan (Heckman & Rubinstein, 2001).

Through many empirical and experimental studies, a strong consensus emphasizes that early characteristics significantly influence later outcomes in several aspects, such as education, economic status, health, and behavior. Research has shown that non-cognitive skills play a vital role in educational and labor market achievements, are influenced by factors like parenting style, and can be more predictive of certain developmental outcomes than cognitive skills (Fletcher and Wolfe, 2016). Heckman et al. (2006) highlighted that there is compelling evidence that non-cognitive skills are critical contributors to success in the labor market, shaping educational paths and employment opportunities. These studies are, however, few in number since cognitive skills are often easier to measure. Building on these gaps, this study focuses on the following questions:

- What early childhood factors have the most significant effect on the development of non-cognitive skills?

- Do different non-cognitive skills share the same "important" predictors, or do these factors vary across skill types?

** Hanoi School of Business and Management - Vietnam National University, Hanoi; email: emmanuelplan@hsb.edu.vn*

2. Related work

According to O'Connell and Shaikh (2008), "Achievement-related skills" can be divided into cognitive and 'non-cognitive' abilities. Molnár, G., & Kocsis (2023) conducted a study to predict academic success in higher education by using cognitive methods, and found that the most significant predictor of later academic success relates not only to cognitive skills but also to non-cognitive skills, such as motivation to learn and the ability to apply effective strategies. Besides, Al-Sheeb et al. (2019) also used both skills to show that a combination of cognitive and non-cognitive factors contributed approximately 30-40%, while non-cognitive factors contributed 25% of the variance in students' GPA.

Several studies have used regression models to investigate the influence of family background and the home learning environment from age 11 on the development of both non-cognitive and cognitive skills, as well as later life outcomes. For instance, Carneiro et al. (2007) examined data from the National Child Development Survey. They reported that anxiety for acceptance or hostility towards adults plays a significant role in various outcomes, and that family background influences both cognitive and non-cognitive skill development. Similarly, Duckworth et al. (2019) employed regression approaches on cognitive ability, grit, and physical ability from 10,000 cadets at the US Military Academy at West Point to predict GPA and binary outcomes like successful graduation. Their study found that grit was a reliable predictor only for completing the initial summer training.

Some studies use machine learning (ML) to predict outcomes from non-cognitive and cognitive skills. Mareckova et al. (2019) used Cluster Analysis and Group Lasso to predict differences in individual unemployment. This study emphasizes the significance of self-esteem, agreeableness, and emotional stability in predicting long-term unemployment. Musso et al. (2013) also proposed using an ANN with a backpropagation multilayer perceptron neural network and predicted different levels of GPA in university students based on cognitive variables like working memory and learning strategies, and non-cognitive factors like family-social background factors and background information. The study found that oriented attention, cognitive resources, time management, and executive control were the most important for predicting. These studies explore and highlight the diverse methods being applied to demonstrate the importance of both cognitive and non-cognitive skills.

While existing research has extensively explored the impact of cognitive and non-cognitive skills on various life outcomes, and some studies have employed ML techniques in this domain (Mareckova et al., 2019), a notable limitation remains. Specifically, there is a scarcity of research that compares the predictors of different non-cognitive skills using advanced ML methodologies. Much of the current understanding relies on traditional statistical methods like regression, which may not fully capture the complex, non-linear relationships and interactions between early childhood factors and non-cognitive abilities.

3. Methodology

3.1. Data collection

This study uses data from the Young Lives Study, a longitudinal study of poverty and inequality that has followed the lives of 12,000 children in four countries (Ethiopia, India, Peru, and Vietnam) since 2001. The Young Lives Study provides data collected across five rounds, from Round 1 in 2001 to Round 5 in 2016, with two cohorts: the Younger cohort (age 1 in Round 1 to age 15 in Round 5) and the Older cohort (age 8 in Round 1 to age 22 in Round 5). The Young Lives Study is particularly well-suited for this research: its longitudinal design allows examination of how early childhood factors at age 1 influence non-cognitive skill development at age 15. The Vietnam dataset was chosen for exploration in this study using data collected from Round 1 and Round 5. Data from Round 1 was divided into 5 main categories associated with early childhood:

- *Child Demographic Characteristics* covers demographic information, with age at 1 year showing minimal variation (Variables: Gender).

