

Article

Supporting Reflective AI Use in Education: A Fuzzy-Explainable Model for Identifying Cognitive Risk Profiles

Gabriel Marín Díaz ^{1,2} 

¹ Faculty of Statistics, Complutense University, Puerta de Hierro, 28040 Madrid, Spain; gmarin03@ucm.es or gabriel.marin@universidadeuropea.es

² Science and Aerospace Department, Universidad Europea de Madrid, Villaviciosa de Odón, 28670 Madrid, Spain

Abstract

Generative AI tools are becoming increasingly common in education. They make many tasks easier, but they also raise questions about how students interact with information and whether their ability to think critically might be affected. Although these tools are now part of many learning processes, we still do not fully understand how they influence cognitive behavior or digital maturity. This study proposes a model to help identify different user profiles based on how they engage with AI in educational contexts. The approach combines fuzzy clustering, the Analytic Hierarchy Process (AHP), and explainable AI techniques (SHAP and LIME). It focuses on five dimensions: how AI is used, how users verify information, the cognitive effort involved, decision-making strategies, and reflective behavior. The model was tested on data from 1273 users, revealing three main types of profiles, from users who are highly dependent on automation to more autonomous and critical users. The classification was validated with XGBoost, achieving over 99% accuracy. The explainability analysis helped us understand what factors most influenced each profile. Overall, this framework offers practical insight for educators and institutions looking to promote more responsible and thoughtful use of AI in learning.

Keywords: generative AI; critical thinking; fuzzy clustering; explainability; AHP; digital maturity; cognitive profiles; education



Academic Editors: Xiaoming Zhai and Shuchen Guo

Received: 4 June 2025

Revised: 3 July 2025

Accepted: 15 July 2025

Published: 18 July 2025

Citation: Marín Díaz, G. (2025). Supporting Reflective AI Use in Education: A Fuzzy-Explainable Model for Identifying Cognitive Risk Profiles. *Education Sciences*, 15(7), 923. <https://doi.org/10.3390/educsci15070923>

Copyright: © 2025 by the author. 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

The integration of artificial intelligence (AI) in education is transforming how knowledge is acquired, processed, and applied. This shift is unfolding in a volatile, uncertain, complex, and ambiguous (VUCA) environment (Cernega et al., 2024), where the pace of innovation often exceeds the capacity for ethical and reflective adaptation.

Although AI tools are increasingly adopted in education, their influence on users' cognitive autonomy, critical engagement, and reflective thinking remains underexplored. While AI can enhance learning, content generation, and decision-making, it may also reduce deep reasoning and metacognitive effort especially when used uncritically.

Over the past decade, AI has evolved from a peripheral tool to a core component of digital ecosystems, particularly in higher education. As highlighted by (UNESCO, 2023), digital innovation has the potential to redefine and complement traditional pedagogical approaches, advancing inclusive and sustainable development in alignment with Sustainable Development Goal 4 (SDG 4) (Boeren, 2019).

Previous studies have extensively analyzed the acceptance of technology in education through models such as the Technology Acceptance Model (TAM) (Davis, 1989). However, a growing concern has emerged: over-reliance on AI tools may reduce cognitive effort and impair critical thinking processes.

These models tend to emphasize adoption-related factors such as perceived usefulness (PU), ease of use (PEU), or social influence (SI), yet they rarely evaluate how AI affects reflective practices, autonomous decision-making, or the risk of cognitive dependency (Lin & Yu, 2023; Marín Díaz et al., 2024).

While the academic literature acknowledges the growing use of AI-based systems for learning support, assessment, and content creation (Crompton & Burke, 2023), few studies have investigated whether such convenience might inadvertently undermine metacognitive development and deep reasoning.

To address this gap, this study proposes a multi-phase mathematical model to profile digital maturity and critical thinking in AI-augmented environments. The methodology integrates both adoption variables and behavioral–cognitive dimensions, including information verification, argumentation, reflection without AI, and the risk of over-reliance on automated outputs.

The objective of this research is to design, implement, and validate a fuzzy multicriteria model that segments users according to behavioral patterns and cognitive tendencies with AI use. This model combines methods from decision theory (AHP) (Saaty, 1977), fuzzy logic (2-tuple linguistic models) (Marín Díaz et al., 2023), unsupervised learning (Fuzzy C-Means) (Bezdek et al., 1984), and explainability (SHAP, LIME) (Molnar, 2019), offering a comprehensive approach that identifies user profiles and informs strategic actions that foster critical thinking in AI-enhanced learning contexts.

Fuzzy C-Means were selected over other clustering methods due to their ability to model the non-binary nature of user behavior. In contrast to hard clustering algorithms, FCM allows users to partially belong to multiple clusters, which is particularly suitable for representing cognitive ambiguity and behavioral overlap in educational contexts.

The study is guided by the following questions:

- Q1: What behavioral and cognitive dimensions are relevant for assessing digital maturity in AI use?
- Q2: Can distinct user profiles be identified using fuzzy clustering and linguistic modeling?
- Q3: How can explainability methods (SHAP, LIME) inform tailored strategies to enhance critical thinking?

The structure of the introduction has been revised to follow a logical sequence from general context to research gaps, objectives, questions, and methodological overview.

The methodology unfolds across several integrated phases. First, a structured questionnaire was developed encompassing five thematic blocks related to AI usage frequency, cognitive effort, decision-making, and openness to reflection. Each response was encoded and normalized for quantitative analysis. Second, the Analytic Hierarchy Process (AHP) was applied to weigh variables both within and across blocks, shaping the maturity index. Third, 2-tuple fuzzy linguistic representation was used to express individual scores in interpretable terms. Fourth, Fuzzy C-Means clustering segmented the aggregated data into cognitive–behavioral user profiles. These clusters were validated using XGBoost, ensuring consistency and predictive reliability. Fifth, SHAP and LIME were applied to interpret the contribution of each variable to clustering outcomes. Finally, the model produced profile-specific recommendations to support critical thinking and digital maturity, aligning pedagogical strategies with user traits.

