AppliedMath

MDPI

*Article*

# Predicting the Cognitive and Social–Emotional Development of Minority Children in Early Education: A Data Science Approach

**Danail Brezov [1,*], Nadia Koltcheva [2] and Desislava Stoyanova [3]**

[1] Department of Mathematics, University of Architecture, Civil Engineering and Geodesy, 1164 Sofia, Bulgaria
[2] Department of Cognitive Science and Psychology, New Bulgarian University, 1618 Sofia, Bulgaria; nkoltcheva@nbu.bg
[3] Master Program 'Cognitive Science and Psychology', New Bulgarian University, 1618 Sofia, Bulgaria
* Correspondence: brezov_fte@uacg.bg

**Abstract**

Our study tracks the development of 105 Roma children between 3 and 5 (median age: 51 months), enrolled in an NGO-aided developmental program. Each child undergoes pre- and post-assessment based on the Developmental Assessment of Young Children (DAYC), a standard tool used to track the progress in early childhood development and detect delays. Data are gathered from three sources, teacher, parent/caregiver and specialist, covering four developmental domains and adaptive behavior scale. There are subjective biases; however, in the post-assessment, the teachers' and parents' evaluations converge. The test results confirm significant improvement in all areas ($p < 0.0001$), with the highest being in cognitive skills 32.2% and the lowest being in physical development 14.4%. We also apply machine learning methods to impute missing data and predict the likely future progress for a given student in the program based on the initial input, while also evaluating the influence of environmental factors. Our weighted ensemble regression models are coupled with principal component analysis (PCA) and yield average coefficients of determination $R^2 \approx 0.7$ for the features of interest. Also, we perform k-means clustering in the plane cognitive vs. social–emotional progress and consider the classification problem of predicting the group in which a given student would eventually be assigned to, with a weighted $F_1$-score of 0.83 and a macro-averaged area under the curve (AUC) of 0.94. This could be useful in practice for the optimized formation of study groups. We explore classification as a means of imputing missing categorical data too, e.g., education, employment or marital status of the parents. Our algorithms provide solutions with the $F_1$-score ranging from 0.92 to 0.97 and, respectively, an AUC between 0.99 and 1.

**Keywords:** childhood development; DAYC; in-service effect; differential learning; principal component analysis; data imputation; predictive modeling

**MSC:** 91E45; 91E10; 62P15

check for updates

Academic Editors: Tommi Sottinen and Takayuki Hibi

Received: 2 June 2025
Revised: 8 July 2025
Accepted: 21 July 2025
Published: 1 September 2025

**Citation:** Brezov, D.; Koltcheva, N.; Stoyanova, D. Predicting the Cognitive and Social–Emotional Development of Minority Children in Early Education: A Data Science Approach. *AppliedMath* **2025**, *5*, 113. https://doi.org/10.3390/appliedmath5030113

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Human development is a dynamic process shaped by the interaction of several core domains: physical development (motor skills), cognitive (thinking, perception, problem-solving), social and emotional (expression, regulation, and interpersonal understanding), and language acquisition. During early childhood, critical neural structures form, laying the foundation for curiosity, self-esteem, and the capacity to engage meaningfully with the surrounding environment. Given the importance of this developmental stage, targeted

interventions can have lasting positive impacts on both individuals and society as a whole. Economic analyses strongly support the cost-effectiveness of early childhood development (ECD) programs. For instance, the High/Scope Perry Preschool Program in the United States demonstrated that long-term economic returns range from USD 7 to USD 16 for every dollar invested [1]. Children who took part in the program demonstrate better high school performance and higher graduation rates, as well as higher earnings and rates of home ownership at age 27. Moreover, they are less likely to rely on welfare, be arrested, or experience out-of-wedlock births. Recent findings further emphasize the value of investing specifically in social and emotional development; for example, ref. [1] highlights that such investments not only support individual well-being, but also contribute to broader societal progress. By equipping children with skills to navigate social interactions and regulate emotions, communities reduce behavioral issues and improve mental health and overall quality of life.

### *1.1. Delays in Early Childhood Development*

Addressing developmental delays early in life can profoundly influence a child's cognitive and socio-emotional trajectory, laying the groundwork for improved academic achievement and long-term psychological well-being. Some scholars [2] emphasize the critical importance of early childhood as a sensitive period for development. They identify the age range between 2 and 5 years as a crucial developmental window, during which speech, autonomy, self-awareness, and cognitive, social, play, and motor skills rapidly emerge and evolve. The first two years of life are also very important: during this time, the human brain triples in size and forms an extensive network of neural circuits necessary for processing sensory input and developing foundational cognitive functions [3]. Language development, for example, begins early in life and reflects a complex interplay between genetic predispositions and environmental influences [4]. The quality of early interactions, especially with primary caregivers, is also crucial. Positive early experiences are foundational for the development of secure social relationships later in life [5,6]. Moreover, the early formation of core executive functions (such as attention, working memory, and self-control) is another very important aspect of neurodevelopment in early childhood [7]. Research shows that children who are subject to high-quality early care and education have better cognitive, social–emotional, and academic outcomes [1,8]. On the other hand, stimulus-poor environments, social deprivation and neglect can have long-lasting adverse effects on brain development.

Cognitive and socio-emotional development in early childhood is a key predictor of later academic achievement, psychological well-being, and social functioning. One of the earliest and most basic developmental milestones in infancy is the ability to recognize and respond to emotional cues, which supports the formation of secure attachments with the parent/caregiver [9]. Brain development during this critical period lays the foundation for future learning and emotional regulation, making it essential to provide a safe, nurturing, and stimulating environment that supports optimal neural growth. Research also shows that children who develop strong emotional regulation and interpersonal skills are better equipped to cope with challenges, establish meaningful relationships, and make positive contributions to society [10]. Children who experience supportive relationship with their parents, teachers and peers are more likely to develop resilience and emotional stability. Conversely, adverse childhood experiences, such as neglect or social exclusion, can impede emotional growth and lead to long-term psychological challenges. Identifying predictors of developmental delays is crucial for enabling timely interventions that could improve long-term outcomes [10].

A study by McCoy et al. [11] identifies three primary risk factors: malnutrition, low parental education, and exposure to psychosocial stressors. Nutritional deficiencies, in particular, are strongly linked to cognitive delays. Evidence from low- and middle-income countries indicates that providing nutritional supplements during pregnancy and early childhood positively influences both physical growth and neurodevelopment [12]. A longitudinal cohort study by Wang et al. [13] based on data for 1245 children from rural China inspects their developmental changes over three time periods. Prospective analysis reveals that those who had experienced persistent developmental delays during the early years show significantly lower cognitive and socio-emotional skills in preschool. Similarly, those whose development had deteriorated before the age of three demonstrate substantial deficits later on. The findings in [13] highlight the role of socio-economic factors, such as household income, parental education, and access to early learning opportunities, in determining whether a child overcomes or remains affected by delays. They reveal that early cognitive and emotional deficits have a strong impact on later preschool performance and highlight the urgent need for targeted early interventions. Children from lower-income households are disproportionately affected by ongoing developmental delays, hence the importance of providing access to early support for marginalized communities.

