Machine Learning Classification Model for Detecting Academic Risk in the Development of Projective Spatial Thinking

Oscar Mendez, Hector Florez*

Universidad Distrital Francisco Jose de Caldas, Bogota, Colombia

Abstract

This study developed and evaluated a Machine Learning classification model for early detection of academic risk in the development of projective spatial thinking. Academic and sociodemographic variables from 153 students in basic and secondary education were analyzed using a Gradient Boosting Classifier algorithm. The model was trained to identify predictors of low performance and classify students according to their risk level. The results showed a good predictive performance (ROC AUC = 0.866). We conclude that the classification system is an effective tool for teachers, facilitating proactive identification of at-risk students and enabling the design of early, personalized pedagogical interventions to strengthen this fundamental cognitive skill.

Keywords

Spatial Thinking, Performance Prediction, Machine Learning, Educational Data Mining, Sociodemographic Factors

1. Introduction

The development of spatial thinking is a fundamental cognitive skill for academic success in various disciplines; however, there is evidence of persistent low performance in this area among primary and secondary school students [1]. Traditional pedagogical approaches, including the use of Information and Communication Technologies (ICT), often prove insufficient as they cannot process the complex network of academic and sociodemographic factors that influence each student's performance [2]. This limitation hinders the creation of effective and personalized intervention strategies.

Faced with this challenge, Educational Data Mining and Machine Learning (ML) techniques emerge as a powerful approach, capable of analyzing large datasets to identify patterns and make predictions [3, 4]. Previous research has demonstrated the feasibility of using ML models for the early detection of students at risk of dropping out or general low performance [3, 5]. Nevertheless, there is a gap in the application of these models to specific cognitive domains such as spatial thinking.

The development of models capable of early detecting students at risk of low performance has proven to be an effective application for predicting dropout and student success. Building on this, the present article addresses a gap in the literature by detailing the development and evaluation of a Machine Learning classification model specifically designed for the early detection of academic risk in the development of the projective space of spatial thinking. The study's objective is to answer the following research question: To what extent do sociodemographic and academic variables, analyzed with Machine Learning techniques, allow for the prediction of risk in the development of projective spatial thinking [1]?

This paper is structured as follows. Section 2 describes the theoretical framework related to this study. Section 3 presents the methodology used to develop the work. Section 4 explains the results obtained from the work. Section 5 presents a discussion of the most insightful elements. Section 6 presents the future work. Some recommendations are offered in section 7. Finally, Section 8 concludes the work.

ICAIW 2025: Workshops at the 8th International Conference on Applied Informatics 2025, October 8–11, 2025, Ben Guerir, Morocco
*Corresponding author.
omendeza@udistrital.edu.co (O. Mendez); haflorezf@udistrital.edu.co (H. Florez)
0009-0002-2353-6185 (O. Mendez); 0000-0002-5339-4459 (H. Florez)

2. Theoretical Framework

2.1. Spatial Thinking as a Psychological Construct

Spatial thinking is defined as the set of cognitive processes through which an individual mentally constructs, represents, and manipulates objects in space, as well as their relationships and transformations [6]. Far from being an innate ability, psychogenetic psychology, led by Jean Piaget, posits that the notion of space is a construct that is progressively developed through the subject's interaction with and action upon the physical world, and not from mere perception [7].

The Piagetian theory, which underpins this study, distinguishes a hierarchy in the development of spatial relations, each with direct implications for the student's capacity for abstraction:

- **Topological Relations:** These constitute the first understanding of space, focused on qualitative properties such as proximity, separation, order, and enclosure. At this stage, the child understands basic concepts of the object itself, without yet considering metrics or external perspectives [8].
- **Projective Relations:** These represent a more advanced stage where the subject is able to coordinate different points of view. This implies the ability to anticipate how an object is perceived from a perspective different from one's own, while preserving its fundamental properties. The analysis of the development of this space is the central focus of our research, as it is a prerequisite for formal geometry [9, 10].
- **Euclidean Relations:** These culminate the development with the ability to establish a coordinated and metric reference system, allowing for the conservation of distances, angles, and surfaces. Here, thinking becomes quantitative, and the foundations for abstract geometric reasoning are established.

