ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

A Hybrid Machine Learning and Mathematical Modeling Approach for Predicting Academic Performance Using Latent Dimensions

Rajoelison Andrianarimalala Abel[1]; Rakotomalala Vololona Harinoro[1]; Rasolomampiandry Gilbert[1]

[1]Doctoral School of Engineering and Innovation Sciences and Technologies
[1]Doctoral Research Team : Cognitive Sciences and Applications
[1]Research Laboratory : Cognitive Sciences and Applications
University of Antananarivo, Madagascar

Publication Date: 2026/05/12

Abstract : Predicting academic performance is a major challenge in higher education, particularly in contexts where student monitoring systems are limited. While machine learning models have demonstrated strong predictive capabilities, their lack of interpretability limits their applicability in educational settings. This study proposes an interpretable model, the Academic Performance Index (API), based on an integrated framework combining mathematical modeling, cognitive sciences, and machine learning. Using real student data, a Random Forest model is employed to identify the most relevant variables and estimate their relative importance. These weights are then used to construct a composite index structured around four latent dimensions : cognitive, motivational, socio-economic, and environmental. The API is subsequently incorporated into a logistic regression model to estimate the probability of academic success. The results show that the proposed model achieves high predictive performance (AUC = 0.897; F1-score = 0.84), comparable to advanced approaches, while providing improved interpretability. The analysis highlights the central role of cognitive factors in academic success.

Keywords : *Machine Learning, Mathematical Modeling, Latent Variables, Random Forest, Academic Performance.*

How to Cite: Rajoelison Andrianarimalala Abel; Rakotomalala Vololona Harinoro; Rasolomampiandry Gilbert (2026) A Hybrid Machine Learning and Mathematical Modeling Approach for Predicting Academic Performance Using Latent Dimensions. *International Journal of Innovative Science and Research Technology*, 11(4), 4181-4190. https://doi.org/10.38124/ijisrt/26apr1548

I. INTRODUCTION

Predicting academic performance has become a major research focus in learning analytics, at the intersection of machine learning, cognitive science, and educational research [1][2]. The ability to anticipate students' trajectories of success or failure represents a strategic lever for improving pedagogical support systems and developing data-driven early warning mechanisms based on reliable empirical evidence [3].

Recent approaches based on machine learning, such as Random Forest and boosting methods, have demonstrated strong predictive performance in analyzing academic outcomes [4][5]. However, their often opaque nature, commonly referred to as "black-box" models, limits their adoption in educational contexts, where understanding decision-making mechanisms is essential to support informed and ethically responsible pedagogical actions [6].

Moreover, academic success cannot be reduced to simple statistical correlations. It relies on complex mechanisms involving cognitive, motivational, and contextual factors, extensively studied in cognitive science and the sociology of education [7]. Despite this, relatively few studies successfully integrate these theoretical foundations into predictive models that are simultaneously accurate, interpretable, and operational.

In this context, this paper proposes a hybrid approach based on the construction of an Academic Performance Index (API), designed as an interpretable mathematical model structured around four latent dimensions : cognitive, motivational, socio-economic, and environmental. The weights associated with these dimensions are empirically estimated using a Random Forest model and subsequently incorporated into a probabilistic framework based on logistic regression. This approach aims to reconcile mathematical rigor, empirical validation through machine

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

learning, and interpretability aligned with the requirements of applied cognitive sciences [8].

The main contributions of this work are threefold: the proposal of an interpretable model derived from a high-performing machine learning algorithm, the mathematical formalization of a composite index grounded in theoretically justified dimensions, and the empirical validation of the proposed model on real-world data, demonstrating a relevant trade-off between predictive performance and interpretability.

II. MATERIALS AND METHODS

Data collection was conducted using an online questionnaire administered through a digital form, allowing for standardized distribution among the targeted student population. This approach was complemented by field visits to facilitate participation, ensure response quality, and contextualize the collected data. Data processing and analysis were carried out in Python within a Jupyter Notebook environment, providing a structured and reproducible workflow. The main libraries used included Pandas for data handling and preprocessing, NumPy for numerical computations, and Matplotlib and Seaborn for visualization. Machine learning methods were implemented using the Scikit-learn library, particularly for feature selection, Random Forest modeling, and logistic regression. Together, these tools ensured consistency throughout the analytical process, from data collection to modeling and performance evaluation.

The methodology adopted in this study is based on an empirical approach combining real-world data analysis and machine learning techniques to identify the key determinants of academic performance. The dataset consists of individual student information, including academic, cognitive, motivational, socio-economic, and environmental variables. These variables were selected in accordance with the literature in educational sciences and learning analytics, which emphasizes the multidimensional nature of academic success [9]. To identify the most relevant predictors, a Random Forest model was employed. This ensemble method, based on multiple aggregated decision trees, enables the estimation of the relative importance of each variable in predicting academic performance while effectively handling non-linear interactions and complex relationships among variables [10]. Variable importance was assessed using standard measures such as impurity reduction (Gini index) and permutation-based accuracy loss, ensuring a robust and empirically grounded feature selection process [11].

The selected variables were subsequently organized into theoretically consistent latent dimensions. Each variable was assigned to one of four dimensions defined in the model : cognitive, motivational, socio-economic, and environmental. This step is grounded both in empirical evidence derived from the Random Forest model and in conceptual frameworks from cognitive science and the sociology of education, ensuring meaningful interpretability of the proposed model [12]. Structuring the variables into dimensions transforms a heterogeneous set of predictors into an organized and interpretable system, facilitating the subsequent construction of a composite academic performance index.