### *1.2. Interventions*

Screening programs that identify at-risk children before the age of three, combined with targeted educational initiatives, can significantly mitigate long-term developmental challenges, as argued in [13]. These findings highlight the critical importance of integrating structured play, emotional intelligence training, and early literacy programs into preschool curricula. The authors also advocate for policy reforms aimed at enhancing early childhood screening, expanding parental education initiatives, and increasing access to preschool interventions—particularly in underserved communities. Future research should explore the effectiveness of culturally tailored interventions and track long-term developmental outcomes beyond the preschool years. Continuously improving and updating early intervention strategies is essential for reducing disparities and fostering the cognitive and emotional development of vulnerable populations [13]. The article highlights the importance of nurturing an environment that supports emotional intelligence, communication skills, and social adaptability [10], while [14] argues that fostering adaptive mechanisms in preschoolers could occur through participation in a variety of interactive games with peers, developing healthy bonds and attachment relationships with adults, active parental social networks, and participation in early childhood education programs. Various early childhood programs report that the benefits for 4-year-olds at the end of the year are focused on vocabulary development, social skills, and a positive approach to learning [8]; on the other hand, benefits were found in all areas of development for 3-year-olds. Some studies, e.g., [15], point out that the effects of program impacts on participants have either increased or remained stable over the years. Another key point of early childhood development programs [16] has to do with fostering community-wide participation and collaboration: building networks of support and engaging locals gives these initiatives a new perspective.

In Bulgaria, where our study takes place, a good example of holistic support for children (from birth to 18 years of age) and families is the Integrated Development Model (IDM) of the Health and Social Development Foundation (HESED). One of its key programs provides incentives and educational support aimed at promoting the well-being of young children aged 3 to 5, with a specific focus on reducing the school readiness gap between minority children and their peers. In recent years, Bulgaria has been making efforts to improve the living conditions and development of Roma children. With the National Strategy for Roma Integration 2012–2020, it has taken important steps to create a national

framework for early childhood development as an integrated approach to support parents and children, thus ensuring their access to quality education, integrated services, health and social care. Despite these efforts, the Roma community continues to face disproportionately high rates of infant mortality, early pregnancy, illiteracy, and unemployment [17,18]. Roma children are significantly more likely to experience malnutrition, lack of adequate healthcare, poverty, and social exclusion. Many do not attend school regularly, which puts them at a substantial disadvantage from an early age. Thus, targeted early education within this particular minority group in Bulgaria is essential for both moral and economic reasons.

### *1.3. Machine Learning for Effective Education*

With the advancement of large language models, like ChatGPT, the use of AI and machine learning applications in education has emerged as a subject of considerable academic and public debate [19–21]. Nevertheless, we do not see the overwhelming abundance of research papers that one might expect. There are numerous ways in which this relatively new technology can be implemented in the curriculum: from statistical analysis of test scores [22,23] to life interactions with learning bots [24,25]. The former is closer to the scope of this paper, particularly regarding applications in early education. We apply methods similar to those used in [26] to predict learning difficulties of children under five years of age, based on parenting strategies and sociodemographic and health conditions. It is also related to studies like [27,28], which access the impact of various factors, such as individual effort, school policy and family background, on students' academic performance and mental condition. There are similar studies in the context of higher education, e.g., [29]; some authors also apply clustering algorithms in order to enhance differential learning strategies [30] and various ML/AI predictive algorithms aimed at optimal outcomes [31,32]. Various aspects of our study have been applied in different research papers before, but not in this particular context. Also note that we work with a smaller and highly specific dataset, so the models cannot be generalized directly; however, once proven effective, they could be built similarly in different settings.

This paper is organized as follows: In Section 2, we comment on the data, statistical tools and machine learning algorithms used in our study. The main findings are then revealed and discussed, with regard to their practical applications and prospects for future research.

## 2. Materials and Methods

### *2.1. Data*

The data in our surveys were collected by the the Health and Social Development Foundation (HESED) within a standard institutional procedure during the years 2014–2018 with the explicit consent of all parties involved and with sensitivity to the Roma cultural context. HESED staff included trained Roma community mediators, respecting community norms and expectations. There are plenty of missing data in the files for various reasons: the respondents refused to answer some of the questions, the children dropped out of the program before final assessments could take place, etc. In some parts of the study, we also use background information about their home environments, such as marital status, education, occupation and income brackets of the parents. Unfortunately, however, there are both absent and misleading data provided by the respondents, especially regarding education, occupation, and household income. Also, in many cases, it is not clear which parent answered the questionnaire. Therefore, we extensively use cross-validation in our estimates.

Our main source of data consists of the much more reliable test results upon entrance and exit, as well as the individual assessment of children's progress given by parents and

teachers and experts. The first noticeable tendency in the latter is, of course, the biased judgment of each party: parents, in particular, tend to overestimate their underperforming children and underestimate the more advanced ones, in both cognitive and social–emotional development, as shown in Figure 1. The shaded areas in the above diagrams correspond to the confidence interval for the regression parameters. These discrepancies tend to converge as the program progresses but still remain visible. The tendency is that teachers provide less biased and dispersed answers compared to parents; however, for both groups, the slope is consistently lower than the one suggested by the experts' opinion, which we use as a benchmark here.

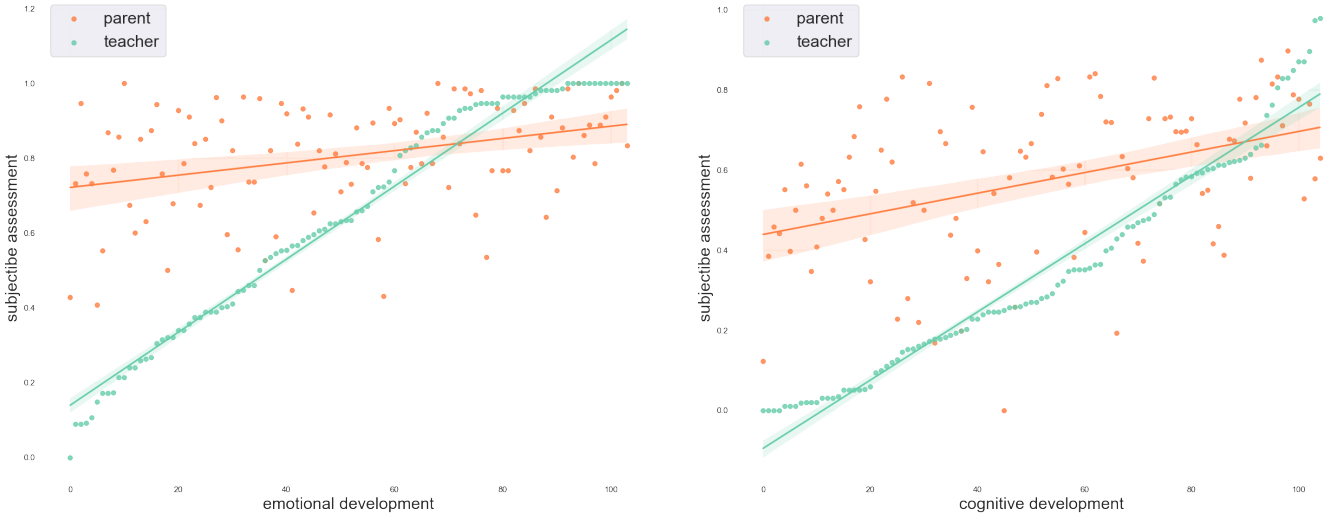



**Figure 1.** Subjective evaluation of children's social–emotional (**left**) and cognitive development (**right**).

We perform a paired $t$-test to assess the $p$-value for the children's progress, according to one or both of their parents, a teacher and a specialist. The results of the survey were averaged for each feature we are interested in—the mean and median progress were assessed, as well as the standard deviation, displayed in Table 1. The 5% $p$-value test was passed in all cases, while the $p < 0.01$ threshold was not, for social–emotional development.