This cognitive development does not occur in a vacuum; it is profoundly influenced by the student's physical, social, and cultural environment. The quality of pedagogical strategies and the richness of interactive experiences with the environment are decisive for the student to successfully transition through these stages.

2.2. Data Mining and Machine Learning in Education

The prediction of academic performance is a complex task due to the multiplicity of factors that affect it, from the family environment to pedagogical strategies. Educational Data Mining (EDM) emerges as a research field that applies computational methods to explore large volumes of educational data and discover patterns that would otherwise remain hidden [11].

Within EDM, Machine Learning (ML) offers a set of techniques for building predictive models from data [12, 13]. For this work, the problem is approached as a **supervised classification** task, where an algorithm "learns" from a set of labeled data (students with known performance) to be able to classify new cases [13]. This study develops a model for the early detection of students at risk of low performance, an application that has proven effective in in the development of projective spatial thinking [2, 3].

To ensure a systematic and reproducible process, this research adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodological framework [13]. As illustrated in Figure 1, this model proposes an iterative life cycle for data mining projects.

The phases of CRISP-DM, followed in this work, are:

1. **Business Understanding**, where the objectives are defined
2. **Data Understanding**, for an initial exploratory analysis
3. **Data Preparation**, which includes cleaning and transformation
4. **Modeling**, where ML algorithms are selected and trained
5. **Evaluation**, to measure the model's performance
6. **Deployment**, which involves integrating the model into a functional system

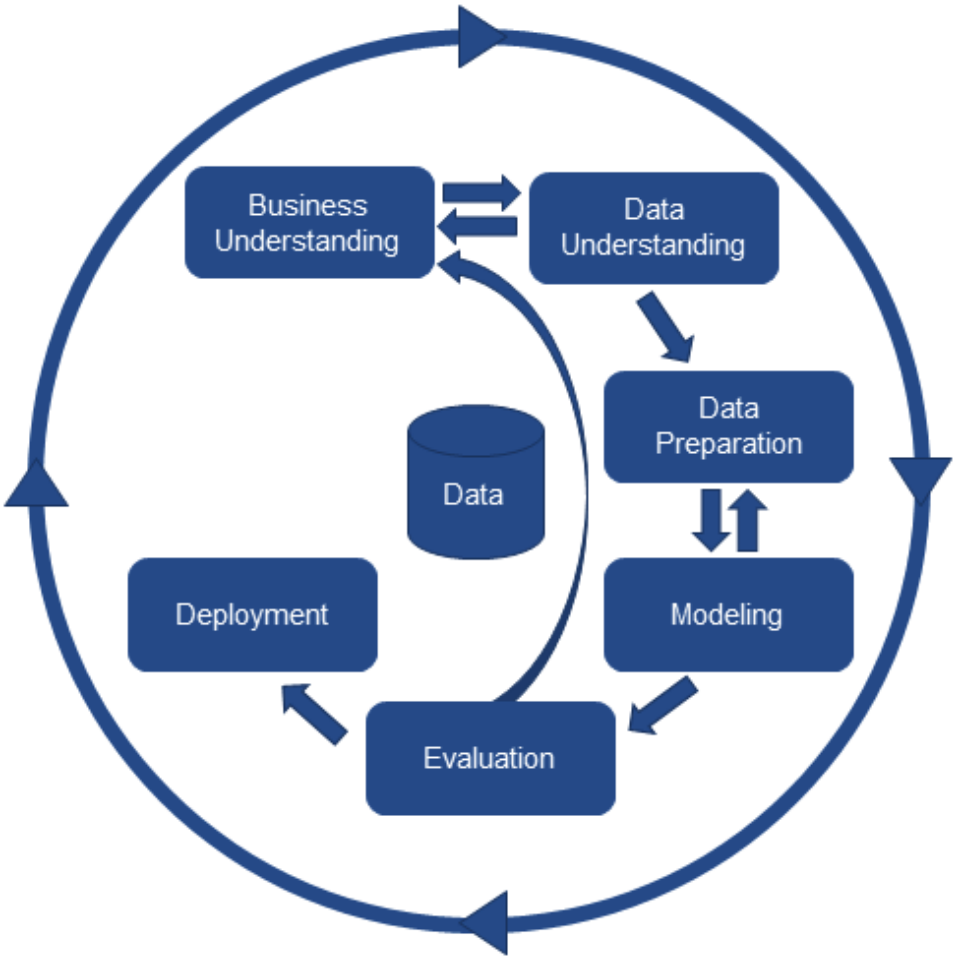

Figure 1: Workflow of the CRISP-DM methodology, showing its six interconnected phases. (Adapted from [1]).