The construction of dimensional scores is based on a weighted aggregation of previously normalized variables belonging to each latent dimension. For each student iii, the cognitive, motivational, and socio-economic dimensions are computed as the average of their respective standardized variables. This standardization ensures comparability across variables of different scales and prevents biases related to measurement units [13][14]. Each dimension is therefore obtained as a normalized average of the variables composing it :

For the cognitive dimension (C) :

$$C_i = \frac{1}{|C|}\sum_{j\in C} X_{ij}^{(std)}$$

For the motivational dimension (M) :

$$M_i = \frac{1}{|M|}\sum_{j\in M} X_{ij}^{(std)}$$

For the socio-economic dimension (S) :

$$S_i = \frac{1}{|S|}\sum_{j\in S} X_{ij}^{(std)}$$

For the environmental dimension (E) :

$$E_i = x_{i,ID\_NUM}$$

Where :

$X_{ij}^{(std)}$: the standardized value of the variable j for student i :

$$X_{ij}^{(std)} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

μ_j : representing the mean of variable j et σ_j : its standard deviation.

The Academic Performance Index (API) is defined as a weighted linear combination of the four latent dimensions. For each student iii, the index is formulated as follows [15][16] :

$$API_i = \alpha S_i + \beta C_i + \gamma M_i + \delta E_i$$

Where the coefficients $\alpha, \beta, \gamma, \delta$ represent the relative contributions of each dimension. These coefficients are estimated from the overall importance measures observed in the data and are constrained by the condition $\alpha + \beta + \gamma +$

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

$\delta = 1$. This constraint ensures both interpretability and stability of the index. Importantly, these coefficients are not arbitrarily defined but are directly derived from the variable importance scores obtained through the Random Forest model. Specifically :

The Random Forest model provides, for each variable X_j, an importance weight w_j such that :

$$\sum_{j=1}^{p} w_j = 1$$

The variables are grouped into four dimensions : S, C, M and E.
For each dimension D ∈ {S, C, M, E}, the total importance is computed as the sum of the importance weights of the variables belonging to that dimension :

$$W_D = \sum_{j \in D} w_j$$

To link the Academic Performance Index (API) and its underlying dimensions to the probability of academic success, a logistic regression model is employed. This choice is justified by its probabilistic nature, statistical robustness, and direct interpretability [17]. The conditional probability of academic success is modeled as follows :

$$P(Y_i = 1|S_i, C_i, M_i, E_i) = \sigma(\theta_0 + \theta_S S_i + \theta_C C_i + \theta_M M_i + \theta_E E_i)$$

Where :

$Y_i \in \{0,1\}$: performance of student i (0 = average performance, 1 = high performance),
θ_0 : Intercept,
$\theta_S, \theta_C, \theta_M, \theta_E$: coefficients estimated from the data,
$\sigma(z)$: logistic function defined as :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The parameters $\theta_S, \theta_C, \theta_M, \theta_E$ are estimated using maximum likelihood estimation with L2 (Ridge) regularization. The objective function to be minimized is given by :

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} [y_i \, log \, p_i + (1 - y_i) \, log \, (1 - p_i)] + \lambda(\theta_S^2 + \theta_C^2 + \theta_M^2 + \theta_E^2)$$

Where :

$p_i = P(Y_i = 1|S_i, C_i, M_i, E_i)$ et $\lambda > 0$ is the regularization parameter.

III. RESULTS AND DISCUSSION

➢ *Comparison of the Six Machine Learning Models :*

The predictive performances obtained on the test set (n=120) are presented in Table 1.

Table 1 Raw Performance Results of the Models

Model	Accuracy	F1-macro	F1-CV (4 folds)
Logistic Regression	0.85	0.85	0.887
RandomForest	0.85	0.85	0.887
Gradient Boosting	0.84	0.84	0.870
SVM (linear)	0.81	0.81	0.888
KNN	0.82	0.82	0.884
Multinomial Naive Bayes	0.82	0.82	0.841

The results indicate that Logistic Regression and Random Forest achieve the highest overall predictive performance. However, given the objective of developing an interpretable model within this study, the Random Forest model is selected as the reference approach. This choice is motivated by its ability to provide variable importance measures, which are subsequently used to construct the latent dimensions S, C, M and E underlying the Academic Performance Index (API).