**Table 1.** Subjective assessment of the children's relative progress and as measured by the test results.

| Category | Parent | Teacher | Expert | Tests |
|---|---|---|---|---|
| adaptation | 9.75% | 20.70% | – | 18.79% |
| communication | 7.16% | 28.79% | 29.85% | 25.63% |
| cognitive | 11.96% | 46.14% | 25.53% | 32.23% |
| physical | 8.60% | 31.65% | 21.58% | 14.36% |
| emotional | 3.79% | 18.33% | 6.65% | 15.19% |

It is interesting to see that verbal and cognitive skills are developed the most according to both the assessment tests and the survey, while the changes in social–emotional and physical development appear less statistically significant. However, half of the parents view their children's verbal skills as negatively influenced during the education program and report negligible average progress across all areas. Teachers tend to underestimate the process even more: in their view, the only statistically significant change is in cognitive abilities, but they see it as being minute and even negative for half of the children. As expected, both parents and teachers are much less homogeneous in their responses compared to experts, as the ratio between the standard deviation and the mean clearly indicates. They

seem to overestimate worse performers while underestimating the better ones in basically all categories. This bias could explain, at least partially, the distrust and negative view of the process many parents seem to share.

In summary, all parties involved agree that there is no significant change in social–emotional development for the whole sample, while there is certainly one for cognition. Parents and experts, on the other hand, see improvement in physical skills, which teachers do not. Unlike parents, teachers and experts agree that children develop communication skills during the program. These biases might be confusing for the objective evaluation of the overall effect of the education process on children, but they provide useful information for our machine learning algorithms. We shall come back to that point later on.

The distribution of the test results for the five skills of interest is quite different: those for cognitive, physical, and communication abilities have some common features, e.g., the lack of significant progress for the top and bottom performers, unlike for those in the middle (see Figure 2). This yields a wider spread on the $x$-axis with peaks at both the top and the bottom, and a much narrower distribution upon exit, due to the significant progress (reaching over 32% on average for cognitive skills). Similarly, adaptive behavior and emotional development are almost normally distributed on both axes and do not show significant improvement during the education process. In order to better assess the interactions between features, we visualize the correlation matrix using heatmap (Figure 3).

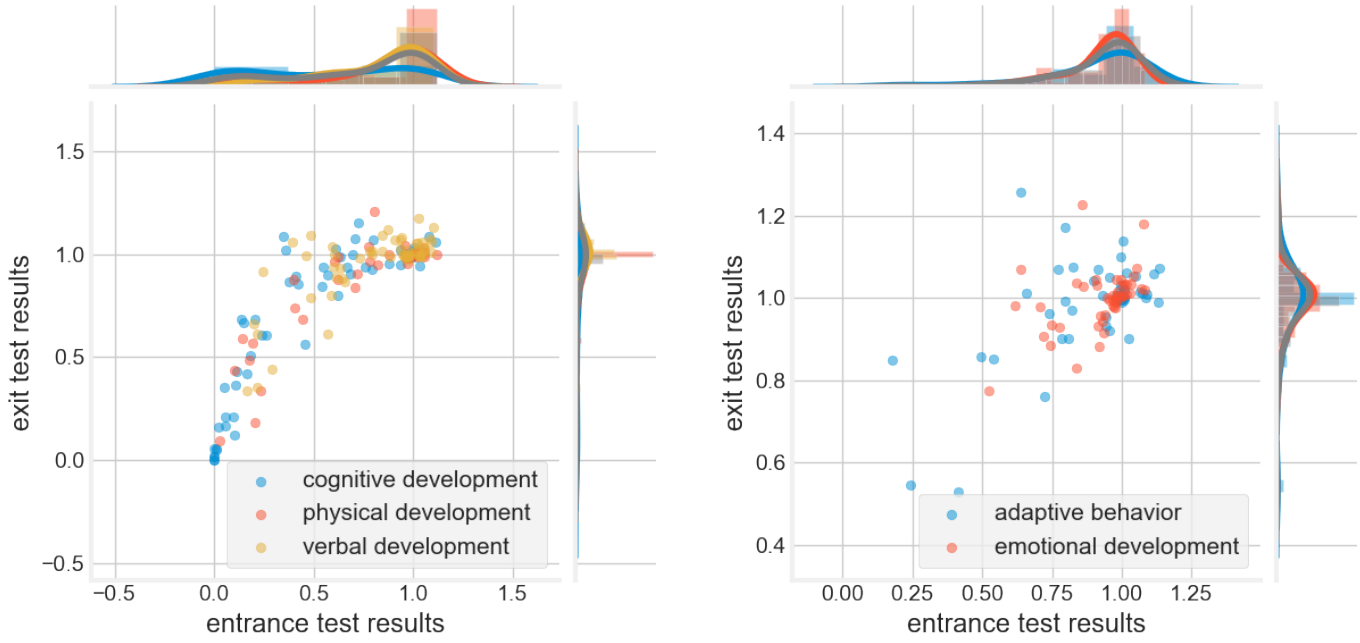



**Figure 2.** Distribution of test scores: cognitive, communication and physical (**left**); social–emotional and adaptive development (**right**).

### 2.2. Methods

We use a large variety of methods, such as descriptive statistics in order to explore the data, inference statistics (e.g., analysis of variance and the paired $t$-test) to test hypotheses and assess the statistical significance of different effects, and machine learning algorithms to generate predictions; thus, we recommend solutions for efficiency improvement. Here, we shall discuss only the latter, since statistical analysis is a standard component on any survey-based research. The first problem it allows us to deal with is restoring damaged or missing data, such as the mother's age, which is absent in over 17% of the questionnaires. For that, we use a simple KNN ($k$ nearest neighbors) imputer based on proximity. For this particular dataset and the tasks we focus on, more advanced techniques such as

iterative imputation (MICE) yield practically identical results as far as the evaluation metrics are concerned. Furthermore, after evaluating the covariance matrix, we apply principle component analysis (PCA) on the test results in order to obtain a better understanding of the internal feature dynamics, which is also carried out using SHAP analysis, and we aim for optimal variance representation that might be useful. Finally, we perform several k-means clustering and test several classification algorithms, such as kernel support vector machines (K-SMV), linear discriminant analysis (LDA), Gaussian Naive Bayes (GNB) and extreme gradient boosting (XGBoost) for imputing missing categorical data and predicting future outcome. Their performance is evaluated using a standard $F_1$-score and 'area under the curve' (AUC):

$$F_1 = \frac{2}{P^{-1} + R^{-1}} = \frac{2T_+}{2T_+ + F_+ + F_-}, \qquad P = \frac{T_+}{T_+ + F_+}, \qquad R = \frac{T_+}{T_+ + F_-} \tag{1}$$

where $P$ and $R$ stand for the so-called 'precision' and 'recall' metrics, while $T_\pm$ and $F_\pm$ denote, respectively, the true/false values—positive/negative predictions of a binary classifier. In the multi-class setting, one uses the micro or macro $F_1$-score instead. The former is calculated using the net $T_\pm$ and $F_\pm$ in Formula (1); for the latter, we simply take the average of class-wise $F_1$-scores. It also has a weighted version where the weights are proportional to the number of samples in each class. As for the AUC metric, we consider the so-called 'receiver-operating characteristic curve' (ROC), which depicts the dependence between the false positive and true positive rates at any possible threshold between 0 and 1. The area beneath its graph is a standard metric for binary classifiers' performance, and like $F_1$, it can be generalized to the multi-class setting via micro, macro, or weighted averaging.