3. Methodology

This research adopted a quantitative, correlational approach with the objective of determining the relationship between academic and sociodemographic variables and the prediction of performance in projective spatial thinking. The entire analysis and modeling process was structured following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodological framework. Figure 2 visually summarizes the complete workflow, from data collection to the final evaluation of the model.

3.1. Population and Sample

The study was conducted at the Institución Educativa Departamental General Carlos Albán, located in Albán, Cundinamarca (Colombia). The total population consisted of 599 students from urban and rural sectors. For this research, a non-probabilistic convenience sample of 153 students who voluntarily participated during the 2019 academic year was selected. The sample was distributed across the eighth, ninth, tenth, and eleventh grades, with ages ranging from 12 to 18 years.

3.2. Data Collection Instruments

Two main instruments were used for data collection:

- **Mock Test:** A 15-question spatial reasoning test, adapted from the 2010 entrance examination of the National University of Colombia, was administered. This instrument was designed to assess the students' level of development in projective space.

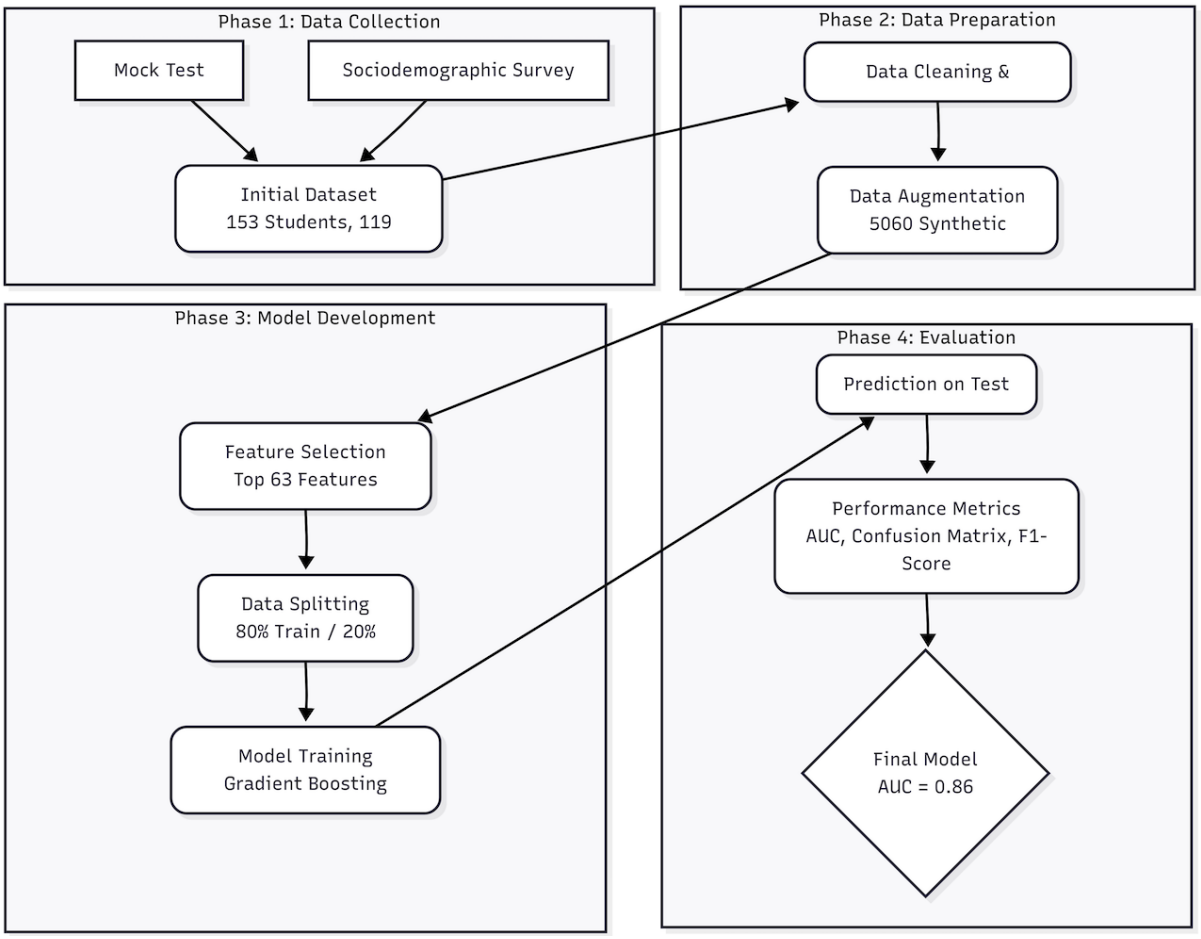

Figure 2: Methodological schematic of the research process.

- **Sociodemographic Factors Survey:** A survey composed of 33 groups of questions was designed and administered to characterize the sociodemographic conditions of the students. The survey variables were based on the dictionary of associated factors from the Colombian Institute for the Evaluation of Education (ICFES).

3.3. Analysis and Modeling Process (CRISP-DM)

The data processing and the construction of the predictive model were carried out in several stages, using the Python programming language and specialized libraries such as Pandas, Scikit-learn, and PyMC3.

3.3.1. Data Preparation

The initial dataset, composed of 119 variables, underwent an exploratory analysis. Missing values were imputed using the mean for numerical data and the mode for categorical data. Given the limited size of the original sample, a statistical simulation based on the Monte Carlo method was applied to generate a more robust synthetic dataset, preserving the original distributions of the variables.