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

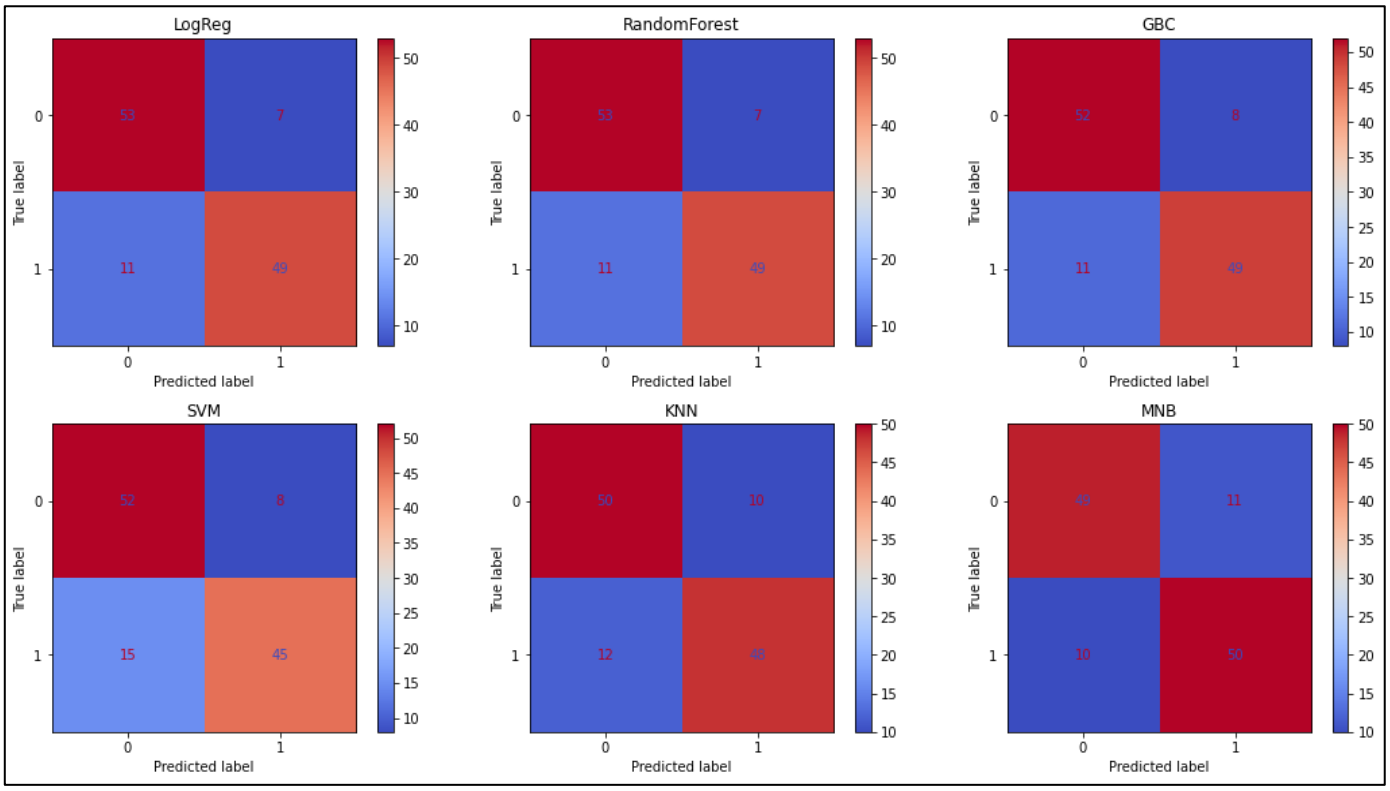

Fig 1 Confusion Matrices of the Tested Models

This figure illustrates the classification performance of each model on the 120 test observations. The main observations are as follows: Random Forest and Logistic Regression exhibit nearly identical classification structures (53/7 and 49/11, respectively). The SVM model shows a higher number of errors for class 1, with 15 false negatives. KNN and Multinomial Naive Bayes present relatively balanced confusion matrices, although with slightly lower overall performance. Gradient Boosting demonstrates good detection of class 1 but produces more misclassifications for class 0. Overall, Random Forest and Logistic Regression maintain the best balance between the two classes.