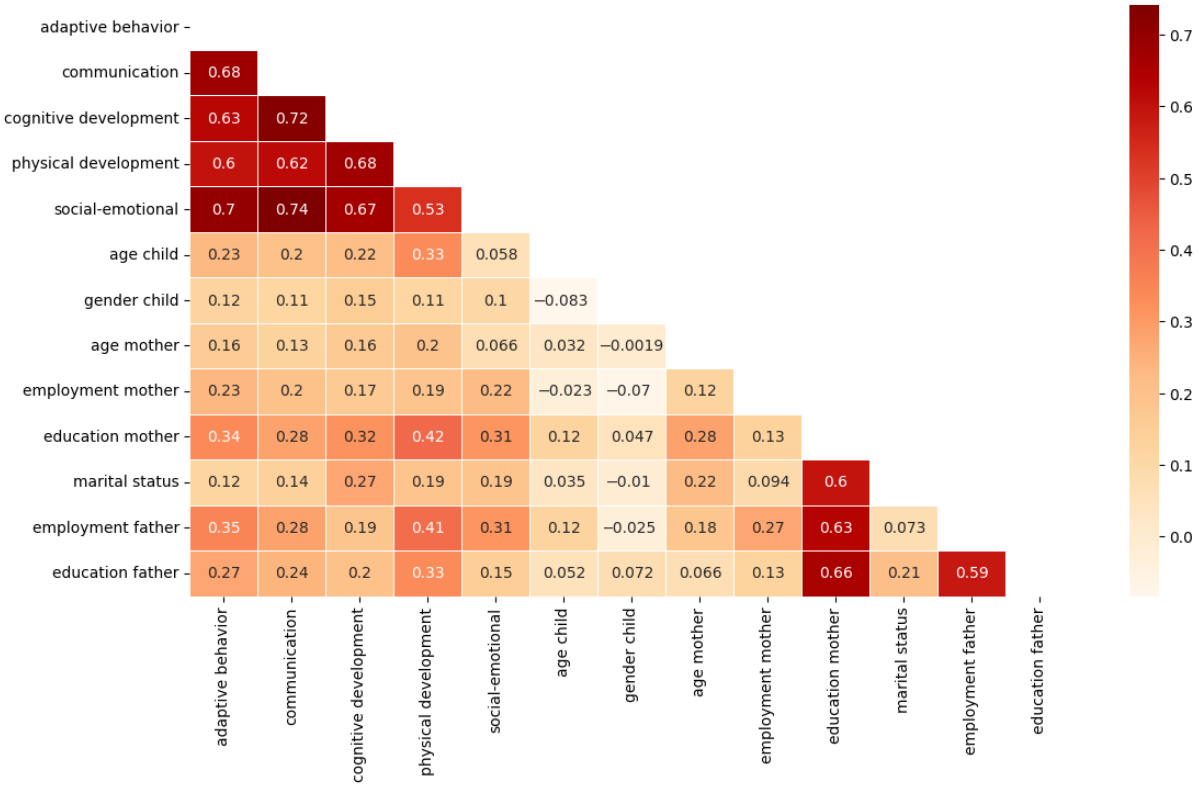



**Figure 3.** Heatmap correlation table of the presented features.

Finally, we attempt to predict values of continuous parameters, namely children's scores difference upon exit and entrance, based on weighted ensemble stack models. For this more ambitious task, we rely on AutoGluon, Amazon's open source Python module,

equipped with great functionality for advanced machine learning. It trains several models independently and ranks them according to a given metric, e.g., the coefficient of determination $R^2$ or the mean absolute percentage error MAPE, given by the following expressions:

$$R^2 = 1 - \frac{\langle \Delta X^2 \rangle}{\langle X^2 \rangle}, \qquad \text{MAPE} = \langle \frac{|\Delta X|}{|X|} \rangle \tag{2}$$

where $\langle X \rangle$ denotes the mean value of the distribution $X$, respectively, and $\langle \Delta X \rangle$ stands for the average deviation of the prediction from the actual value of $X$. Similarly, $\langle \Delta X^2 \rangle$ is the average square error and $\langle |\Delta X| \rangle$ is the average absolute error of our prediction. Also note that these quantities are assessed via cross-validation (CV), i.e., the training and the test sets are determined multiple times and the corresponding estimates are then averaged in order to avoid misleading evaluations due to pure luck; thus, CV values are usually lower.

In order to compensate for the small dataset, we need to wisely use the capabilities of machine learning algorithms. AutoML tools [33] provide a great opportunity in that respect: they test various models, assess them, and tune their hyperparameters without human interference. Then, the models are ranked according to several performance criteria, and the best ones are combined into a weighted sum, which tends to reduce the overall error. This process, known as 'stacking' [34], can be executed repeatedly, which we refer to as 'multi-level stacking'. Figure 4 shows a simplified flow chart of a single level.

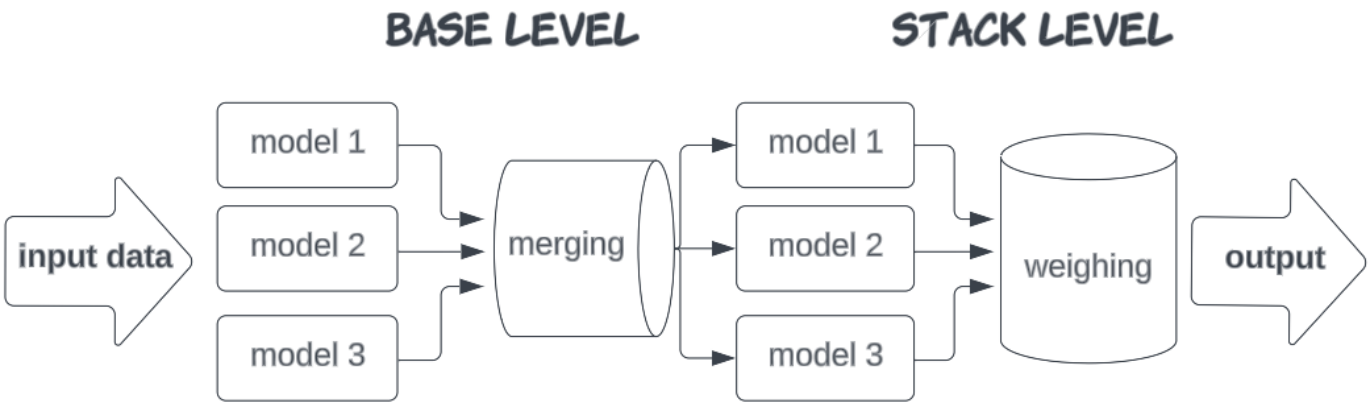



**Figure 4.** Merging and stacking of multiple models in AutoML.

In the case of classification, some kernel models perform quite well, as shown on Table 2. For regression problems, our 'top learners' are neural networks (NeuralNetTorch, NeuralNetFastAI) and ensemble algorithms based on decision trees. They promote the concept of collective decisions, showing, in practice, that a large number of rough biased models (weak learners) often beat the accuracy of highly sophisticated algorithms. The most common example, Random Forest, uses an ensemble of decision trees trained in parallel with biased sub-samples, replicated via bootstrapping (sampling with replacement). Then, aggregation yields the output by either simple majority vote, or a weighted sum. This two-step solution is known as bagging. It is powerful and easy to handle (no need for pre-processing or taking care of multicollinearity), suitable for parallel processing and remarkably resistant to overfitting. However, it might need lots of resources, especially when it comes to hyperparameter tuning and optimization. Boosting algorithms provide significant improvement in that respect, relying on iterative corrections of statistical weights, assigned to each data instance in the sample, thus giving the model a chance to focus on its flaws in the next iteration, which saves lots of resources. That is how adaptive boosting (AdaBoost) works, but there are other variations as well, e.g., gradient boosting algorithms, such as XGBoost, which uses gradient descent to optimize the weights at each iteration,

similarly to neural networks. This makes it quite fast and successful in balancing the so-called 'bias–variance trade-off'. Just like bagging algorithms, it does not require data normalization and works well even with missing values, but seems to be more sensitive to outliers, so one needs to pay attention. Still, XGBoost is one of the best-known solutions used for both classification and regression machine learning problems. CatBoost and LightGBM are two other well-known examples of gradient boosting algorithms with many applications.