As a result, the model assigns a digital maturity score to each segment, providing a foundation for the design of adaptive pedagogical and institutional strategies. The methodology was validated using a dataset of 1273 structured user profiles, selected to ensure demographic diversity and behavioral variability across cognitive dimensions.

This research contributes to ongoing debates around AI adoption and cognitive development by offering a scalable, interpretable, and human-centered diagnostic model applicable across educational and professional settings.

2. Related Work

2.1. AI Adoption in Education: Beyond Perceived Usefulness

The Technology Acceptance Model (TAM) (Davis, 1989), and its subsequent extensions TAM2 (Venkatesh & Davis, 2000), TAM3 (Venkatesh & Bala, 2008), and the Unified Theory of Acceptance and Use of Technology (UTAUT) (Andreas, 2012), has been widely used to evaluate technology adoption in educational and professional settings. These models focus on key constructs such as perceived usefulness (PU), perceived ease of use (PEU), attitude toward usage (ATT), and social influence (SI), which have traditionally served as predictors of users' intentions to adopt digital tools.

While these models offer solid foundations for understanding initial adoption behavior, they fall short in capturing the long-term cognitive implications of sustained AI use. They do not account for how AI technologies may influence users' ability to think critically, verify information independently, or engage in deep analytical reasoning.

Moreover, current adoption models often overlook the potential risks associated with AI-mediated learning, such as over-reliance on automation, decreased cognitive effort, and the erosion of metacognitive engagement during tasks that require problem-solving or critical analysis (Lin & Yu, 2023). These limitations highlight the need for complementary models that address functional adoption, cognitive maturity, and reflective practices in AI-enhanced learning environments.

2.2. Critical Thinking and Reflection in Automated Environments

Critical thinking is a foundational competence in education, particularly in the context of increasingly complex and rapidly evolving technology. It involves the ability to challenge assumptions, evaluate arguments, analyze information from multiple sources, and make well-reasoned decisions. However, the growing reliance on artificial intelligence (AI) for tasks such as information retrieval, content generation, and problem-solving raises new challenges for preserving and developing these cognitive abilities.

Several authors have expressed concern about the potential for cognitive offloading associated with the widespread use of AI tools (Risko & Gilbert, 2016; Salomon & Perkins, 1998). When users delegate analytical and decision-making tasks to algorithms, there is a risk of reducing both the frequency and depth of reflective thinking, especially among students with low metacognitive awareness.

Recent educational research has begun to explore how AI may influence learners' autonomy and critical engagement with content. For example, the importance of preserving human-centered pedagogical approaches in the age of intelligent systems has been emphasized in (Zawacki-Richter et al., 2019). However, other studies highlight the paradox that while AI can support learning, it can also encourage passive consumption of information if not properly integrated into reflective practice (Abdulmunem, 2023).

Moreover, there is growing interest in developing strategies that promote intentional reflection in AI-mediated learning environments. These include prompting users to verify outputs, engaging in debates about generated content, or designing activities that require independent argumentation without automated support (Facione, 2011; Qian & Clark, 2016).

Recent studies from 2022 to 2024 on critical thinking and AI in education (Favero et al., 2024; Hutson et al., 2022; Mintz et al., 2023), along with the latest UNESCO reports (UNESCO, 2023), have been included to strengthen this discussion. These works address the cognitive, ethical, and pedagogical implications of AI integration in higher education and help to situate the study within the ongoing academic discourse.

2.3. Multicriteria and Fuzzy Models in Educational Evaluation

Educational contexts often involve complex decision-making processes that are not fully captured in binary or deterministic models. In recent years, multicriteria decision-making (MCDM) techniques have gained relevance in educational research due to their ability to integrate diverse qualitative and quantitative indicators. Among them, the Analytic Hierarchy Process (AHP) has been widely used to prioritize factors in contexts such as curriculum design, educational technology evaluation, and student assessment (Chen et al., 2015; Saaty, 2008).

The use of fuzzy logic in combination with MCDM techniques has further enhanced the capacity to deal with subjective and uncertain data, which is inherent in educational evaluation. The 2-tuple linguistic representation model, proposed by (Herrera & Martinez, 2000a), allows for the expression of nuanced human judgments and enables more interpretable evaluations in settings where crisp scoring may not reflect the underlying complexity.

These models have been applied to various educational problems, including teacher performance evaluation (Chen et al., 2015), intelligent tutoring systems (Sottolare et al., 2014), and learning analytics models (Queiroga et al., 2020). Their ability to accommodate ambiguity makes them suitable for evaluating latent constructs such as critical thinking, digital maturity, or cognitive engagement, especially when user self-perception and behavioral indicators must be analyzed jointly.

Despite their proven effectiveness, these approaches have not yet been fully exploited to assess the intersection between AI adoption and human cognition. This gap motivates the integration of fuzzy and multicriteria techniques in the proposed model to evaluate user profiles across behavioral and cognitive dimensions.

2.4. User Profiling Through Clustering: Towards Personalized Models

In recent years, there has been growing interest in the personalization of educational experiences, driven by the need to adapt teaching and learning strategies to the diversity of learners' cognitive styles, motivations, and technological behaviors. Clustering techniques have played a key role in identifying latent user profiles based on behavioral, emotional, and academic data (Dutt et al., 2017; Romero & Ventura, 2010).

Among the clustering methods, Fuzzy C-Means (FCM) stands out for its capacity to model soft boundaries between groups, which is especially useful when individuals exhibit overlapping traits across dimensions such as motivation, reflection, or AI usage. Unlike traditional hard clustering algorithms (e.g., K-Means), FCM allows for degrees of membership in multiple clusters, better reflecting the nuance of real user behavior (Bezdek et al., 1984).

In educational contexts, fuzzy clustering has been applied to detect patterns in student engagement, identify at-risk learners, and inform adaptive learning systems (Al-Shabandar et al., 2017). However, most of these approaches focus exclusively on performance metrics, without incorporating broader concepts such as critical thinking, cognitive maturity, or ethical considerations in interactions with AI.