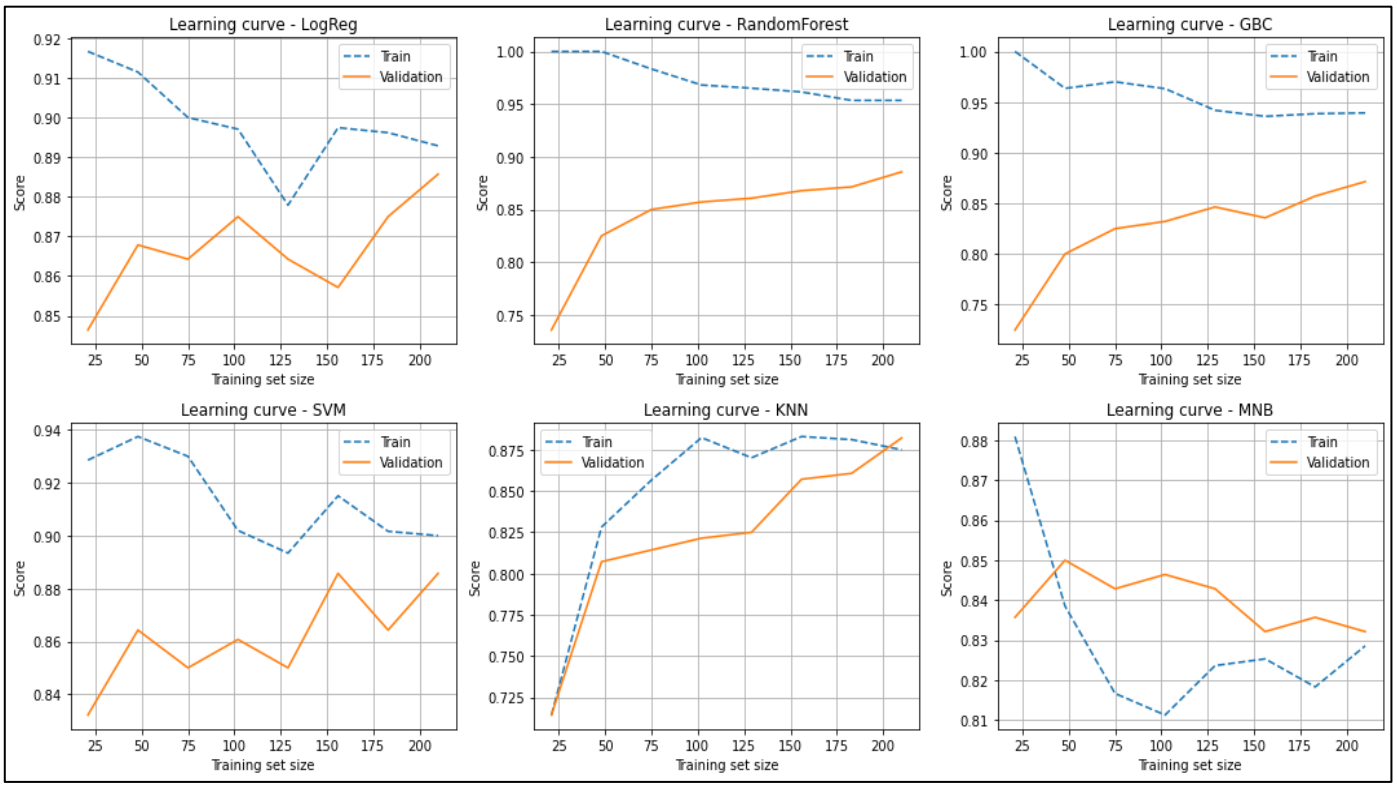

Fig 2 Learning Curves of the Six Models

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

This figure provides insight into model stability, the presence of overfitting, and performance trends as the sample size increases. The learning curves reveal distinct behaviors across the models. Logistic Regression shows closely aligned training and validation curves, indicating strong generalization capacity, with slight improvement as the sample size grows. Random Forest achieves a very high training score, close to 1, while the validation score stabilizes around 0.88, suggesting mild overfitting but overall stable performance, confirming its ability to capture complex variable interactions. Gradient Boosting exhibits a gradual increase in validation performance, reflecting stable learning behavior from moderate sample sizes. The SVM model presents a moderate gap between training and validation curves, with acceptable performance that reaches a plateau relatively early. KNN achieves a very high training score, while its validation performance improves progressively with increasing data, indicating sensitivity to noise and lower stability. Finally, the Multinomial Naive Bayes model shows significant variability in training performance, suggesting a weaker fit to the underlying data structure. Overall, Random Forest emerges as the most balanced model in terms of robustness, stability, and predictive performance.

- *Feature Importance Based on the Selected Model (Random Forest) :*

The optimized Random Forest model provides the following variable importance rankings :

Table 2 Most Influential Variables

Rank	Variable	Importance	Dimension
1	COM_SC (Scientific communication skills)	0.236	C
2	AI (Use of AI in studies)	0.205	C
3	NTIC (ICT proficiency)	0.102	C
4	ID_NUM (Digital identity / online resources)	0.070	E
5	MENT_BAC (High school honors)	0.067	C
6	MAT_REDOUB (History of repeating a year)	0.060	C
7	TMP_ETU (Study time)	0.059	M
8	TMP_LOIS (Leisure time)	0.026	M
9	INFO (Academic information tracking)	0.026	C
10	PRI_TACH (Task prioritization)	0.020	M
11	NIV_MERE (Mother's education level)	0.019	S
12	NIV_PERE (Father's education level)	0.017	S
13	LIEU_RES (Place of residence)	0.015	S
14	SIT_ECO (Socio-economic status)	0.014	S
15	SX (Gender)	0.013	S
16	SERIE_BAC (High school track)	0.012	C
17	STAT_PARENT (Parental status)	0.008	S
18	AG (Age)	0.008	S
19	TRAV_GRP (Group work)	0.007	M
20	EMP_ETU (Student employment)	0.006	M
21	COURS_SUP (Supplementary courses)	0.005	M
22	CLUB_ASS (Association involvement)	0.004	M

The results reveal a clear hierarchy among the predictors of academic performance. The three most influential variables COM_SC (scientific communication skills), IA (AI-related competencies), and NTIC (ICT proficiency), all reflect cognitive and digital capabilities directly mobilized in higher education. These findings indicate that, within the studied sample, academic success is primarily driven by the ability to structure scientific reasoning, effectively use digital tools, and engage with emerging technological concepts. This strongly supports the central role of the cognitive dimension in explaining performance.

Building on the importance scores derived from the Random Forest model, a structured interpretation can be developed across the four latent dimensions S, C, M, and E.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