3.3.2. Feature Selection

To optimize the model's performance and avoid overfitting, a feature selection process was carried out. A *Random Forest Classifier* algorithm was used to evaluate and rank the predictive importance of

each variable. A final subset of 63 variables was selected, which together explained 95% of the model's predictive capacity, a practice recommended in the literature for optimizing classifiers.

3.3.3. Model Building and Evaluation

Several supervised classification algorithms were evaluated, with the **Gradient Boosting Classifier** offering the best performance for this problem. The dataset was split into a training set (80%) and a test set (20%) to validate the model. The final performance was measured using the ROC (Receiver Operating Characteristic) curve and the area under it (AUC), a standard metric for evaluating the discriminative ability of a binary classifier.

4. Results

4.1. Diagnostic Performance in Spatial Thinking

The administration of the mock test to the sample of 153 students revealed a widespread low performance in the development of projective spatial thinking. The results indicated that a high percentage of students across all evaluated grades fell into the "Low Level" category. Specifically, it was found that 88% of eighth-grade students, 76% of ninth-grade, 98% of tenth-grade, and 84% of eleventh-grade students obtained an insufficient score. These findings empirically validated the existence of a significant educational problem in the area of study and provided the baseline performance data for training the model [1].

4.2. Feature Analysis and Model Selection

To build a robust model, a feature importance analysis was first conducted to identify the variables with the greatest predictive power. As shown in Figure 3, the most influential variables were not academic but rather sociodemographic and related to personal habits. The top five predictors were the time the student spends reading, the number of siblings, the number of people in the household, the time spent browsing the internet, and the Sisbén score.