- *Child Health Status* is vital since physical health affects energy levels and engagement, influencing non-cognitive skills. Stunting has been linked to deficits in motor skills and social skills. (Variables: Short height for age, Underweight status, BCG vaccination status, Measles vaccination status, BMI index,…).

- *Parental/Caregiver Characteristics* are important for understanding how early caregiving influences skill development. Berger et al. (2019) found that parental education correlates with children's mental health, a proxy for non-cognitive skills. (Variables: Caregiver's ability to read, Caregiver's education level, Gender of caregiver, Age of the father at R1,…).

- *Household Characteristics* give insight into socioeconomic status. Fletcher and Wolfe (2016) found that lower family income is associated with reduced non-cognitive skills. (Variables: Household size, Age of the household head, Education level of the household head, Access to safe drinking water, Access to electricity, Access to services index,…).

- *Perceptions and Relationships* captures parental perceptions and child-parent relationships. (Variables: Potentially life-threatening injury in early childhood, … Child's health is perceived as worse than other children's, Children see their mother every day)

Next, this study uses the Round 5 dataset as the result of the target variables (z_ refers to standardization):

• **chhealth**: Cognitive and non-cognitive skills can significantly influence socioeconomic trajectories and health.

• **z_selfefficacy**: Self-efficacy is the belief in one's ability to succeed in specific situations or accomplish tasks.

• **z_agency**: Agency refers to the capacity of individuals to act independently and make their own choices, measured as a standardized score.

• **z_selfsteem**: Self-esteem is fundamental to mental health, influencing behavior, relationships, and performance, a key non-cognitive skill.

• **z_peers**: This variable measures aspects of peer relationships or interactions, such as social skills or peer acceptance.

• **z_pride**: Pride is a positive emotion associated with self-accomplishment or group identity.

• **z_relationparents**: This measures the quality of the parent-child relationship, such as closeness or conflict.

3.2. Artificial Neural Networks and Shapley Values

This study proposes to apply Artificial Neural Networks(ANNs) and Shapley Values to explore and understand the power of early childhood factors to predict different non-cognitive skills. ANNs are a computational structure consisting of several interconnected computational elements, known as neurons, and each unit carries out a very simple operation on its inputs and transfers the output to a subsequent node or nodes in the network topology (Specht, 1991). An ANN architecture includes an input layer (representing independent variables), one or more hidden layers, weight connections between nodes, and an output layer (representing the dependent variables). ANNs are particularly effective at capturing nonlinear, high-dimensional data, such as the complex interactions between young children's traits and non-cognitive skills. The predictive capability of the ANN model as used here was improved by modifying the parameters that determine the rate of learning, the persistence, momentum, and stopping criteria, and the type of functions used for weight adjustments.

This study uses Shapley values to address key research questions regarding the influence of early childhood characteristics on non-cognitive skills. Shapley values provide a mathematically rigorous approach to fairly distribute the predictive contribution of an ANN model among its input features, offering clear insights into their relative importance. This method arises from game theory, wherein it fairly attributes the total payoff from a cooperative game to the game's players (Shapley, 1953). The use of Shapley values in machine learning for social sciences and education is gaining traction due to their interpretability.

4. Results

4.1. Descriptive Analysis, data preprocessing, and dimension reduction

The Vietnam dataset includes 1980 children's information from Round 1 and Round 5. The data was first preprocessed: null values and duplicates were removed, and the data was properly formatted. After cleaning, the final sample consists of 1192 records. Sample descriptive statistics for the dataset are reported in Table 1 below (full data are not shown here but are available upon request).