Furthermore, clustering outputs are often interpreted as fixed groupings, with little effort invested in making the grouping criteria transparent. This limitation underscores the

need for integrated models that combine clustering with interpretability layers (e.g., SHAP, LIME), enabling educators and policymakers to understand not just *how* users are grouped, but also *why*.

2.5. Explainability and Motivation for an Integrated Evaluation Framework

With artificial intelligence becoming more embedded in the education systems, concerns about transparency and interpretability in decision-making have grown significantly. It is no longer enough to generate accurate results; it is also essential to understand the factors behind those predictions, especially when they influence teaching practices or evaluation processes.

Explainable artificial intelligence (XAI) has emerged as a response to the “black-box” nature of many machine learning models, particularly those with high complexity such as ensemble methods or deep learning architectures (Adadi & Berrada, 2018). In educational contexts, XAI serves multiple purposes: it helps educators interpret student clustering or risk predictions, supports learners in understanding feedback generated by AI tutors, and promotes ethical accountability in the design of adaptive systems (Holstein et al., 2019).

Among the most widely used explainability techniques are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP provides global and local interpretation of feature contributions based on cooperative game theory (Lundberg & Lee, 2017), while LIME approximates the behavior of complex models by training interpretable models locally around individual predictions (Ribeiro et al., 2016). These tools have been successfully applied in domains such as healthcare, finance, and increasingly in education, where transparency is essential to build trust and fairness in algorithmic evaluations.

Despite their relevance, few studies integrate XAI techniques within broader educational approaches that also include fuzzy modeling, decision-making processes, or behavioral assessments (Casalino et al., 2024; Chimatapu et al., 2021; Mencar, 2019). Most existing models in education remain methodologically fragmented: some emphasize technology acceptance (e.g., TAM, UTAUT) without addressing cognitive engagement; others explore critical thinking or decision-making but ignore explainability or data-driven personalization.

In addition, educational models often rely on binary or linear scoring systems that oversimplify the complexity of human behavior. Very few approaches incorporate fuzzy modeling, which is better suited to capture the ambiguity inherent in constructs such as digital maturity, cognitive effort, or trust in AI-generated content (Chen et al., 2015; Herrera & Martinez, 2000b). Furthermore, while XAI methods such as SHAP and LIME are gaining traction, they are frequently applied in isolation, without being part of a larger diagnostic or assessment-oriented strategy.

These limitations highlight the importance of developing integrated approaches that, in addition to assessing user behavior, provide meaningful explanations of model results. Bridging behavioral analysis and algorithmic interpretability is essential to ensure that AI-based educational tools are transparent and pedagogically relevant.

2.6. Research Gap and Study Contribution

Although interest in AI-supported education continues to grow, the existing literature remains fragmented when it comes to connecting cognitive engagement, user profiling, and algorithmic transparency. Most current models focus either on technology acceptance (Marín Díaz et al., 2024) or on specific cognitive traits (Chimatapu et al., 2021), without bringing them together in explainable, data-driven frameworks. In addition, many cluster-

ing methods lack transparency, and only a few studies employ fuzzy logic to address the uncertainty and nuance of human behavior in learning environments (Casalino et al., 2024).

Furthermore, while explainable AI (XAI) techniques have been explored in educational settings (Karpouzis, 2024), their integration into models aimed at profiling cognitive dispositions remains limited. The absence of structured evaluation pipelines that combine cognitive constructs with interpretable modeling restricts the development of targeted pedagogical responses.

These limitations underscore the need for more integrative and interpretable approaches to support informed decision-making in AI-enhanced educational contexts. The present work addresses this gap by integrating fuzzy multicriteria analysis (AHP), unsupervised fuzzy clustering, and XAI techniques (SHAP, LIME) into a unified framework for cognitive profiling and pedagogical design.

3. Methodology

This study introduces the CRITIC-AI model, a structured methodology designed to assess digital maturity and critical thinking in the use of artificial intelligence. Given the current lack of publicly available datasets that simultaneously capture behavioral, cognitive, and decision-making aspects related to AI adoption in educational contexts, the model was validated using data collected from 1273 user profiles through a structured questionnaire. The dataset was intended to reflect authentic patterns of AI usage, cognitive engagement, and reflective practices, based on empirically grounded constructs and observed behavioral correlations.

The responses exhibit realistic behavioral variability informed by both the empirical literature and expert judgment. The dataset captures meaningful relationships among constructs such as cognitive effort, reflection without AI support, trust in AI-generated content, and information verification behaviors. This empirical foundation enabled a rigorous evaluation of the framework's internal consistency, interpretability, and practical relevance for educational contexts.

The methodology unfolds in the following phases, as shown in Figure 1:

1. **Survey and Data Design:** A structured questionnaire was designed encompassing five thematic areas: user characteristics, frequency and context of AI usage, trust and verification behaviors, cognitive effort and reflective practices, and strategies aimed at fostering critical thinking. The model was evaluated using data from 1273 user profiles collected in a controlled application setting. The sample was structured to ensure both internal consistency and behavioral diversity across cognitive and engagement dimensions. The responses reflect varying levels of digital maturity, trust in AI, and critical reflection, based on empirically defined domains and observed behavioral patterns.
2. **Multicriteria Analysis Using AHP:** The Analytic Hierarchy Process (AHP) was applied to assign weights to individual variables within each criterion, based on pairwise comparisons. The structure and comparisons were designed to reflect plausible expert reasoning and were later adjusted using insights from feature importance analysis through machine learning.
3. **Fuzzy Normalization and Linguistic Modeling:** The responses were transformed into normalized scores and then expressed using the 2-tuple linguistic representation model. This approach allows for interpretable representations of vague and gradual constructions such as “moderate reflection” or “high trust in AI”.
4. **Fuzzy Clustering and User Segmentation:** Fuzzy C-Means (FCM) was employed to identify latent user profiles based on the weighted maturity scores. The optimal number of clusters was determined using Partition Coefficient (PC) and Silhouette

- Index metrics. The resulting clusters revealed diverse patterns of AI usage and reflection, from over-reliant users to critically engaged users.
5. Validation with Supervised Learning: An XGBoost classifier was trained to predict cluster membership from the raw features, achieving high accuracy. This validation step confirmed the internal consistency of the fuzzy segmentation model and allowed for deeper analysis of discriminative features.
 6. Explainability via SHAP and LIME: Both global and local interpretability techniques were applied to understand the contribution of each feature to the cluster assignments. SHAP provided a ranked importance of variables across all users, while LIME offered localized explanations for representative individuals in each cluster.
 7. Cluster-Based Recommendations and Strategic Interpretation: Based on the profiles identified, a set of pedagogical and institutional recommendations was developed. These aim to strengthen critical thinking in each segment, for instance, by designing reflective activities for users with high automation dependency or promoting peer mentoring in profiles with strong cognitive autonomy.