**Table 2.** Subjective assessment of the children's progress measured by the *p*-value.

| Feature | $F_1$ Score | AUC | Best Model |
|---|---|---|---|
| employment father | 0.92 | 0.99 | Random Forest |
| education mother | 0.96 | 0.99 | Gaussian NB |
| marital status | 0.97 | 1.00 | LDA |
| cluster type | 0.86 | 0.94 | XGboost/K-SMV |

## 3. Results

After data imputation and normalization, we attempt to find a regression model for some of our features. As the above heatmap indicates, cognitive development appears to be highly correlated with verbal and physical skills among other personal characteristics. As for the influence of domestic environments, the parents' education (especially that of the father) dominates all other factors, followed by their marital status and occupation. The family income is not significant in that respect, but it plays a role in the development of verbal skills, along with the father's education. The mother's education appears highly correlated with adaptive behavior, while that of the father with social–emotional development; however, in both areas, the marital status plays an equally important role. This becomes apparent in the feature importance tables in regression models, too. For example, cognitive and physical development entrance scores can be 'guessed' using standard ensemble learning algorithms (XGBoost and Random Forest), with accuracy reaching, respectively, 86% and 95%. Another highly correlated feature is the infant's age. For this factor, the accuracy exceeds 91% with hyperparameter tuning. It seems more likely to predict children's weight and height at birth (with accuracy about 88% and 93%, respectively) than household income (with error almost 30%), which may be due to intentional misleading answers in the survey and a large amount of missing data: less than 68% actually responded to that question.

### *3.1. Classification Results*

As for the classification problem, using different algorithms, we obtain excellent predictions for the education, marital or employment status of the parents (see Figure 5). Both the accuracy and the $F_1$-score are about 0.9, particularly for the Random Forest and Naive Bayes classifiers, while it is quite difficult to infer the child's gender from the available data. Also, we perform cluster analysis on the data, this time focusing on the differences between entrance and exit level skill assessments. In particular, we are interested in the possibility to identify children who are likely to demonstrate a significant improvement in a certain skill during the course. To do that, we study the differences in the initial and final assessment of all five features and try to find a pattern. For example, if we proceed as such regarding the progress of cognitive and social–emotional progress, we can see a nice separation into three clusters, as illustrated in Figure 6. The bottom left segment shows no particular improvement in any dimension, while the top one corresponds to a balanced improvement, weighing on the social–emotional side. Finally, children in the right cluster demonstrate significant progress on the cognitive front and almost none on the emotional one (some of them even regress). Assigning children to the proper cluster is a different

classification problem, as it attempts to predict future events, unlike the previous examples, which were suited for data imputation. The result is therefore expectedly less accurate (Table 2), but it is much more valuable when it comes to strategies for differential learning as a means of optimization.

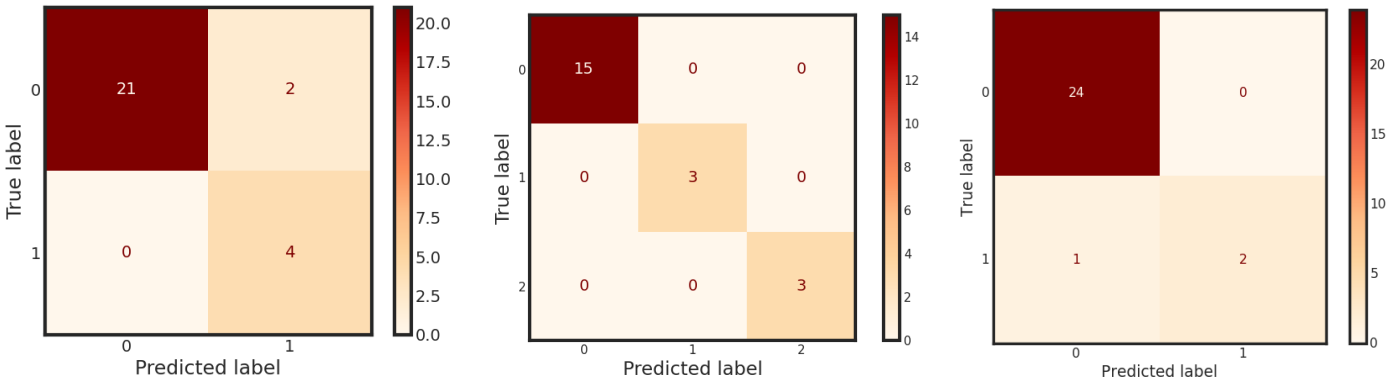



**Figure 5.** Classification for the father's employment status (**left**), as well as the mother's education (**center**) and marital status (**right**).

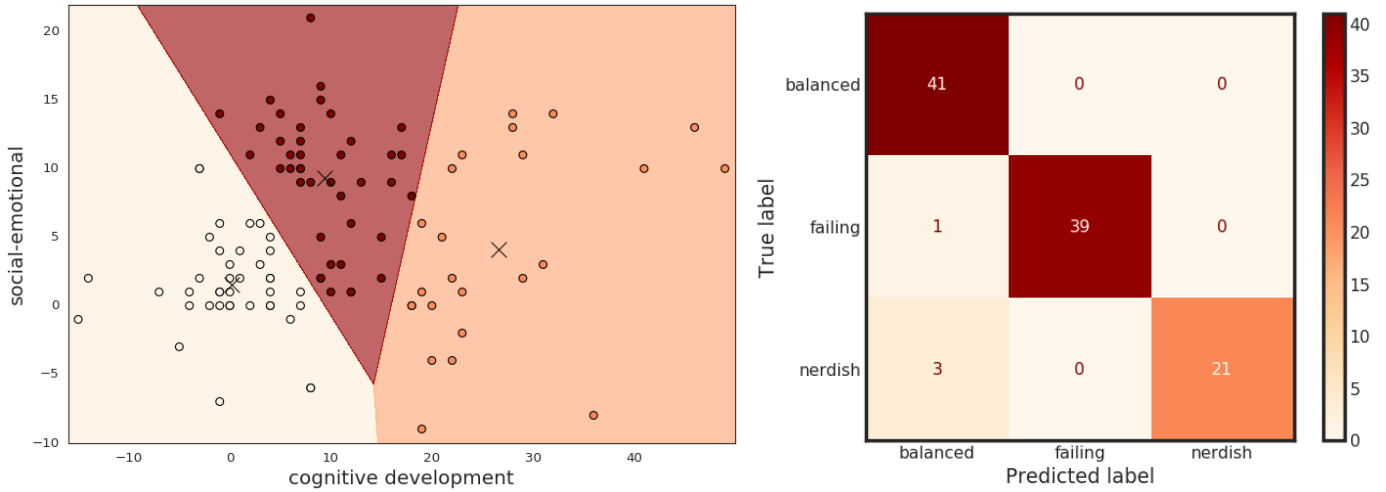



**Figure 6.** K-means clusters for cognitive vs. emotional development throughout the program (**left**) and the confusion matrix for prediction of cluster classes, based on entrance-level assessment (**right**). Note that the extreme clusters were flawlessly predicted.

### *3.2. Predictions Form Weighted Ensemble Regression*

After exploring the possibilities of binary and multi-class predictions for different categorical features of our data, let us approach an even more ambitious task: modeling continuous values via regression. In particular, it would be useful to have predictions for the expected progress of each child within the program, with respect to the five areas of interest. The measure of progress is given by the difference between exit and entrance test scores and can take positive as well as negative values. Although not so helpful for the classification problem, the biased assessment of teachers and parents at admission has proven quite valuable for the regression task (as it reflects hidden background dynamics), which is much harder considering the small amount of data in our survey. Cognitive and physical development seem unlikely to predict, with a coefficient of determination of $R^2 < 0.4$; on the other hand, progress in communication and social–emotional development is much more easily 'tamed', with $R^2 \approx 0.8$ using deep learning with NeuralNetTorch (Figure 7). It is interesting to note that with regard to feature importance, the father's education seems

to be a good predictor for the former and the mother's age for the latter. Similarly, for adaptive behavior, level 3 stacking yields $R^2 \approx 0.7$, while for cognitive progress, significant predictors appear to be the parents' education and the age of both the mother and the child. However, the data are too scarce to establish a strong enough correlation. The results are summarized in Table 3.