Data was split using an 80-20 split and was normalized. For each non-cognitive skill, a RandomForestRegressor was trained based on the training data due to computationally efficient and robust feature importance estimates. Then, the absolute mean Shapley Values across all training samples were calculated for each feature. Features with importance above the 40th percentile are selected for each output, which reduces the input dimensionality for each output. Seventeen features were chosen for each skill.

Table 1. Summary Statistics for select variables

Variable	Mean	Std
female	0.4823	0.499899
stunting	0.115772	0.262430
bmi	8.404910e-16	1.000420
dadedu	0.482143	0.281724
momedu	0.452661	0.282187
sees_dad_daily	0.910235	0.285965
sees_mom_daily	0.997483	0.050125
z_selfefficacy_r5	0.042852	1.003331
z_agency_r5	0.094939	0.957992
z_selfsteem_r5	-0.006257	1.009678
z_peersr5	0.009710	1.027226
z_pride_r5	0.057601	0.999476
z_relationparents_r5	0.052514	1.002729
chhealth5	0.589346	0.152944

4.2. Neural network analyses

Next, to predict all non-cognitive skills simultaneously, a multi-output neural network was designed with two hidden layers (64 and 32 neurons, ReLU activation), 30% dropout, and 7 output neurons. Besides, a single-output model following the same structure was trained for each non-cognitive skill. Models were optimized and trained for up to 200 epochs with a batch size of 32, and early stopping

(patience of 20 epochs) to prevent overfitting. With the Adam optimizer (learning rate 0.001) and MSE loss, the model provided the best result, with an average for all non-cognitive skills R2 is 72%, and a separate model for each output with selected features, which gives an average R2 is 80%. After that, feature importance was quantified as the percentage contribution of each feature to predictions, based on Shapley Values. Importance was averaged across outputs, and the top 5 features were identified in Figure 1.

Figure 1. Average importance of features (Top 5)

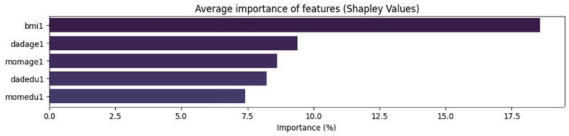


Table 2 shows that bmi has the highest average influence, highlighting the substantial impact of body mass index on later outcomes. Notably, several other features in the top 15, such as parental age (dadage and momage) and parental education (dadedu and momedu), also demonstrate considerable influence. This suggests that parental background and investment, alongside health factors like bmi, play a crucial role in shaping later life outcomes.

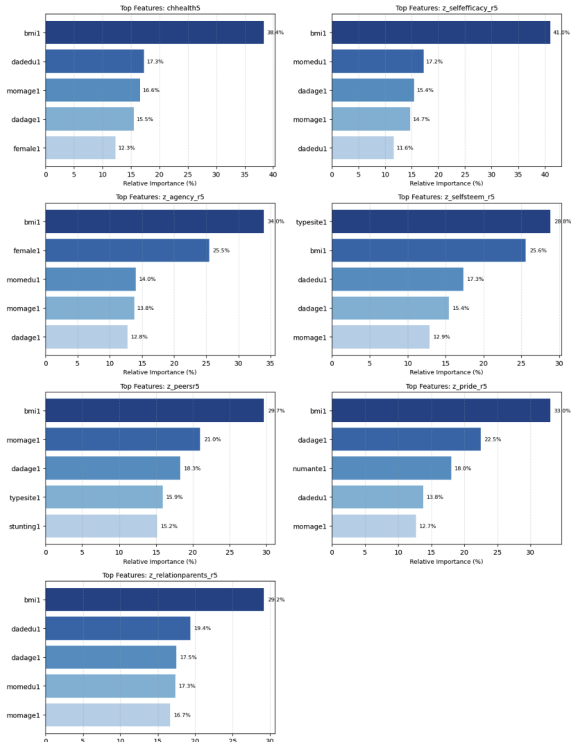
Table 2: Top 5 important features

Features	Mean importance(%)	Category
bmi	18.5595	Child Health Status
dadage	9.3820	Caregiver Characteristics
momage	8.6016	Caregiver Characteristics
dadedu	8.2219	Caregiver Characteristics
momedu	7.4014	Caregiver Characteristics