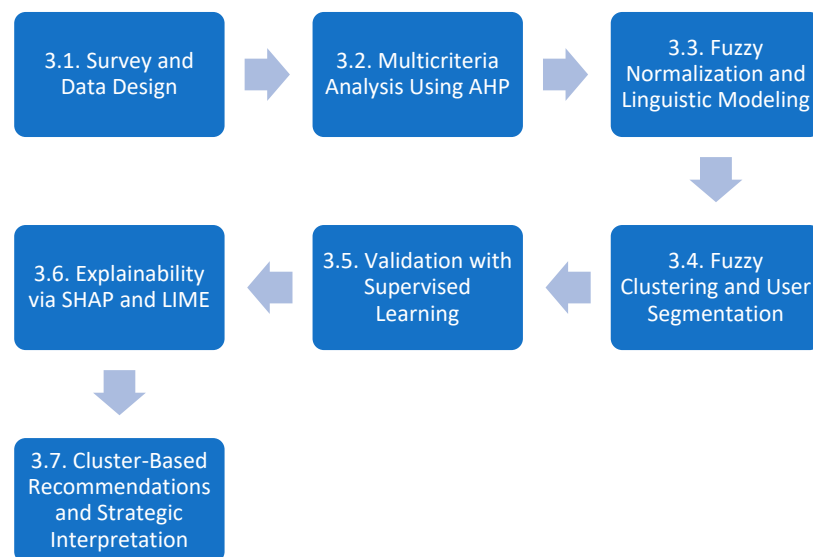


Figure 1. CRITIC-AI framework.

As shown in Figure 1, the CRITIC-AI framework is composed of seven sequential phases, each addressing a specific component of the evaluation model. These stages are described in detail in the following subsections (from Sections 3.1–3.7), ensuring a structured presentation of the full methodological pipeline.

3.1. Survey and Data Design

To validate the CRITIC-AI framework in a controlled environment, a dataset was collected through a purpose-built questionnaire. This instrument was specifically designed to assess the relationship between users' interaction with AI technologies and their levels of digital maturity, reflective capacity, and critical thinking.

Data collection took place in real-world educational settings under controlled conditions, involving undergraduate students from various academic disciplines at institutions where the researcher holds teaching responsibilities. A total of 1273 anonymized user profiles were gathered through a structured online questionnaire designed to capture behavioral and cognitive dimensions related to AI use. No personal identifiers were collected, and participation was entirely voluntary.

In accordance with national regulations on anonymous, non-interventional research, ethical approval and informed consent were not required. The study complies with the ethical principles outlined in the Declaration of Helsinki (2013 revision).

While the dataset is not publicly available due to privacy restrictions, it constitutes a robust and diverse sample for validating the proposed model. Future research will explore its application across different populations and educational contexts.

The questionnaire was structured around five core dimensions informed by the study's conceptual model and served as the foundation for subsequent analytical phases. These dimensions were derived from established models, such as the Technology Acceptance Model (TAM), metacognitive theory, and critical thinking models and previous empirical studies on user interaction with AI in educational settings. They capture essential behavioral and cognitive factors that influence how individuals adopt, evaluate, and engage critically with AI tools.

- Block 1 gathered demographic and contextual data, including age range, education level, and main context of AI use (academic, professional, or personal).
- Block 2 examined the frequency, intensity, and purposes of AI use, such as content generation, data analysis, learning support, and decision-making.
- Block 3 assessed users' trust in AI-generated content and their habits of verifying information with external sources.
- Block 4 explored cognitive engagement, including perceived effort, the tendency to accept automated responses without critical evaluation, and argumentation skills.
- Block 5 captured participants' preferences for strategies aimed at fostering critical thinking in AI-rich environments (see Table 1).

Table 1. Mapping of survey blocks and items to constructed evaluation variables.

Block	Survey Items	Description
Block 1 User Profile and Context	1.1 Age group 1.2 Educational level 1.3 Main AI use context (academic, professional, personal)	Demographic and contextual segmentation only.
Block 2 Use Frequency and Purpose	2.1 Frequency of AI use 2.2 Purpose (e.g., summarization, decision-making) 2.3 Daily usage time	Evaluates intensity and scope of AI reliance.
Block 3 Trust and Verification	3.1 Verifies information 3.2 Trust in AI outputs 3.3 Thinks less since using AI 3.4 Reflects on responses 3.5 Dependency perception	Assesses user vigilance, critical posture, and effort.
Block 4 Reflection and Reasoning	4.1 Consults AI for decisions 4.2 Evaluates alternatives 4.3 Argumentation ability 4.4 Reflects without AI	Focus on cognitive autonomy and argumentative capacity.
Block 5 Strategic Attitudes	5.1 Need to foster critical thinking 5.2 Strategies to improve analysis (e.g., debates, source checking)	Collected for interpretive purposes and final recommendations.

Only Blocks 2, 3, and 4 contributed to the Digital Maturity and Critical Thinking Index (DMCTI) through AHP-based weighting of six derived variables.

Given the lack of public datasets combining cognitive, behavioral, and attitudinal constructs related to AI engagement, we developed and administered a structured questionnaire grounded in theoretical models and expert input. This instrument yielded 1273 real user profiles, ensuring content validity and construct coherence.