Next, we look at the dimensionality reduction, or rather the orthoganlization problem, using principal component analysis (PCA). It diagonalizes the covariance matrix of the individual features in its canonical basis and rank-orders the eigenvalues, allowing us to select only the leading ones for simplicity. This technique has been used, for instance, to construct the famous 'Big five' among all personality traits: openness, conscientiousness, extroversion, agreeableness, and neuroticism. In our case, three components are quite sufficient to 'explain' about 85% of the variance, as Figure 8 shows. With two components, we only cover about 64% which is insufficient for a high-precision study. It is easy to see that one of the components is related mostly to social–emotional and verbal skills, while the other two are related to cognitive and physical development, respectively (see Figure 8).

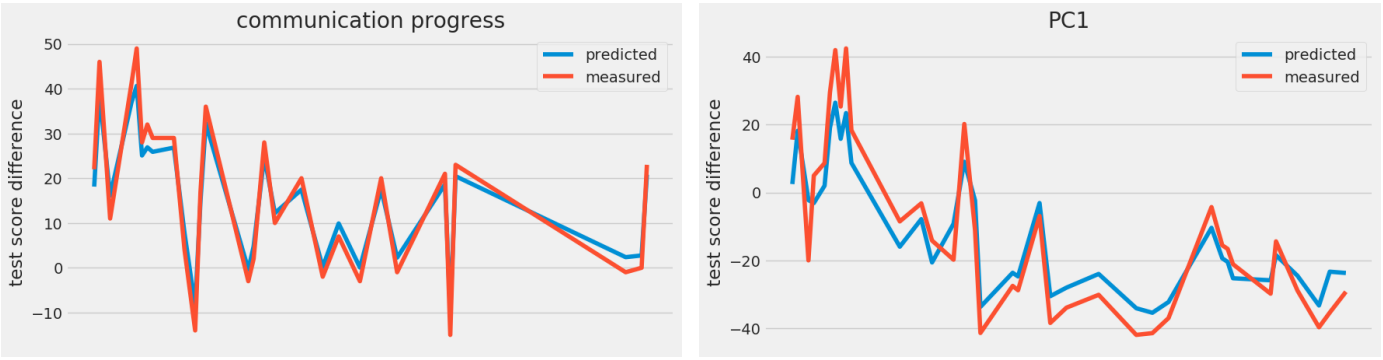



**Figure 7.** Validation of the regression models for communication (**left**) and PC1 'soft skills' progress (**right**).

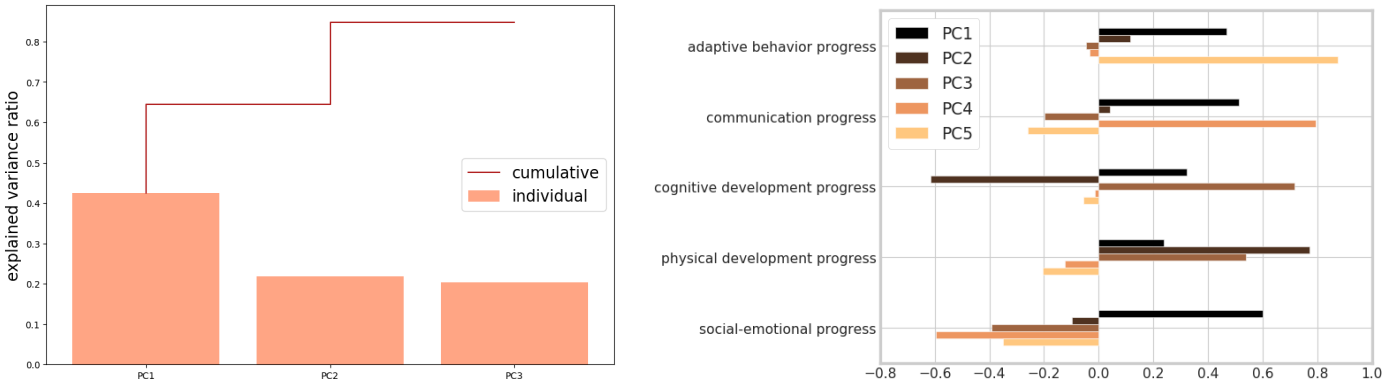



**Figure 8.** PCA cumulative explained variance for test results (**left**) and the contributing factors for the first three components (**right**).

However, one should note that PCA is merely a coordinate transformation. It presents a trade-off, rather than a cure, gaining better separation (e.g., for graphs and diagrams) at the expense of explainability (some of the principal components may be hard to associate with human words). Also, contrary to what we expected, our classification algorithms did not improve significantly when assisted by PCA (or the standard algorithms for imbalanced data). Regression results, on the other hand, improve in some areas and regress in others. Cognitive skills are better modeled as the third component in PCA coordinates, while PC1, associated mostly with 'soft' skills (see Figure 8), is less predictable

than pure social–emotional development, probably due to the 'smearing' that takes place. On the other hand, PC2 is just as hard to model as the original feature 'physical skills', to which it is directly linked. Even with PCA, however, predicting cognitive progress remains difficult—probably because it depends strongly on the learning process and environment, and not nearly as much on the initial circumstances.

Note that both the PCA technique and multi-level stacking may be used for the classification task as well. While the former does not contribute much, as mentioned before, we see a significant improvement with the latter. Ensemble learning and neural networks with precise hyperparameter tuning allow us to determine the correct cluster for each child in the validation set, as shown in Figure 6 (the confusion matrix becomes diagonal, so disregard this picture in this instance). This compensates, at least partially, for our inability to accurately predict the cognitive progress of individual children: After all, preschool programs work with groups and proper separation is essential for their efficiency. Being able to predict individual scores is of less value in this context, while it can be a major focus when it comes to admission to high school or university.

Another interesting aspect of this study is the evaluation of the relevance different features have on the predicted label (see Table 3). For example, parents' education seems to be a strong predictor for communication, physical and cognitive development. The mother's age also has an impact on these features, as well as on the 'soft' skills (PC1, social–emotional development and adaptation), along with that of the child. Gender also plays a minor role in social–emotional development, but parents' features clearly dominate overall: Figure 9 illustrates that. As for cluster classification, adaptive behavior is the most crucial predictor.

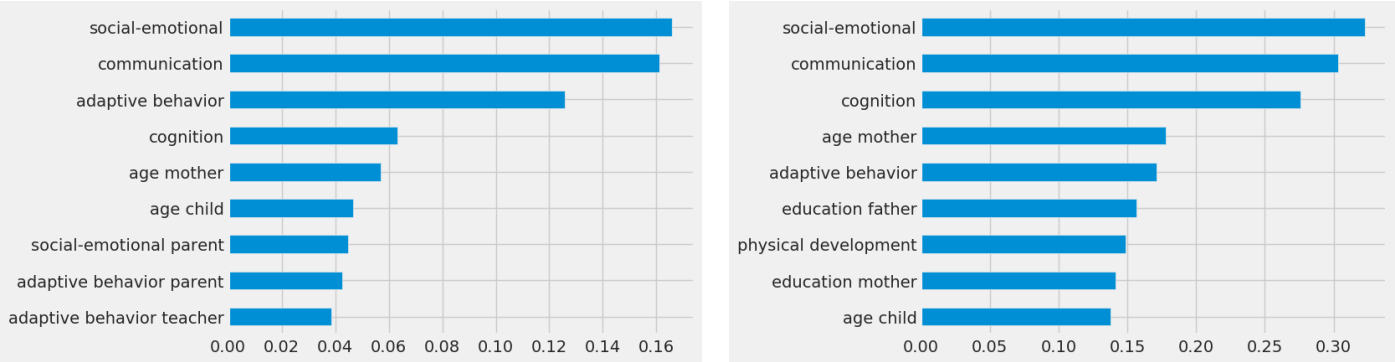



**Figure 9.** Feature importance charts for the PC1 'soft skills' (**left**) and PC3 'cognitive development' (**right**) regression models.