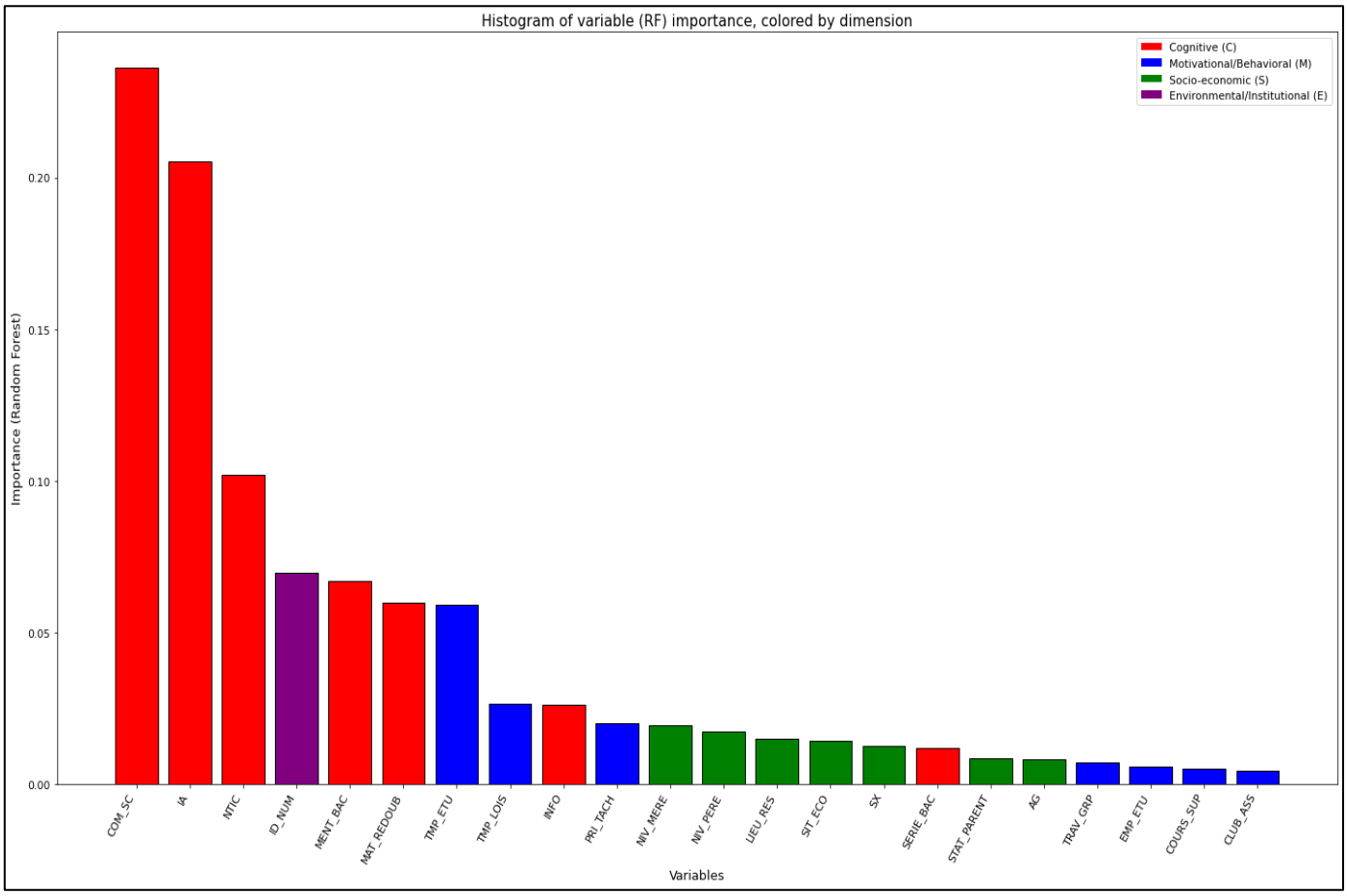

Fig 3 Importance of Variables (RF)

The cognitive dimension (red) clearly dominates the model. Variables such as COM_SC and IA emerge as the most influential predictors, followed by NTIC, MENT_BAC, MAT_REDOUB, INFO, and SERIE_BAC. This concentration at the top of the ranking indicates that academic performance primarily depends on students' ability to mobilize advanced cognitive resources, particularly in scientific reasoning and digital environments. These competencies reflect both prior academic capital and the capacity to adapt to modern tools for knowledge production and analysis.

The motivational dimension (blue) occupies a secondary but meaningful position. Variables such as TMP_ETU (study time) and TMP_LOIS (leisure time) highlight the role of time management and daily organization. Other factors, including PRI_TACH, TRAV_GRP, EMP_ETU, COURS_SUP, and CLUB_ASS, contribute to a lesser extent but collectively describe behavioral regulation patterns. This dimension captures students' ability to plan, collaborate, and balance academic and non-academic commitments, suggesting that performance is also shaped by structured learning behaviors.

The socio-economic dimension (green) shows a moderate contribution. Variables such as parental education levels, socio-economic status, residence, and family context indicate that background conditions still influence academic outcomes. However, their relative importance remains lower than that of cognitive factors, suggesting that while structural inequalities persist, they are not the primary drivers of performance within this framework.

The environmental dimension (purple) is mainly represented by the variable ID_NUM, which reflects administrative stability and institutional integration. Although its contribution is smaller, it highlights the importance of consistent engagement with institutional systems and digital infrastructures, acting as an indirect indicator of continuity in the academic pathway.

Overall, the figure reveals a structured hierarchy of determinants: cognitive factors dominate, followed by motivational influences, while socio-economic and environmental dimensions play complementary but non-negligible roles. This organization justifies the adopted modeling approach, which aggregates the 22 variables into four latent dimensions prior to constructing a composite and interpretable academic performance index.

➢ *Mathematical Modeling of Academic Performance :*

- *Structural Coefficients (α, β, γ, δ)*

The aggregated weights derived from the Random Forest results are as follows : $W_S = 0{,}094539$; $W_C = 0{,}708104$; $W_M = 0{,}127620$ et $W_E = 0{,}069737$.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

Accordingly, the weights of the Academic Performance Index (API), derived from the Random Forest model, are estimated as: $\alpha = W_S \approx 0{,}095$, $\beta = W_C \approx 0{,}708$, $\gamma = W_M \approx 0{,}128$, $\delta = W_E \approx 0{,}070$. These coefficients directly reflect the relative contribution of each dimension to academic performance, as empirically learned from the data. They enable the construction of a structural model grounded not in prior assumptions, but in the extraction of latent relationships captured by a machine learning algorithm, which constitutes the central contribution of this approach.