To ensure the robustness of the model, response patterns were analyzed to confirm expected relationships among key variables. For example, cognitive effort, reflection without AI, information verification, and argumentation showed positive associations, frequently linked to higher levels of digital maturity. Conversely, strong trust in AI and its use for decision-making were negatively associated with critical reflection and cognitive effort, aligning with concerns about over-reliance on automated systems.

The final dataset preserves the diversity and internal consistency required for the validation of the proposed model. Its empirical richness provides a robust foundation for the subsequent analytical phases, which include multicriteria scoring, fuzzy clustering, and explainability.

3.2. Multicriteria Analysis Using AHP

Figure 2 and Table 2 present the six core criteria and their associated subcriteria derived from the survey instrument. Some survey items contribute to more than one variable, as they reflect distinct facets of the cognitive and reflective constructs being assessed.

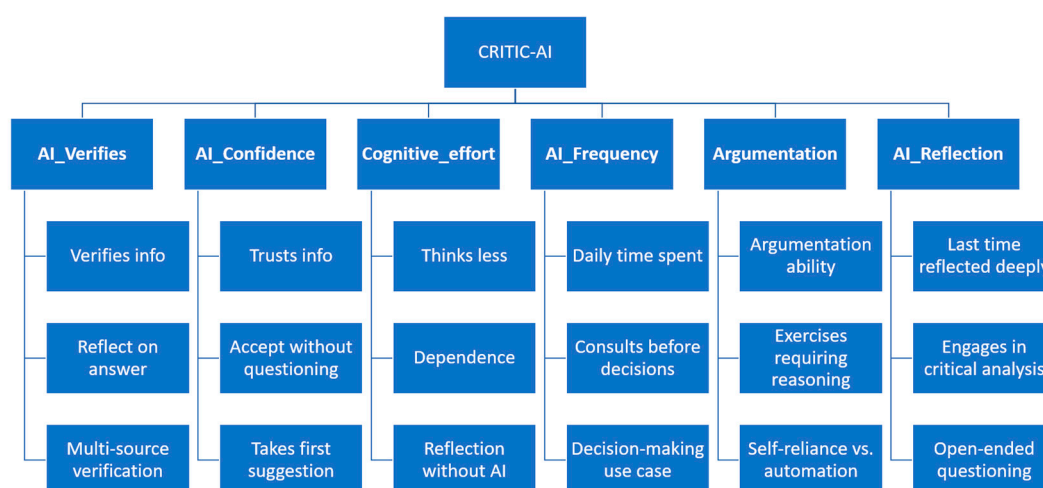


Figure 2. Hierarchical structure of the CRITIC-AI model.

Table 2. Construction of the six core variables from survey items.

Variable Name	Description	Associated Survey Items
AI_Verifies	Degree to which users verify AI-generated information using external sources.	3.1 (verifies info) 3.4 (reflect on answers) 5.2 (multi-source verification)
AI_Confidence	Level of uncritical trust placed in AI outputs.	3.2 (trusts info) 3.4 (accept without questioning) 4.2 (takes first suggestion)
Cognitive_effort	Perceived mental engagement and depth of thinking when using AI.	3.3 (thinks less) 3.5 (dependence) 4.4 (reflection without AI)

Table 2. Cont.

Variable Name	Description	Associated Survey Items
AI_Frequency	Frequency and importance of using AI to support or guide decision-making.	2.3 (daily time spent) 4.1 (consults before decisions) 2.2 (decision-making use case)
Argumentation	Ability to formulate arguments without relying on AI tools.	4.3 (argumentation ability) 5.2 (exercises requiring reasoning) 3.5 (self-reliance vs. automation)
AI_Reflection	Capacity to engage in deep thought and critical analysis independently from automated outputs	4.4 (last time reflected deeply) 3.4 (engages in critical analysis) 5.2 (open-ended questioning)

The six core variables were designed based on theoretical models in educational psychology, cognitive science, and AI literacy. For example, the construct *AI_Confidence* draws from prior work on trust in AI systems and perceived reliability of automated outputs (Marín Díaz et al., 2024; Rajki et al., 2025). *Cognitive_effort* is based on the literature on metacognitive regulation and mental engagement when interacting with external tools (Flavell, 1979; Kember et al., 2007). Similarly, *AI_Reflection* and *Argumentation* are rooted in research on critical thinking and epistemic practices (Chinn et al., 2011; Facione, 2011). These theoretical underpinnings support the internal validity of the constructs used in the segmentation process.

This structure supports the application of the AHP method to calculate individualized weighted scores for each dimension.

The Analytic Hierarchy Process (AHP) is a structured multicriteria decision-making methodology that enables the weighting of alternatives through pairwise comparisons. For a set of n criteria $C = \{c_1, c_2, \dots, c_n\}$, a reciprocal comparison matrix $A = [a_{ij}]$ is constructed, where each element a_{ij} represents the relative importance of criterion c_i over c_j , with $a_{ji} = \frac{1}{a_{ij}}$ and $a_{ii} = 1$.

$$\begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ \frac{1}{a_{12}} & 1 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{1n}} & \frac{1}{a_{2n}} & \dots & 1 \end{bmatrix} \quad (1)$$

The priority vector $w = [w_1, w_2, \dots, w_n]^T$ is calculated as the normalized principal eigenvector of the matrix A , satisfying the following:

$$w = \lambda_{max} w \quad (2)$$

where λ_{max} is the maximum eigenvalue of A . The resulting weights w_i indicate the relative importance of each criterion.

To assess the consistency of the comparisons, the Consistency Index (CI) and Consistency Ratio (CR) are computed:

$$CI = \frac{\lambda_{max} - n}{n - 1} \quad (3)$$

and

$$CR = \frac{CI}{RI} \quad (4)$$

where RI is the Random Index, an empirical constant based on matrix size n . A value of $CR < 0.10$ indicates an acceptable level of consistency in the judgments.

In this model, AHP is applied at a single level to derive weights for each set of three subcriteria linked to the six main criteria (Table 2). These criteria are then treated as independent weighted dimensions and are not aggregated into a global score.