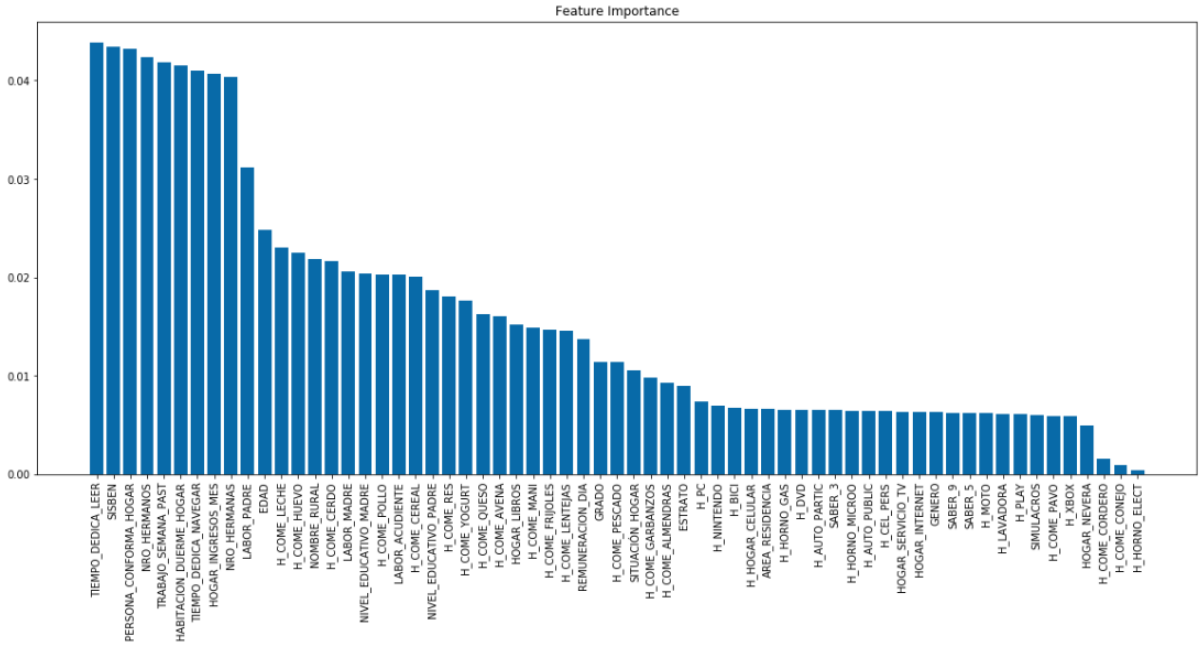

Figure 3: Importance of the 63 selected features. Sociodemographic variables and student habits proved to be the strongest predictors.

Subsequently, four classification algorithms were evaluated to select the most suitable one. Table 1 summarizes the performance of each. The **Gradient Boosting Classifier** was selected as the final model due to its superior performance (AUC of 0.86) on the test dataset.

Table 1
Performance Comparison (AUC) among Classification Models.

Algorithm	AUC on Test Set
K-Nearest Neighbors (KNN)	0.75
Support Vector Machine (SVM)	0.78
Random Forest	0.82
Gradient Boosting Classifier	**0.86**

4.3. Performance and Evaluation of the Final Model

The final *Gradient Boosting Classifier* model was optimized by tuning its hyperparameters, determining that 100 estimators offered the best balance between performance and overfitting (see Figure 4). The final evaluation of the optimized model yielded an **AUC = 0.86**, as observed in the ROC Curve in Figure 5, confirming a high discriminatory power.

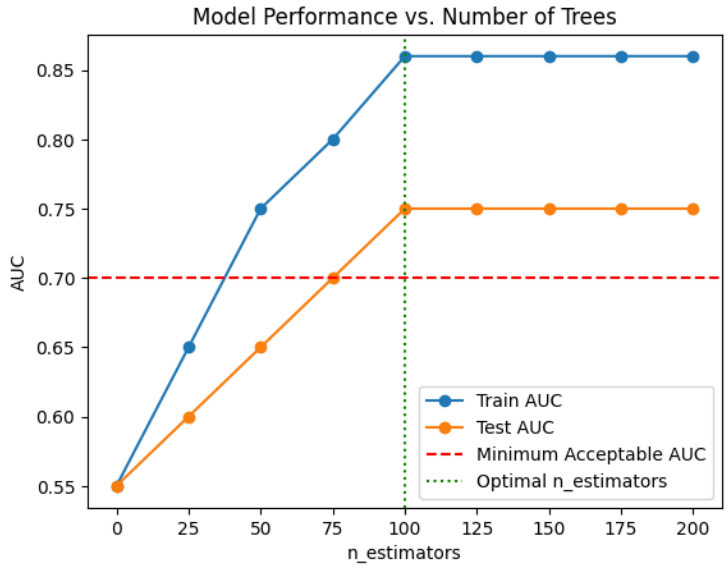

Figure 4: Tuning of the *n_estimators* hyperparameter.