**Table 3.** Performance of our regression models on various targets.

| Target Feature | $R^2$ | Best Model | Background Features |
|---|---|---|---|
| social–emotional | 0.77 | NeuralNetTorch | age mother/child |
| communication | 0.75 | NeuralNetTorch | education/age |
| adaptive behavior | 0.68 | WE_L2 | age mother/child |
| PC1 ('soft' skills) | 0.68 | WE_L3 | age mother/child |
| PC2 (physical) | 0.40 | WE_L2 | education/age |
| PC3 (cognitive) | 0.47 | WE_L2 | education father |

## 4. Discussion

The results obtained in this study provide valuable insights into the developmental progress of Roma children enrolled in an early educational program and demonstrate

how machine learning tools can enhance our understanding and predictive capability of individual learning trajectories. Overall, the findings confirm the significant potential of structured early interventions in improving cognitive, communication, and, to some extent, social–emotional outcomes for children of disadvantaged backgrounds.

### *4.1. Linking Outcomes with Theory*

From a developmental perspective, the observed improvements align with well-established theoretical frameworks emphasizing early childhood as a critical period for brain development and plasticity [2,9]. The highest gains—in cognitive skills 32% and communication 26%—are consistent with the literature, suggesting that early experiences shape neurocognitive architecture, particularly in areas related to language, attention, and executive function [4,35]. While physical and socio-emotional development showed more modest improvements, this is not surprising, given that these domains are more sensitive to relational quality and environmental stability, which may not shift as quickly or uniformly during programmatic interventions. Furthermore, the convergence between teacher and parent evaluations post-intervention suggests increasing reliability in subjective assessments as the program progresses. This convergence may reflect both genuine developmental changes and improved adult–child interactions fostered by the program. However, discrepancies, particularly parental under- or over-estimations of children's progress, highlight the continuing need for objective measurement tools and caregiver support to develop accurate expectations and developmental literacy.

### *4.2. The Role of Socio-Economic Factors*

Our findings also support previous research on the role of socio-economic factors in early development [11,13]. Parental education emerges as a key predictor of both cognitive and social–emotional outcomes, while marital status also plays a significant role in emotional development and adaptive behavior. This reinforces the argument that early interventions must be contextualized within broader eco-systems, recognizing the influence of family structure, educational background, and home stability on child development. In particular, the father's education is linked to the cognitive and verbal outcomes of children. This may reflect cultural patterns in Roma families, where paternal influence over schooling could be more pronounced, and it is consistent with the findings of other studies [36]. Interestingly, income was not a strong predictor, likely due to under-reporting or unreliability in self-reported data—an important limitation for surveys in marginalized populations. Nevertheless, its role in verbal development suggests that financial resources may still impact specific learning environments, e.g., access to books, media, structured play.

### *4.3. Implications and Future Directions*

These findings underscore the need for integrated early childhood strategies that combine developmental screening, parental education, and culturally sensitive pedagogues. Programs like HESED's Integrated Development Model offer promising frameworks for replicable interventions tailored to Roma communities and other marginalized groups. Beyond program design, this study also emphasizes the value of interdisciplinary approaches, bridging developmental psychology and data science, for scalable and evidence-based educational reform. Future research should expand on our findings with larger and more diverse samples, and evaluate long-term developmental outcomes in primary school and adolescence. This would align with long-term studies such as [11,13]. Moreover, incorporating more granular measures of home learning environments and caregiver–child interaction quality could further enhance predictive modeling. Despite limitations in sample size, our

findings demonstrate that small-scale studies, when carefully designed and validated, are capable of providing meaningful insights and informing scalable educational interventions.

### *4.4. Promoting Interdisciplinary Research*

The present paper deals with a very specific problem in developmental psychology; however, the approach we use to deal with it can be extended to different areas of the field, as well as to multidisciplinary research. That is, we show that combining simple clustering and dimensionality reduction techniques with powerful multilayer stacking algorithms for regression and classification allows us to efficiently use scarce resources at our disposal. In particular, we managed to achieve an $F_1$-score of $\approx$0.9 and an AUC of $\approx$1 for classification, respectively, coefficient of determination of $R^2 \approx 0.8$ for regression tasks with only about a hundred data instances. What is more important, we effectively predict the outcome for a child enrolled in the program with reasonable accuracy, which allows us to implement various differential learning strategies in order to optimize the process by allocating resources more efficiently and thus boost the net benefit. This small-scale study can also be used as a base for future research involving data at the national or regional level and aiming for higher precision. However, studies with small datasets are also quite useful as they demonstrate how much can be completed with so little—if we use resources wisely.

**Author Contributions:** Conceptualization, D.B. and N.K.; methodology, D.B. and N.K.; software, D.B.; validation, D.B., N.K. and D.S.; formal analysis, D.B. and D.S.; investigation, N.K. and D.S.; resources, N.K.; data curation, N.K. and D.S.; writing—original draft preparation, D.B.; writing—review and editing, D.B., N.K. and D.S.; visualization, D.B.; supervision, N.K.; project administration and funding acquisition. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Department of Cognitive Science and Psychology, New Bulgarian University, N 281/21.05.2020.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in this study.

**Data Availability Statement:** The data used in this study are not publicly available.

**Acknowledgments:** The authors are grateful to the HESED Foundation for providing the raw data.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| PCA | Principal component analysis (dimensionality reduction algorithm) |
| LDA | Linear discriminant analysis (dimensionality reduction and classification) |
| KNN | *k*-nearest neighbors (here used for classification and data imputation) |
| NB | Naive Bayes (classification algorithm) |
| XGBoost | Extreme gradient boosting (classification and regression algorithm) |
| CV | Cross-validation (for the evaluation metric) |
| WE_L2/L3 | Weighted ensemble, level two/three (stacking of trained models) |
| K-SVM | Kernel support vector machine (classification algorithm) |
| ROC | Receiver operating characteristic |
| AUC | Area under the curve |
| MICE | Multiple imputation by chained equations (data imputation algorithm) |