Although a second-level AHP model could in principle be applied to compute an overall maturity index by assigning weights to the six main criteria, this step is intentionally excluded in the present study. The aim is not to produce a final aggregate score, but rather to obtain relative weights for each criterion and subcriterion. These weights are used to preserve the multidimensional structure necessary for profile-based clustering and interpretability, without reducing the evaluation to a single global measure.

3.3. Fuzzy Normalization and Linguistic Modeling

The second phase of the CRITIC-AI framework consists of transforming user responses into fuzzy linguistic representations to preserve the gradual and imprecise nature of constructs such as cognitive effort, reliance on AI, and reflection. Unlike binary or rigid scoring systems, fuzzy modeling enables nuanced evaluation, enhancing both human interpretability and mathematical formalism.

Fuzzy logic offers an essential advantage in educational modeling: it accommodates the ambiguity inherent in human behavior and cognitive states. Constructs such as “moderate reflection” or “partial trust in AI” cannot be meaningfully represented using binary or crisp categories. The 2-tuple linguistic representation allows for nuanced evaluations that align more closely with subjective human judgment, providing richer insight in contexts like digital maturity and critical thinking assessment.

To begin with, all raw response values were rescaled to the unit interval $[0, 1]$ using min–max normalization. For each input value x , the normalized score x' was calculated as follows:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

This step ensures that all variables operate on a consistent numerical domain and can be compared or aggregated across different constructs.

Subsequently, the normalized values were transformed into fuzzy linguistic representations using the 2-tuple fuzzy linguistic model proposed by (Herrera & Martinez, 2000b). This model is particularly advantageous for interpretability, as it encodes each quantitative value $x' \in [0, 1]$ as a pair (s_i, α) , where s_i is the closest linguistic label in a predefined term set, and $\alpha \in (-0.5, 0.5)$ quantifies the deviation from that label.

Let $S = \{s_0, s_1, \dots, s_g\}$ be a linguistic term set of cardinalities $g + 1$. In this study, a set of five elements was used:

$$S = \{VL, L, M, H, VH\} \quad (6)$$

The transformation from a normalized score x' to its 2-tuple representation proceeds as follows:

- The symbolic translation index is computed as, $\beta = x' \cdot g$.
- The nearest linguistic label is identified using, $i = \text{round}(\beta)$.
- The deviation from the label is obtained as, $\alpha = \beta - i$.
- The final 2-tuple is then expressed as, (s_i, α) .

This formulation guarantees the preservation of information and avoids the need for further defuzzification.

For instance, a user with a normalized cognitive effort score of $x' = 0.82$ would yield $\beta = 3.28$, $i = 3$, and $\alpha = 0.28$, resulting in the 2-tuple representation $(H, 0.28)$. This indicates a high effort with a moderate deviation toward the upper bound of that linguistic

category. In contrast, a value of $x' = 0.38$ would result in $(L, -0.12)$, indicating a level just below “Low”.

Applying this transformation across all variables, each user profile is described by a vector of interpretable fuzzy values, which facilitates downstream modeling with fuzzy logic, clustering, and explainable AI. This strategy also supports the conceptual objective of preserving human-like reasoning in educational assessment, offering a flexible and insightful alternative to conventional numeric scoring.

3.4. Fuzzy Clustering and User Segmentation

Fuzzy C-Means (FCM) clustering was selected over traditional hard clustering methods due to its ability to represent users’ partial membership across multiple cognitive-behavioral profiles. In educational contexts, learners often exhibit overlapping traits, e.g., being confident in AI use while still engaging in occasional verification, making rigid segmentation inadequate. FCM enables a more nuanced understanding of such variability by assigning membership degrees rather than binary labels, aligning closely with the fuzzy modeling of digital maturity and critical engagement.

To identify latent user segments based on their digital maturity and cognitive posture toward AI, the Fuzzy C-Means (FCM) algorithm was applied. This unsupervised clustering technique allows everyone to belong to multiple clusters with varying degrees of membership, reflecting the continuous and overlapping nature of behavioral traits in educational and digital contexts.

Let $X = \{x_1, x_2, \dots, x_n\}$ be the set of n user profiles, each represented as a vector in R^p , where $p = 6$ corresponds to the six aggregated criteria:

AI_Verifies, AI_Confidence, Cognitive_effort, AI_Frequency, Argumentation, AI_Reflection.

The objective of FCM is to minimize the following cost function:

$$J_m = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \cdot |x_i - c_j|^2 \quad (7)$$

where

c is the number of clusters; u_{ij} is the degree of membership of x_i to cluster j , satisfying $\sum_{j=1}^c u_{ij} = 1$ for all i ; c_j is the centroid of cluster j ; and $m > 1$ is the fuzzification parameter (commonly set to $m = 2$).

The algorithm iteratively updates the membership degrees and the cluster centroids using the following update rules:

- Membership update:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{|x_i - c_j|}{|x_i - c_k|} \right)^{\frac{2}{m-1}}} \quad (8)$$

- Centroid update:

$$c_j = \frac{\sum_{i=1}^n u_{ij}^m \cdot x_i}{\sum_{i=1}^n u_{ij}^m} \quad (9)$$

The process is repeated until convergence, typically when the change in the membership matrix U falls below a predefined threshold.

To determine the optimal number of clusters c , two evaluation metrics are computed:

- Partition Coefficient (PC):

$$PC = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c u_{ij}^2 \quad (10)$$

which indicates the degree of fuzziness, with higher values reflecting more compact clustering.

- Silhouette Index (S), which compares intra-cluster cohesion with inter-cluster separation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (11)$$

where $a(i)$ is the average dissimilarity of sample i to other points in the same cluster, and $b(i)$ is the minimum dissimilarity to points in a different cluster.

These metrics support the selection of a clustering configuration that balances interpretability and discriminative power, ensuring consistency in subsequent analytical stages.

3.5. Validation with Supervised Learning

To evaluate the internal consistency and predictive relevance of the fuzzy clustering results, a supervised learning approach was employed using an Extreme Gradient Boosting (XGBoost) classifier. The objective was to assess whether the soft cluster assignments obtained from Fuzzy C-Means (FCM) could be accurately predicted from the original set of features. High classification accuracy would indicate strong internal coherence of the identified segments and confirm the discriminative capacity of the features.