For each non-cognitive skill, the top 5 features were also identified (Figure 2). Notably, body mass index (bmi) influences all seven of the measured outcomes. BMI (bmi1) appears as the most contributing factor, affecting all outputs, with particularly strong impacts on self-efficacy (23.28%) and general health perception (22.60%). The broad influence of BMI indicates that interventions targeting physical health may have cascading benefits across multiple developmental domains. Furthermore, father's age (dadage) affects four outcomes, father's education (dadedu) impacts three, while mother's age (momage) and mother's education (momedu) each influence two outcomes.

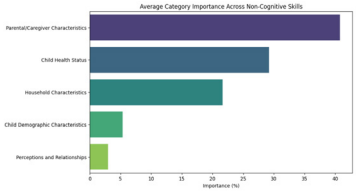
Parental characteristics show differentiated impact patterns. Father's age affects four skills, with the strongest influence on pride (13.34%), while mother's age impacts only two skills, most notably peer relationships (10.70%). This gender difference in parental influence may reflect distinct socialization roles, where fathers' characteristics relate more to internal motivational factors (self-efficacy, pride) while mothers' age shows a stronger association with social outcomes. Similarly, education levels demonstrate parent-specific effects: father's education primarily relates to health and relationship outcomes, whereas mother's education shows stronger connections with self-efficacy and agency.

Figure 2. Top 5 features for each skill



Lastly, since inputs were grouped into five categories, this study also wants to explore which groups have the most influence on each non-cognitive skill. Importance (measured by cumulative Shapley values for each category) was normalized to percentages to enable comparison across outputs, then the top category for each output was identified, and categories ranking in the top three for multiple outputs were analyzed to detect broad-impact predictors (Figure 3). The results show that Caregiver Characteristics make up the greatest level among other categories at more than 40%, and perceptions and Relationships have the least contribution to later outcomes.

Figure 3. Categories important across non-cognitive skills



5. Discussion and Conclusion

This study aimed to explore both applications of using ANN to predict factors affecting children at an early age that have the most impact on later outcomes,

and to identify and understand the contribution of each factor to each outcome. The neural networks achieved a good predictive performance, with average R^2 scores of 72% for the full model (using all features) and 80% for single-output models (using Shapley-selected features). These results indicate that child health indicators, caregiver education, and household socio-economic status can reliably predict non-cognitive skills measured later. The single-output models, which used a reduced set of features selected via Shapley Values, performed even better, suggesting that feature selection effectively reduced model complexity without sacrificing accuracy. This balance between parsimony and performance is particularly valuable in applied settings, where simpler models enhance interpretability for policymakers.

The feature importance analysis revealed that certain early-life factors play a disproportionate role in shaping non-cognitive skills. The top feature overall, BMI, with an average importance of 18.5%, underscores the critical influence of health on developmental outcomes. This agrees with previous results where higher body weight is associated with lower social and emotional skills (Guerra et al. 2022). Moreover, mom and dad's education level (with 8.2% and 7.4% respectively) shows results aligning with prior research where parental education is a key determinant of child psychological and social development (Heckman, 2006). Other high-ranking features, such as household size, stunting, and service index, point to the combined importance of health and environmental factors.

The identification of common features between pairs of non-cognitive skills, such as self-efficacy & self-esteem, which share 5 features (e.g., mom and dad's age and edu, BMI), highlights overlapping predictors. For instance, BMI contributed to all outcomes, suggesting that health fosters almost all outcomes, such as confidence, self-worth, and pride, through similar mechanisms. These cross-skill relationships suggest that non-cognitive skills are not developed in isolation but are interconnected through shared environmental and social influences.