The confusion matrix (Figure 6) details this performance on the 5060 samples of the test set. The model correctly identified **1626** students at risk (True Positives) and **2354** students not at risk (True Negatives). The classification report (Table 2) summarizes the key metrics, highlighting a *Recall* of 0.71 for the "Risk" class and an overall *Accuracy* of 79%.

Table 2
Detailed Classification Report of the Final Model.

Class	Precision	Recall	F1-Score	Support (Samples)
No Risk (0)	0.78	0.85	0.81	2769
Risk (1)	0.80	0.71	0.75	2291
Accuracy			**0.79**	**5060**
Macro Avg	**0.79**	**0.78**	**0.78**	**5060**
Weighted Avg	**0.79**	**0.79**	**0.78**	**5060**

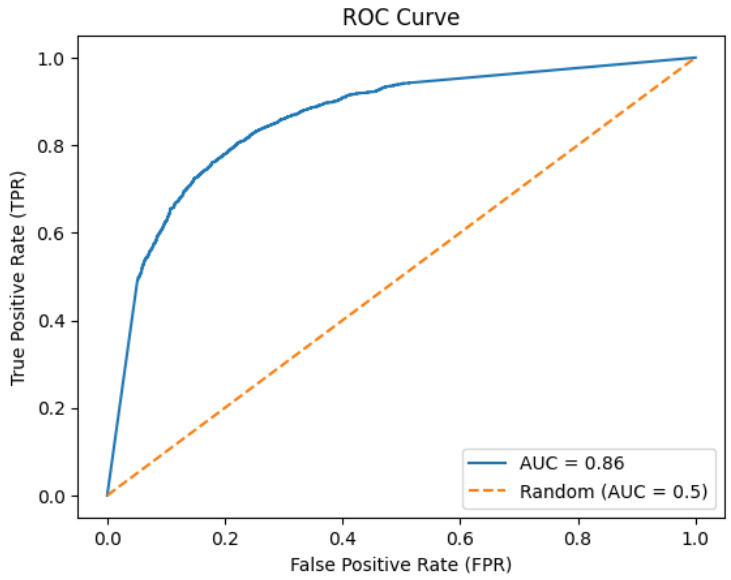

Figure 5: ROC Curve of the final model (AUC = 0.86).

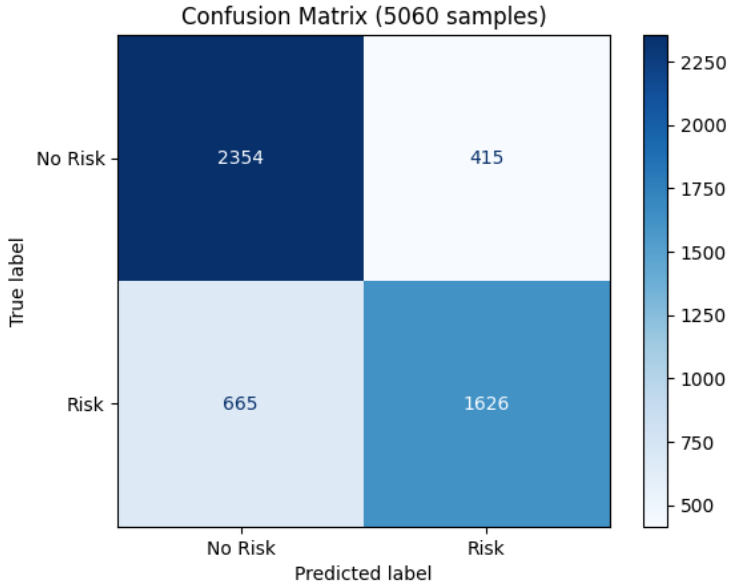

Figure 6: Confusion matrix of the model on the test set.

5. Discussion

The results of this research demonstrate that it is feasible to predict the risk of academic performance in projective spatial thinking with a high degree of certainty (AUC = 0.86) using a Machine Learning model. Beyond the technical validation of the model, the findings offer a deep insight into the factors that modulate cognitive development in primary and secondary school students.