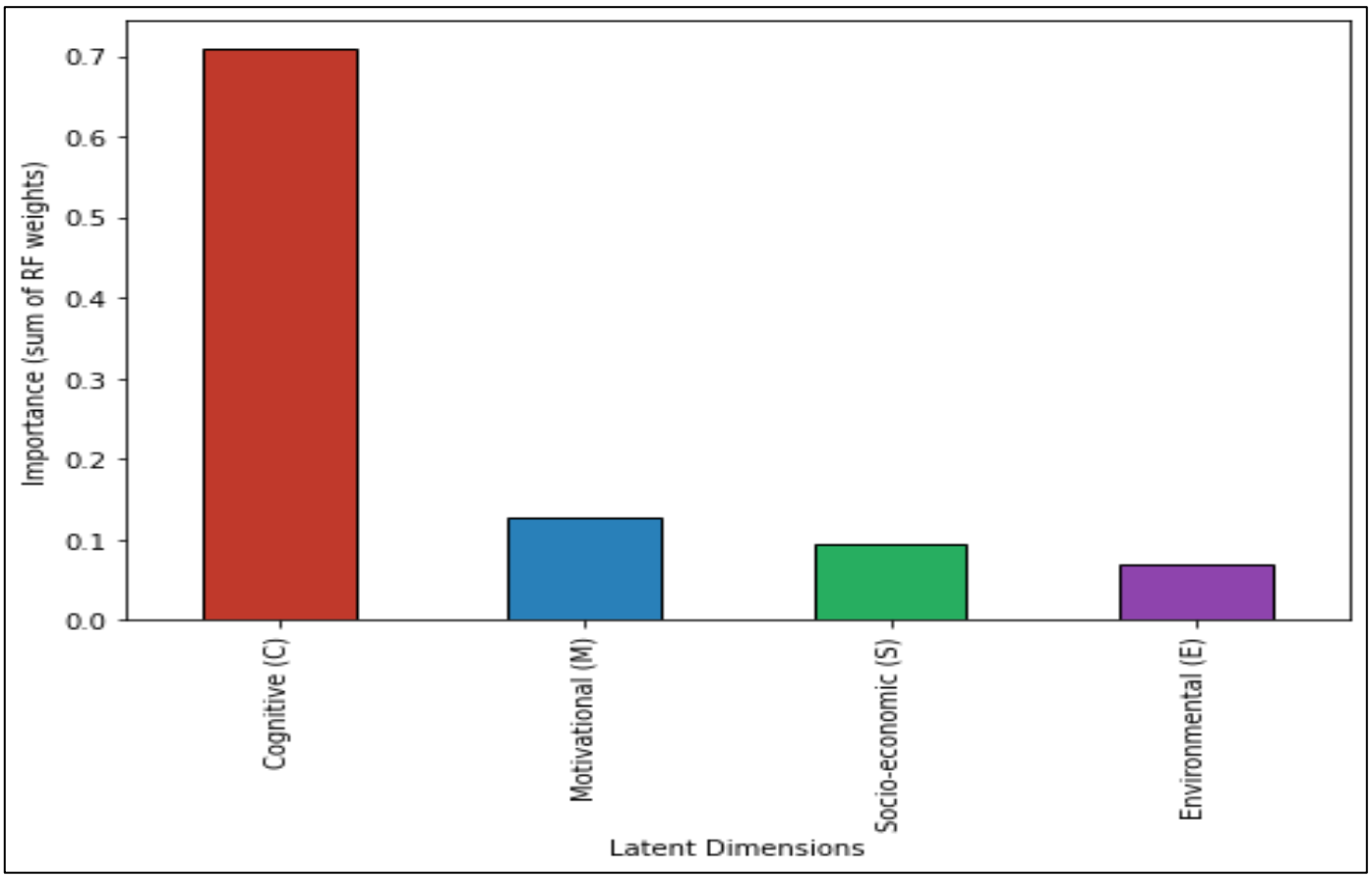

Fig 4 Contribution of Latent Dimensions in the API

The figure clearly highlights the predominance of the cognitive dimension (C) within the Academic Performance Index. With an aggregated weight of approximately 0.71, the cognitive component alone accounts for more than two-thirds of the predictive information captured by the Random Forest model. In practical terms, variables related to prior academic achievement, such as high school track and honors, repetition history, scientific communication skills, and proficiency in computing, ICT, and artificial intelligence play a decisive role in the probability of academic success. This observation is consistent with existing literature, which emphasizes that prior academic preparation and digital literacy strongly condition students' ability to engage with higher education requirements and adopt effective learning strategies.

The motivational dimension (M) ranks second, with a weight of approximately 0.13. Although substantially lower than the cognitive contribution, it remains meaningful. Variables such as study time, motivation for pursuing advanced studies, engagement in group work, and task prioritization reflect students' behavioral and self-regulatory engagement. This suggests that a highly motivated student may partially compensate for certain cognitive limitations through consistent effort and effective learning habits.

The socio-economic (S) and environmental (E) dimensions, with respective weights of approximately 0.09 and 0.07, play a more moderate but still relevant role. Factors such as family economic status, parental education level, place of residence, and administrative stability define a contextual framework that may either support or constrain academic engagement. While these dimensions do not dominate the predictive structure, they contribute to shaping the conditions under which learning takes place.

Overall, the distribution of weights confirms that the proposed API is primarily structured around a strong cognitive core, modulated by motivational factors and contextualized by socio-economic and institutional conditions. The model does not ignore structural inequalities, but suggests that, within the studied sample, academic success is primarily driven by prior academic capital and university-related competencies, and subsequently influenced by the way students engage with their learning environment.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

- *Dimensional Scores S, C, M, E and the Academic Performance Index (API)*

The dimensional scores are computed as averages of standardized variables (z-scores) within each dimension. They are centered around zero and therefore directly comparable. By substituting the coefficients derived from the Random Forest model, the Academic Performance Index for student i is defined as :

$$API_i = 0{,}095S_i + 0{,}708C_i + 0{,}128M_i + 0{,}070E_i$$

This methodological choice ensures that the coefficients are directly calibrated from the Random Forest importance scores, while maintaining a sum close to unity, which facilitates interpretability. As a result, the API evolves on a consistent scale, generally ranging between −1 and +1 in our dataset. Students classified as high performers tend to exhibit positive API values, whereas those with lower academic performance are associated with negative scores. The API can thus be interpreted as a synthetic indicator of academic profile : the higher its value, the more favorable the overall profile for academic success.

- *Estimation and Regularization*

Logistic regression applied to the four dimensions (S, C, M, E) yields the following coefficients : $Intercept$: $\theta_0 = -0{,}207$; $\theta_S = 0{,}236$; $\theta_C = 3{,}187$; $\theta_M = 0{,}454$ and $\theta_E = 0{,}716$. The estimated model is therefore expressed as :

$$P(Y_i = 1|S_i, C_i, M_i, E_i) = \frac{1}{1 + \exp[-(-0{,}207 + 0{,}236S_i + 3{,}187C_i + 0{,}454M_i + 0{,}716E_i)]}$$

The coefficients provide a clear interpretation of the relative influence of each dimension. The cognitive coefficient ($\theta_C = 3{,}187$) is by far the most dominant, indicating that a one-unit increase in the cognitive score leads to a substantial rise in the log-odds of academic success. This finding is consistent with the Random Forest results, where the cognitive dimension also carried the highest weight ($\beta \approx 0{,}708$). The motivational ($\theta_M = 0{,}454$) and environmental ($\theta_E = 0{,}716$) coefficients are positive, suggesting complementary contributions to performance. In contrast, the socio-economic coefficient ($\theta_S = 0{,}236$) remains positive but comparatively weaker, indicating a more indirect effect.