Let the input feature matrix be denoted as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d} \quad (12)$$

where n is the number of users, and d is the number of raw features related to AI usage and critical thinking (e.g., *AI_Confidence*, *Reflection*, etc.).

The fuzzy cluster assignment is defined by the membership matrix:

$$X = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1k} \\ u_{21} & u_{22} & \dots & u_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nk} \end{bmatrix} \in [0, 1]^{n \times k}, \sum_{j=1}^k u_{ij} = 1 \quad (13)$$

For classification purposes, each user i is assigned to the cluster with the highest membership degree:

$$y_i = \underset{j}{\operatorname{argmax}} u_{ij} \quad (14)$$

Thus, a vector of labels is obtained— $y = [y_1, y_2, \dots, y_n]^\top \in \{1, \dots, k\}^n$ —to be predicted from the features X .

The XGBoost classifier $f: \mathbb{R}^d \rightarrow \{1, \dots, k\}$ is trained to minimize the multiclass classification loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^n l(y_i, f(x_i; \theta)) + \Omega(\theta) \quad (15)$$

where $\mathcal{L}$ is SoftMax multiclass cross-entropy loss, $\Omega(\theta)$ is the regularization term to control model complexity, and θ are the parameters of the additive tree ensemble model.

The final model is an additive function:

$$f(x) = \sum_{m=1}^M \alpha_m h_m(x), \quad h_m \in \mathcal{H} \quad (16)$$

where each h_m is a regression tree from the space of possible trees $\mathcal{H}$, and α_m are the corresponding weights learned during training.

The performance of the model was evaluated using the following classification metrics:

- Confusion matrix: to visualize the distribution of true vs. predicted labels across clusters.
- Accuracy: the proportion of correctly classified users.

$$Accuracy = \frac{1}{n} \sum_{i=1}^n I(f(x_i) = y_i) \quad (17)$$

- F1-score (macro): the mean harmonic of precision and recall across all classes, balancing performance on imbalanced groups.

$$F1_{macro} = \frac{1}{k} \sum_{j=1}^k \frac{2 \cdot Precision_j \cdot Recall_j}{Precision_j + Recall_j} \quad (18)$$

- Recall (macro):

$$Recall_{macro} = \frac{1}{k} \sum_{j=1}^k \frac{TP_j}{TP_j + FN_j} \quad (19)$$

Furthermore, the XGBoost model enabled feature importance analysis through SHAP values, highlighting the most discriminative dimensions for cluster differentiation, further enriching the interpretability of the segmentation.

3.6. Explainability via SHAP and LIME

To enhance the transparency and interpretability of the classification model, both global and local explainability techniques were applied. These techniques aimed to understand how individual features influenced the prediction of cluster membership and to validate the semantic coherence of the segments defined by the fuzzy clustering.

The model $f: R^d \rightarrow \{1, \dots, k\}$, trained using XGBoost, is an ensemble of decision trees that is not inherently interpretable. To address this, two post hoc interpretability approaches were used.

3.6.1. Global Interpretability with SHAP

SHAP (SHapley Additive exPlanations) assigns a contribution score to each feature x_j for a given prediction $f(x)$, based on cooperative game theory. The SHAP value ϕ_j represents the marginal contribution of feature j to the model's output:

$$f(x) = \phi_0 + \sum_{j=1}^d \phi_j \quad (20)$$

where $\phi_0 = E[f(x)]$ is the expected prediction over all samples, and ϕ_j quantifies the effect of including feature x_j on the model output.

By aggregating SHAP values across all users, a global importance ranking of the characteristics is obtained. This ranking indicates which variables contributed most to the discrimination between clusters.

3.6.2. Local Interpretability with LIME

LIME (Local Interpretable Model-agnostic Explanations) was used to provide individualized explanations for representative users from each cluster. The approach fits a simple interpretable model g (e.g., linear regression) locally around an instance x , approximating the decision boundary of the complex model f :

$$g(x') = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (21)$$

where x' is a binary representation of x , π_x defines a local proximity kernel around x , $\mathcal{L}$ measures the fidelity of g in approximating f , and $\Omega(g)$ is a complexity penalty for g .

LIME explanations allow identification of which characteristics drove the prediction for a given individual, enabling in-depth analysis of prototype users by cluster. This allows validation of user-centered segmentation and provides practical information for adaptive educational or technological interventions.

3.7. Cluster-Based Recommendations and Strategic Interpretation

Following the identification of user profiles through fuzzy clustering and their validation via supervised learning and explainable AI methods, a strategic interpretive model was developed. This approach leverages the differentiated behavioral patterns and cognitive styles associated with each cluster to generate targeted pedagogical and institutional recommendations.

Each cluster reflects a distinct relationship between users' AI usage patterns and their critical thinking dispositions. These relationships provide a conceptual basis for personalized interventions, which aim to reinforce metacognitive engagement, reduce over-reliance on automation, and foster the development of reflective and evaluative habits.

The methodology assumes that users benefit from context-sensitive recommendations, aligned with their initial position on the spectrum of cognitive autonomy and AI reliance. Thus, cluster-specific strategies were formulated along two complementary axes:

- Cognitive axis: Emphasizing individual thought processes, such as reflection, argumentation, and verification.
- Social-institutional axis: Enabling structural support, such as mentoring systems, formative feedback, or peer exchange.

For example, users exhibiting high automation confidence but low reflection may benefit from guided activities that prompt deliberate questioning and critical reasoning. Conversely, users with strong self-regulation and cognitive effort can serve as peer mentors or contribute to reflective learning environments, reinforcing collective development of higher-order thinking skills.

This strategic layer transforms the clustering output into actionable knowledge for educational design, allowing institutions to adopt a data-driven approach to curriculum adaptation, support systems, and the promotion of critical digital literacy.

4. Understanding and Enhancing Critical Thinking in AI Users

The proposed model was developed and validated within the context of a broader research initiative aimed at analyzing the interplay between artificial intelligence usage and critical thinking among university-level students. The central research question addressed was as follows:

How can differentiated user profiles be identified based on their AI-related behaviors and cognitive dispositions, and how can this knowledge be used to design strategic pedagogical interventions?