A first significant finding emerges when contrasting the students' low performance on the diagnostic test with Piaget's theoretical framework [7]. Although the students were in an age range where they theoretically should have consolidated concrete operations, the majority showed deficiencies in projective tasks. This indicates that the development of spatial skills is not an automatic maturational process linked solely to age, but is strongly conditioned by the environment and the stimulation received, suggesting that educational and contextual experience plays a more critical role than might be assumed.

The model's feature importance analysis (Figure 3) strongly reinforces this idea. The model did not

identify previous grades as the strongest predictors, but rather sociodemographic variables and personal habits such as time spent reading, the number of people in the household, and the Sisbén score. This result aligns with studies that have demonstrated the profound influence of cultural and economic capital on academic performance [14, 15], and suggests that the student's environmental conditions are a determining factor in the development of complex cognitive skills.

Finally, the model's performance as a classification tool has direct practical implications for teaching. The classifier's ability to identify students "At Risk" (Recall = 0.71) and those "Not at Risk" (Recall = 0.85), as detailed in Table 2, makes it an effective early warning system [16, 17]. Although the model is not infallible—the 665 cases of False Negatives (Figure 6) represent at-risk students who were not detected and deserve special attention—it does provide a first layer of analysis that allows educators to focus their resources and design personalized pedagogical interventions. Instead of applying homogeneous strategies, teachers can use the model's results to nurture the potential of high-achieving students and, crucially, to offer focused support to those most likely to face difficulties.

6. Future Work

It is essential to acknowledge a methodological limitation of this study: the use of synthetic data generated from a small original sample to achieve sufficient volume for model training. Although this technique allowed for the construction of a robust classifier and the validation of the hypothesis, it introduces a potential bias, as the generated data inherits the characteristics and distribution of the initial sample.

Therefore, while the model has shown high potential, the generalization of these results should be considered with caution. The reproducibility of the use case in other populations or educational contexts will require empirical validation with a larger and more diverse organic dataset. Future lines of research should focus on applying and retraining this model with data collected on a larger scale to confirm and expand upon the findings presented here.

7. Recommendations

Based on the findings and limitations identified, the following recommendations are proposed:

- **Prioritize Organic Data in Future Iterations:** The most important recommendation is that future replications of this study prioritize the use of non-synthetic data. To validate and generalize the findings, it is essential to train and test the model on a larger and more diverse organic dataset, thereby eliminating the bias inherent in the simulation technique used.
- **Adopt a "Data Culture":** It is recommended that educational institutions foster an organizational culture that values data as a strategic asset. This involves the systematic and ethical collection of information to use predictive models like the one presented here and thus design personalized and proactive pedagogical strategies.
- **Deepen Causal Analysis:** Future research should go beyond prediction and delve into the causal analysis of the most influential variables. Understanding why "time spent reading" is such a strong predictor, for example, could generate valuable pedagogical knowledge.
- **Integration into Educational Systems:** Work should be done to integrate the predictive model into an interactive module within school information systems (SIS) or learning management systems (LMS). This would allow teachers and counselors to consult risk predictions in real-time, transforming the model from a research artifact into a daily support tool for decision-making [18].
- **Creation of an Alert and Manual Review Module:** It is recommended to develop an automated alert module. This system should establish a clear policy for cases where a student's predicted risk score exceeds a predefined threshold. Upon triggering an alert, the system should facilitate a protocol for a detailed manual review by the teacher, allowing for a human-centric intervention that complements the model's automated diagnosis.

8. Conclusions

This study affirmatively answers the research question, demonstrating that the application of Machine Learning techniques is an effective tool for predicting the risk of academic performance in the development of projective spatial thinking. The classification model validated the hypothesis that sociodemographic variables and personal habits are highly significant predictors. However, it is imperative to contextualize these good results, acknowledging that they were achieved using a dataset augmented through synthetic simulation. While this was a necessary methodological strategy for the analysis in this use case due to the small size of the original sample, it introduces a potential bias that must be considered.

Therefore, it is also concluded that the success of such projects depends on robust interdisciplinary collaboration, where education experts guide the interpretation of the context and technical results. The model should not be seen as an end in itself, but as a starting point for deeper inquiry.