➢ *Comparative Performance*

The comparative evaluation of the six machine learning models Logistic Regression, Random Forest, Gradient Boosting, SVM, KNN, and Naive Bayes highlights a key result. Random Forest emerges as the most accurate predictive model, with an accuracy of 0.85, an F1-score of 0.85, and an AUC of approximately 0.89. The proposed API model, designed for interpretability, achieves very similar performance, with an accuracy of 0.84, an F1-score of 0.84, and an AUC of approximately 0.897, confirming its ability to provide an effective trade-off between predictive performance and interpretability.

Table 3 Performance Comparison

Model	Accuracy	F1-score	AUC	Interpretability
Random Forest (RF)	0.85	0.85	0.89	Very low
Logistic Regression	0.85	0.85	0.89	Moderate
Gradient Boosting	0.84	0.84	0.87	Low
SVM	0.81	0.81	0.87	Very low
KNN	0.82	0.82	0.85	Very low
Multinomial Naive Bayes	0.82	0.82	0.84	Moderate
API model (proposed)	0.84	0.84	0.897	Excellent

The performance gap between Random Forest (0.85) and the API model (0.84) remains minimal, indicating that the interpretable model successfully reproduces most of the predictive behavior of the Random Forest despite its significantly lower complexity. Moreover, the AUC of the API model (0.897) slightly exceeds that of the Random Forest, demonstrating a strong ability to discriminate between high-performing and low-performing students. Overall, the API model provides a particularly relevant compromise, combining high predictive accuracy with direct interpretability, an advantage not offered by models such as Random Forest, SVM, or Gradient Boosting, which remain difficult to interpret in practice.

- *Visualization of Weights and the Academic Performance Index (API)*

The observed API values show a clear separation between performance groups. Students with high academic performance (Y=1) typically exhibit API values ranging from approximately +0.35 to +0.70, whereas those with average performance (Y=0) are mostly distributed between −0.60 and −0.10. This distinction suggests that the API functions as a meaningful latent indicator of academic performance.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

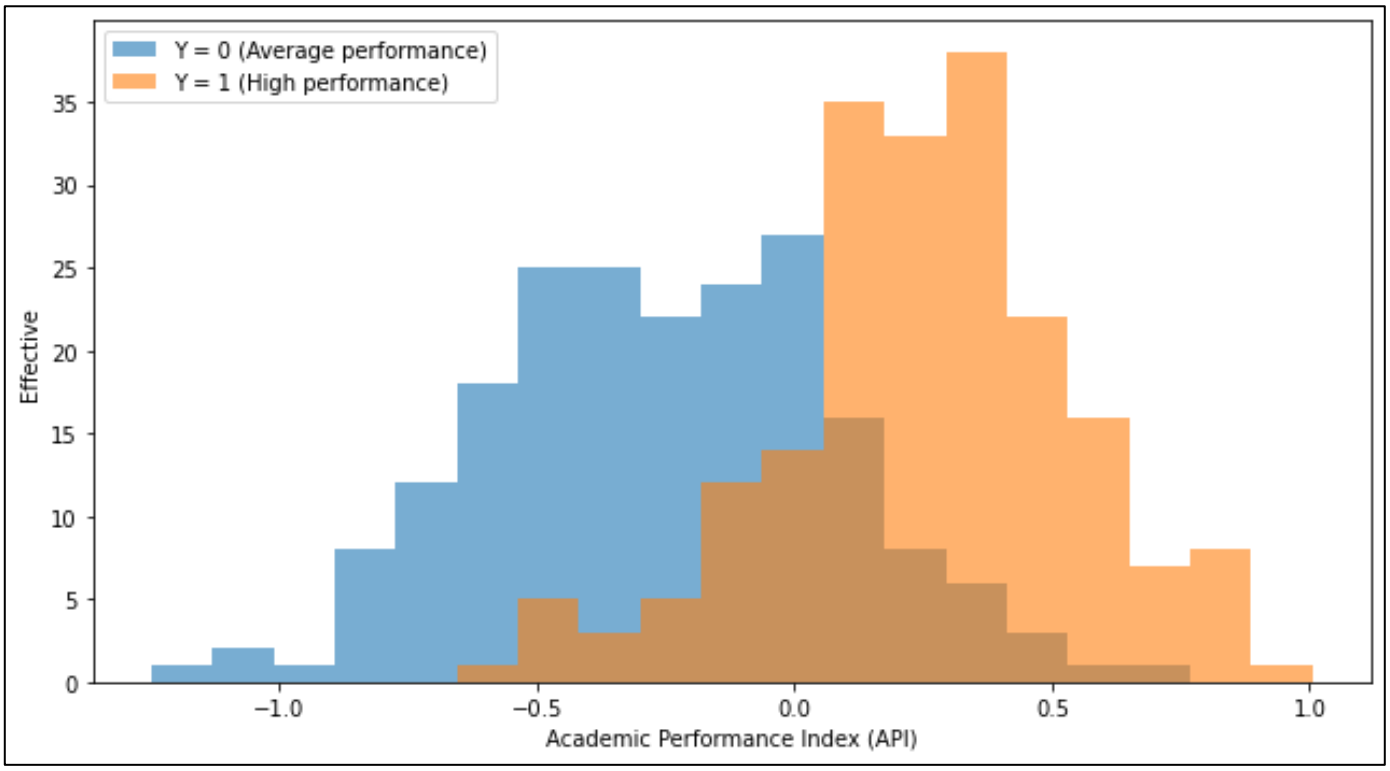

Fig 5 Distribution of the API by Performance Level

The histogram reveals a noticeable shift between the two API distributions. Students in class Y=0 (average performance) are predominantly concentrated on the lower end of the scale, while those in class Y=1 (high performance) are largely located on the higher end. In practical terms, an increase in the API is associated with a higher probability of belonging to the high-performing group. This confirms that the index effectively synthesizes information from the four latent dimensions S, C, M, and E, and operates as a coherent latent score of academic performance.