## References


1. Schweinhart, J.L. The High/Scope Perry Preschool Study Through Age 40. Summary, Conclusions, and Frequently Asked Questions. High/Scope Educational Research Foundation. 2005. Available online: https://highscope.org/wp-content/uploads/2024/07/perry-preschool-summary-40.pdf (accessed on 30 May 2025).
2. Grantham-McGregor, S.M.; Cheung, Y.B.; Cueto, S.; Glewwe, P.; Richter, L.; Strupp, B.; The International Child Development Steering Group. Child development in developing countries 1: Developmental potential in the first 5 years for children in developing countries. *Lancet* **2007**, *369*, 60–70. [CrossRef] [PubMed]
3. Giedd, J.N. Structural Magnetic Resonance Imaging of the Adolescent Brain. *Ann. N. Y. Acad. Sci.* **2004**, *1021*, 77–85. [CrossRef]
4. Bishop, D.V.M.; Snowling, M.J. Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychol. Bull.* **2004**, *130*, 858–886. [CrossRef]
5. Thompson, R.A. Early attachment and later development. In *Handbook of Attachment: Theory, Research, and Clinical Applications*; Cassidy, J., Shaver, P.R., Eds.; Guilford: New York, NY, USA, 1999; pp. 265–286.
6. Liu, D.; Diorio, J.; Tannenbaum, B.; Caldji, C.; Francis, D.; Freedman, A.; Sharma, S.; Pearson, D.; Plotsky, P.M.; Meaney, M.J. Maternal care, hippocampal glucocorticoid receptors, and hypothalamic–pituitary– adrenal responses to stress. *Science* **1997**, *277*, 1659–1662. [CrossRef]
7. Diamond, A. Executive Functions. *Annu. Rev. Psychol.* **2013**, *64*, 135–168. [CrossRef]
8. Head Start Impact Study. Final Report. U.S. Department of Health and Human Services. Administration for Children and Families. 2010. Available online: https://acf.gov/sites/default/files/documents/opre/hs_impact_study_final.pdf (accessed on 30 May 2025).
9. Marzola, P.; Melzer, T.; Pavesi, E.; Gil-Mohapel, J.; Brocardo, P.S. Exploring the Role of Neuroplasticity in Development, Aging, and Neurodegeneration. *Brain Sci.* **2023**, *13*, 1610. [CrossRef]
10. Hanish, L.D. The grand challenge in social and emotional development: Building a harmonious world for children and adolescents. *Front. Dev. Psychol.* **2024**, *1*, 1130880. [CrossRef]
11. McCoy, D.C.; Peet, E.D.; Ezzati, M.; Danaei, G.; Black, M.M.; Sudfeld, C.R.; Fawzi, W.; Fink, G. Early Childhood Developmental Status in Low- and Middle-Income Countries: National, Regional, and Global Prevalence Estimates Using Predictive Modeling. *PLoS Med.* **2016**, *13*, e1002034. [CrossRef]
12. Prado, E.; Larson, L.; Cox, K.; Bettencourt, K.; Kubes, J.; Shankar, A. Do effects of early life interventions on linear growth correspond to effects on neurobehavioural development? A systematic review and meta-analysis. *Lancet Glob. Health* **2019**, *7*, e1398–e1413. [CrossRef] [PubMed]
13. Wang, W.; Dong, Y.; Liu, X.; Bai, Y.; Zhang, L. The effect of parents' education on the academic and non-cognitive outcomes of their children: Evidence from China. *Child. Youth Serv. Rev.* **2020**, *117*, 105307. [CrossRef]
14. Cabrera, J.N. Positive Development of Minority Children. *Soc. Policy Rep.* **2013**, *27*, 1–22. Available online: https://www.srcd.org/sites/default/files/file-attachments/spr_272_final.pdf (accessed on 30 May 2025). [CrossRef]
15. Dodge, K.; Bai, Y.; Ladd, H.; Muschkin, G.C. Impact of North Carolina's Early Childhood Programs and Policies on Educational Outcomes in Elementary School. *Child Dev.* **2016**, *88*, 996–1014. [CrossRef]
16. Arango, M.; Nimnicht, G.; Penaranda, F. Twenty Years On: A report of the PROMESA programme in Colombia. Early Childhood Development: Practice and Reflections, No. 22, The Hague, Bernard van Leer Foundation. 2004. Available online: https://issuu.com/bernardvanleerfoundation/docs/twenty_years_on_a_report_of_the_promesa_programme_ (accessed on 30 May 2025).
17. Mladenov, C.; Dimitrov, E.; Kazakov, B. Demographical development of Bulgaria during the transitional period. *Méditerranée* **2008**, *110*, 117–123. [CrossRef]
18. Imdorf, C.; Ilieva-Trichkova, P.; Stoilova, R.; Boyadjieva, P.; Gerganov, A. Regional and Ethnic Disparities of School-to-Work Transitions in Bulgaria. *Educ. Sci.* **2022**, *12*, 233. [CrossRef]
19. Ersozlu, Z.; Taheri, S.; Koch, I. A review of machine learning methods used for educational data. *Educ. Inf. Technol.* **2024**, *29*, 22125–22145. [CrossRef]
20. Frias-Martinez, E. Machine Learning for Education: Current State and Future Prospects. In *Radical Solutions for Artificial Intelligence and Digital Transformation in Education. Lecture Notes in Educational Technology*; Burgos, D., Branch, J.W., Tlili, A., Huang, R., Jemni, M., Stracke, C.M., de la Higuera, C., Looi, C.-K., Berrada, K., Eds.; Springer: Singapore, 2024. [CrossRef]
21. Ren, X. A hybrid model combining environmental analysis and machine learning for predicting AI education quality. *Sci. Rep.* **2025**, *15*, 12577. [CrossRef] [PubMed]
22. Song, C.; Shin, S.-Y.; Shin, K.-S. Implementing the Dynamic Feedback-Driven Learning Optimization Framework: A Machine Learning Approach to Personalize Educational Pathways. *Appl. Sci.* **2024**, *14*, 916. [CrossRef]
23. Menon, H.K.D.; Janardhan, V. Machine Learning Approaches in Education. *Mater. Today Proc.* **2021**, *43*, 3470–3480. [CrossRef]
24. Labadze, L.; Grigolia, M.; Machaidze, L. Role of AI chatbots in education: Systematic literature review. *Int. J. Educ. Technol. High Educ.* **2023**, *20*, 56. [CrossRef]

25. Davar, N.F.; Dewan, M.A.A.; Zhang, X. AI Chatbots in Education: Challenges and Opportunities. *Information* **2025**, *16*, 235. [CrossRef]
26. Mim, M.A.; Khatun, M.R.; Hossain, M.M.; Rahman, W.; Munir, A. Exploring Early Learning Challenges in Children Utilizing Statistical and Explainable Machine Learning. *Algorithms* **2025**, *18*, 20. [CrossRef]
27. Jin, X. Predicting academic success: Machine learning analysis of student, parental, and school efforts. *Asia Pac. Educ. Rev.* **2023**. [CrossRef]
28. Deng, J.; Huang, X.; Ren, X. A Multidimensional Analysis of Self-Esteem and Individualism: A Deep Learning-Based Model for Predicting Elementary School Students' Academic Performance. *Meas. Sens.* **2024**, *33*, 101147. [CrossRef]
29. Chu, T.S.; Ashraf, M. Artificial Intelligence in Curriculum Design: A Data-Driven Approach to Higher Education Innovation. *Preprints* **2025**, 2025050041. [CrossRef]
30. Sharif, S.; Theeng Tamang, M.; Mani, D.; Elmedany, W. Application of Clustering Algorithms to Enhance Personalized Learning through Recommendation Model. In Proceedings of the International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, Sakhir, Bahrain, 17–19 November 2024. [CrossRef]
31. Chen, Y.; Sun, J.; Wang, J.; Zhao, L.; Song, X.; Zhai, L. Machine Learning-Driven Student Performance Prediction for Enhancing Tiered Instruction. *arXiv* **2025**. [CrossRef]
32. Herath, D.; Dinuwan, C.; Ihalagedara, C.; Ambegoda, T. Enhancing Educational Outcomes Through AI Powered Learning Strategy Recommendation System. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 10. [CrossRef]
33. Hutter, F.; Kotthoff, L.; Vanschoren, J. *Automated Machine Learning: Methods, Systems, Challenges*; Springer: Berlin/Heidelberg, Germany, 2019; ISBN 9783030053185.
34. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
35. Knudsen, I.E. Sensitive Periods in the Development of the Brain and Behavior. *J. Cogn. Neurosci.* **2004**, *16*, 1412–1425. [CrossRef] [PubMed]
36. Mercan, Z.; Akpınar, Ü.; Kocaarslan, M. Cognitive abilities in early childhood: An exploration across gender, age group, school type, and parental educational status. *Bartın Univ. J. Fac. Educ.* **2024**, *13*, 705–721. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.