Finally, it is crucial to address the ethical dimension raised by the high predictive power of socioeconomic variables, such as the Sisbén score. While effective for the model, their uncritical use risks reinforcing systemic biases and potentially stigmatizing students from vulnerable backgrounds. Therefore, it is emphatically concluded that the model must not be used as an automatic labeling tool. Instead, it should function as an early warning system that serves as a starting point for a teacher's inquiry. Its role is to **complement, not replace**, the professional judgment of the educator, who provides the indispensable human context and understanding required for any fair and effective intervention.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Ó. A. Méndez Aguirre, J. G. López Martínez, Técnicas de Machine Learning para la predicción de desempeño académico en el Desarrollo del espacio proyectivo del Pensamiento Espacial, Universidad Pedagógica Nacional, Bogotá, D.C., 2019. Master Thesis.

[2] P. H. Santoso, B. Setiaji, Y. Kurniawan, Wahyudi, S. Bahri, Fathurrahman, M. Kusuma, I. U. Wusqo, N. D. Muldayanti, A. D. Kurniawan, J. Syahbrudin, Students' performance dataset for using machine learning technique in physics education research, Scientific Data 12 (2025). doi:10.1038/s41597-025-04913-0.

[3] S. Ma, N. Jia, X. Wei, W. Zhang, Constructing a predictive model of negative academic emotions in high school students based on machine learning methods, Scientific Reports 15 (2025) 19183.

[4] J. Lawson, Data Science in Higher Education: A Step-by-Step Introduction to Machine Learning for Institutional Researchers, CreateSpace Independent Publishing Platform, 2015.

[5] O. Mendez, H. Florez, Applying the flipped classroom model using a vle for foreign languages learning, in: International Conference on Applied Informatics, Springer, 2018, pp. 215–227. doi:https://doi.org/10.1007/978-3-030-01535-0_16.

[6] A. Mukuka, J. K. Alex, Student teachers' knowledge of school-level geometry: Implications for teaching and learning., European Journal of Educational Research 13 (2024).

[7] J. Piaget, B. Inhelder, Psicología del niño, Morata, Madrid, 1997.

[8] J. Piaget, La representación del Espacio en el Niño, Presses Universitaires de France (PUF), Paris, 1947.

[9] J. Piaget, La Geometría Espontánea en el Niño, Presses Universitaires de France (PUF), Paris, 1948.

[10] Y. Abuhasanein, A. Rejeb, K. Jemai, Challenges to learning mathematical concepts among sixth-grade students in primary education: A teachers' perspective, International Electronic Journal of Mathematics Education 20 (2025) em0818.

[11] G. Rabinowitz, K. S. Moore, S. Ali, M. Weckel, I. Lee, P. Gupta, R. Chaffee, Study of an effective machine learning-integrated science curriculum for high school youth in an informal learning setting, International Journal of STEM Education 12 (2025). doi:10.1186/s40594-025-00543-5.

[12] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, O'Reilly Media, Inc., 2019.

[13] Y. Lou, K. F. Colvin, Performance prediction using educational data mining techniques: a comparative study, Discover Education 4 (2025) 1–14.

[14] Y. Velázquez Narváez, M. A. González Medina, Factores institucionales y socioeconómicos en el rendimiento académico: Un estudio longitudinal en universidades públicas mexicanas (2018-2023), Revista de la Educación Superior 53 (2024) 45–68. doi:10.33412/rev.educ.sup.2024.210.3.

[15] V. autores, Predicting academic performance in public universities using hybrid ai models, IEEE Transactions on Learning Technologies 17 (2024) 102–115.

[16] A. Villar, C. R. V. de Andrade, Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study, Springer Nature Computer Science 4 (2024). doi:https://doi.org/10.1007/s44163-023-00079-z.

[17] D. Perez Bravo, M. A. Zanata Alves, L. A. Ensina, L. E. Soares de Oliveira, Evaluating strategies to predict student dropout of a bachelor's degree in computer science, Anais do Symposium on Knowledge Discovery, Mining and Learning (KDMiLe) 11 (2024) 1–8. doi:10.5753/kdmile.2023.232763.

[18] O. Mendez, J. Camargo, H. Florez, Machine learning operations applied to development and model provisioning, in: International Conference on Applied Informatics, Springer, 2024, pp. 73–88. doi:https://doi.org/10.1007/978-3-031-75144-8_6.