A partial overlap between the two distributions can nonetheless be observed. Some students with intermediate API values are classified as either Y=0 or Y=1. This overlap is expected in educational contexts, where academic outcomes are also influenced by unobserved factors such as examination conditions, personal circumstances, or health-related issues. From a methodological perspective, this indicates that the API should be interpreted as a probabilistic rather than deterministic measure: higher values increase the likelihood of success without guaranteeing it.

Finally, the clear separation between the central tendencies of the two groups, where the mean API for Y=1 is significantly higher than that for Y=0, demonstrates the strong discriminative capacity of the index. This supports the relevance of the proposed model : through a single continuous score, the API enables student classification, identification of at-risk profiles (low or intermediate scores), and the potential guidance of targeted academic interventions.

➢ *Discussion*

This study aimed to move beyond the traditional trade-off between predictive performance and interpretability in academic success prediction. The findings suggest that these two objectives can be reconciled through a hybrid modeling approach that combines machine learning, mathematical formalization, and cognitive science foundations.

From a theoretical perspective, the dominance of the cognitive dimension aligns with well-established explanations of academic achievement. Skills related to information processing, digital tool usage, and scientific reasoning have become central in higher education environments. This observation is consistent with cognitive load theory and broader cognitive approaches to learning, which emphasize that the ability to efficiently manage information is a key determinant of performance in complex and technology-rich academic settings.

From a methodological standpoint, the main contribution of this work lies in transforming a “black-box” model into an explicit, stable, and reproducible mathematical structure. Rather than relying solely on predictive accuracy, the proposed approach extracts latent relationships learned by a Random Forest model and reformulates them into an interpretable index. This directly addresses current challenges in explainable artificial intelligence, where the goal is not only to predict but also to understand and justify model outputs.

From an applied perspective, the Academic Performance Index offers practical value for higher education institutions. Its formal simplicity and interpretability make it suitable for pedagogical monitoring, early identification of at-risk students, and the design of

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1548

targeted support strategies. Unlike purely statistical models, the API provides actionable insights by linking predictions to theoretically grounded cognitive and behavioral dimensions.

Overall, this study shows that integrating machine learning within a structured theoretical framework allows for a shift from purely predictive models toward interpretable and operational tools. The proposed model is not intended to replace high-performance algorithms, but rather to complement them by enhancing their transparency and usability in real educational contexts.

IV. CONCLUSION

This research proposed an integrated approach to academic performance prediction based on the construction of an interpretable Academic Performance Index (API), combining mathematical modeling, machine learning, and insights from cognitive science. The originality of the model lies in its ability to balance predictive performance with interpretability, through the use of latent dimensions and empirically derived weights obtained from a Random Forest model.

The results demonstrate that the API achieves a level of predictive performance comparable to advanced machine learning models, while providing a clearer understanding of the factors influencing academic success. In particular, the analysis highlights the central role of the cognitive dimension, reinforcing the importance of learning processes and academic competencies in higher education outcomes. The inclusion of motivational, socio-economic, and environmental dimensions further contributes to a more comprehensive and nuanced representation of student performance.

From a methodological perspective, this study illustrates the value of a hybrid approach in which machine learning techniques are not used solely for prediction, but also as tools for constructing interpretable and theoretically grounded models. This perspective opens promising directions for the development of decision-support systems in educational contexts, particularly for student monitoring and targeted intervention strategies.

However, several limitations should be acknowledged. The generalizability of the findings depends on the size and characteristics of the dataset. In addition, certain dimensions, particularly the environmental component, could be further refined through the inclusion of more detailed variables. Future research may explore the integration of longitudinal data, the use of more advanced interpretable models, or the adaptation of the API framework to different educational contexts.

REFERENCES

[1]. Siemens, G. & Baker, R.S. (2012). Learning analytics and educational data mining: towards communication and collaboration. LAK. https://doi.org/10.1145/2330601.2330661
[2]. Romero, C. & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Trans. SMC. https://doi.org/10.1109/TSMCC.2010.2053532
[3]. Arnold, K.E. & Pistilli, M.D. (2012). Course signals at Purdue: using learning analytics to increase student success. LAK. https://doi.org/10.1145/2330601.2330666
[4]. Breiman, L. (2001). Random Forests. Machine Learning. https://doi.org/10.1023/A:1010933404324
[5]. Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD. https://doi.org/10.1145/2939672.2939785
[6]. Ribeiro, M.T., Singh, S. & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD. https://doi.org/10.1145/2939672.2939778
[7]. Pintrich, P.R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. JEP. https://doi.org/10.1037/0022-0663.95.4.667
[8]. Lundberg, S.M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS. https://doi.org/10.48550/arXiv.1705.07874
[9]. Baker, R.S. & Inventado, P.S. (2014). Educational data mining and learning analytics. In Learning Analytics. https://doi.org/10.1007/978-1-4614-3305-7_1
[10]. Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. R News. https://doi.org/10.32614/RJ-2002-022
[11]. Strobl, C. et al. (2007). Bias in random forest variable importance measures. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-8-25
[12]. Eccles, J.S. & Wigfield, A. (2002). Motivational beliefs, values, and goals. Annual Review of Psychology. https://doi.org/10.1146/annurev.psych.53.100901.135153
[13]. Han, J., Kamber, M. & Pei, J. (2012). Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5
[14]. Jolliffe, I.T. & Cadima, J. (2016). Principal component analysis: a review and recent developments. https://doi.org/10.1098/rsta.2015.0202
[15]. Breiman, L. (2001). Statistical modeling: The two cultures. https://doi.org/10.1214/ss/1009213726
[16]. Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013). Applied Logistic Regression. https://doi.org/10.1002/9781118548387
[17]. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. https://doi.org/10.5555/1953048.2078195