At the category level, Caregiver Characteristics, Child Health Status, and Household Characteristics emerged as the most influential across all non-cognitive skills. While the feature importance analysis highlighted BMI within the Child Health Status category as the single most influential factor, the Caregiver Characteristics category as a whole demonstrates the greatest cumulative influence on non-cognitive skills. This category, encompassing factors like caregiver literacy and parental education, underscores the impact of the caregiving environment. The Child Health Status category was unsurprisingly dominant for chhealth5. For psychological skills, Caregiver Characteristics was the leading category alongside health factors, further emphasizing the role of nurturing relationships and parental background in fostering confidence and self-worth. Household Characteristics also contributed to later non-cognitive skills.

This study has implications for social, policy, and research domains by stressing the impact of early health and parental factors in shaping non-cognitive skills. Targeted interventions can be done towards nutrition, education, and household stability to foster a resilient future generation. Investments in early childhood nutrition programs, quality preschool education, and initiatives improving parental education and mental health are crucial, aligning with research suggesting the enduring impact of the early environment. Besides, businesses and policymakers can leverage these insights within the framework of the Sustainable Development Goals, thus providing targeted social programs and long-term community investments.

Lastly, this study, while insightful, faces limitations due to a limited dataset from Vietnam. To strengthen these results, future research should aim to confirm them using larger and more varied groups of people, such as data from other countries (Ethiopia, India, Peru) within the same study. Considerably, the potential to utilize data from other rounds (6 and 7) of The Young Lives Study could provide a longitudinal perspective on the development of these relationships over time. It would also be valuable to investigate the cause-and-effect link between factors identified and the development of non-cognitive skills.

Acknowledgements: *The authors would like to thank Micole De Vera, Javier Garcia-Brazales, and Clark Kendrick Go for providing data access and useful discussions.*

References:

Al-Sheeb, B. A., Hamouda, A. M., & Abdella, G. M. (2019). Modeling of student academic achievement in engineering education using cognitive and non-cognitive factors. Journal of Applied Research in Higher Education, 11(2), 178-198. https://doi.org/10.1108/JARHE-10-2017-0120

Anghel, B., & Balart, P. (2017). Non-cognitive skills and individual earnings: new evidence from PIAAC. SERIEs, 8(4), 417-473. https://doi.org/10.1007/s13209-017-0165-x

Berger, L. M., & Houle, J. N. (2019). Rising household debt and children's socioemotional well-being trajectories. Demography, 56, 1273-1301. https://doi.org/10.1007/s13524-019-00800-7

Carneiro, P., Crawford, C., & Goodman, A. (2007). The impact of early cognitive and non-cognitive skills on later outcomes.

Duckworth, A. L., Quirk, A., Gallop, R., Hoyle, R. H., Kelly, D. R., & Matthews, M. D. (2019). Cognitive and noncognitive predictors of success. Proceedings of the National Academy of Sciences, 116(47), 23499-23504. https://doi.org/10.1073/pnas.1910510116

Fletcher, J. M., & Wolfe, B. (2016). The importance of family income in the formation and evolution of non-cognitive skills in childhood. Economics of education review, 54, 143-154. https://doi.org/10.1016/j.econedurev.2016.07.004

Guerra, C., & Huneeus, F. (2022). Bodyweight and human capital development: Assessing the impact of obesity on socioemotional skills during childhood in Chile. Economics & Human Biology, 46, 101146. https://doi.org/10.1016/j.ehb.2022.101146

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. American economic review, 91(2), 145-149. 10.1257/aer.91.2.145

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor economics, 24(3), 411-482. 10.1086/504455

Molnár, G., & Kocsis, Á. (2023). Cognitive and non-cognitive predictors of academic success in higher education: a large-scale longitudinal study. Studies in Higher Education, 49(9), 1610–1624. https://doi.org/10.1080/03075079.2023.2271513

Musso, M. F., Kyndt, E., Cascallar, E. C., & Dochy, F. (2013). Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks. Frontline Learning Research, 1(1), 42-71. 10.14786/flr.v1i1.13

Shapley, L. S. (1953). A value for n-person games. doi.org/10.1515/9781400829156-012