4.1. Context

The widespread adoption of AI tools in higher education has reshaped how students' access, process, and evaluate information. Although these tools enhance efficiency and productivity, they may, in some cases, foster automation bias, reduce cognitive engagement, and inhibit the development of critical and reflective thinking. In this context, educational institutions face the challenge of balancing technological innovation with the cultivation of intellectual autonomy.

This section illustrates the application of the CRITIC-AI framework in a real-world academic setting. The analysis follows the methodological pipeline detailed in Section 3 (Figure 1), from the construction of cognitive dimensions to clustering, validation, interpretability, and strategic recommendations based on user segmentation.

4.2. Data and Dimensions

This stage corresponds to the initial phases of the methodological pipeline, Section 3.2 (Multicriteria Analysis Using AHP) and Section 3.3 (Fuzzy Normalization and Linguistic Modeling), in which the dimensions and their hierarchical structure were defined and weighted.

A structured instrument was employed to collect user data across six core dimensions aligned with the dual axis of AI interaction and cognitive engagement, as illustrated in Figure 2 and detailed in Table 2:

- *AI_Verifies*: Degree to which users independently verify AI-generated content.
- *AI_Confidence*: Level of trust placed in AI tools and recommendations.
- *Cognitive_effort*: Self-perceived mental effort during tasks involving AI.
- *AI_Frequency*: Regularity and intensity of AI tool usage.
- *Argumentation*: Tendency to justify decisions and construct reasoned arguments.
- *AI_Reflection*: Propensity to engage in metacognitive reflection after using AI.

To enhance the interpretability and decision relevance of the input variables, hierarchical structuring of dimensions and subdimensions was implemented using the Analytic Hierarchy Process (AHP). Each of the six primary dimensions was decomposed into three subcriteria, which, in real-world applications, should be evaluated through pairwise comparison matrices constructed by specialists in educational technology, pedagogy, and cognitive science, as they are best positioned to assess the relevance of each dimension in their specific institutional context.

The AHP weighting process was informed by a panel of four faculty specialists with extensive experience in AI-supported learning environments, educational technology, and cognitive science. Experts participated in pairwise comparison exercises to prioritize subcriteria within each dimension of the proposed model. Consistency Ratios were verified to ensure the reliability of the matrices, and the final weights reflected a consensus across expert judgments.

The relative importance of each subcriterion was derived from the normalized eigenvector of the comparison matrix. To ensure methodological robustness, the Consistency Ratio (CR) was calculated for each matrix. All matrices reported a $CR < 0.1$, indicating a satisfactory level of consistency in judgments. The final weights for each set of subcriteria are presented in Table 3.

Table 3. Construction and weighting of the six core variables from survey items.

Criterion	Subcriterion	Weight
AI_Verifies	Reflects_on_Answers	0.540
	Multi_Source_Verification	0.297
	Verifies_Info	0.163
AI_Confidence	Trusts_AI	0.500
	Accepts_without_Questioning	0.250
	Takes_First_Suggestion	0.250
Cognitive_effort	Reflection_without_AI	0.600
	Thinks_Less	0.200
	Dependence	0.200
AI_Frequency	Decision_Use_Case	0.600
	Daily_Time_Spent	0.200
	Consults_before_Decisions	0.200
Argumentation	Argumentation_Ability	0.333
	Exercises_Requiring_Reasoning	0.333
	Self_Reliance_vs_Automation	0.333
AI_Reflection	Last_Reflected	0.333
	Critical_Analysis	0.333
	Open_Ended_Questioning	0.333

Each criterion score was then computed as a weighted aggregation of its corresponding subcriteria. For example, within the “*AI_Verifies*” dimension, the subcriterion “*Reflects_on_Answers*” received the highest weight, as determined by the expert-based pairwise comparisons described in Section 3.3. This hierarchical integration process yielded a single interpretive score per dimension and user.

To support further analysis, all resulting scores were normalized to enable comparability across dimensions. Table 4 reports the normalized scores for five representative user profiles, obtained after applying the weights derived from the pairwise comparison matrices of the subcriteria within each main criterion, resulting in the final weighted scores for each of the six core dimensions. Table 5 provides their corresponding 2-tuple linguistic representations obtained via fuzzy modeling, enabling more interpretable and semantically rich descriptions of user profiles.

Table 4. AHP-weighted normalized scores.

ID	AI_Verifies	AI_Confidence	Cognitive_Effort	AI_Frequency	Argumentation	AI_Reflection
0001	0.580838	0.4375	0.461538	0.55	0.500000	0.545455
0002	0.353021	0.8125	0.230769	0.70	0.333333	0.363636
0003	0.080838	0.6250	0.307692	0.70	0.166667	0.272727
0004	0.625204	0.5625	0.384615	0.35	0.416667	0.545455
0005	0.735710	0.3125	0.615385	0.25	0.750000	0.909091

Table 5. Fuzzy linguistic modeling (2-tuple representation).

ID	AI_Verifies	AI_Confidence	Cognitive_Effort	AI_Frequency	Argumentation	AI_Reflection
0001	(M, 0.081)	(M, −0.062)	(M, −0.038)	(M, 0.05)	M	(M, 0.045)
0002	(L, 0.103)	(H, 0.062)	(L, −0.019)	(H, −0.05)	(L, 0.083)	(L, 0.114)
0003	(VL, 0.081)	(M, 0.125)	(L, 0.058)	(H, −0.05)	(L, −0.083)	(L, 0.023)
0004	(H, −0.125)	(M, 0.062)	(M, −0.115)	(L, 0.1)	(M, −0.083)	(M, 0.045)
0005	(H, −0.014)	(L, 0.062)	(M, 0.115)	L	(H, −0.0)	(VH, −0.091)

These variables served as input features for subsequent clustering and classification models. Their distribution across the user population is visualized in Figure 3, offering a diagnostic overview of how users are positioned along the key cognitive–AI dimensions.

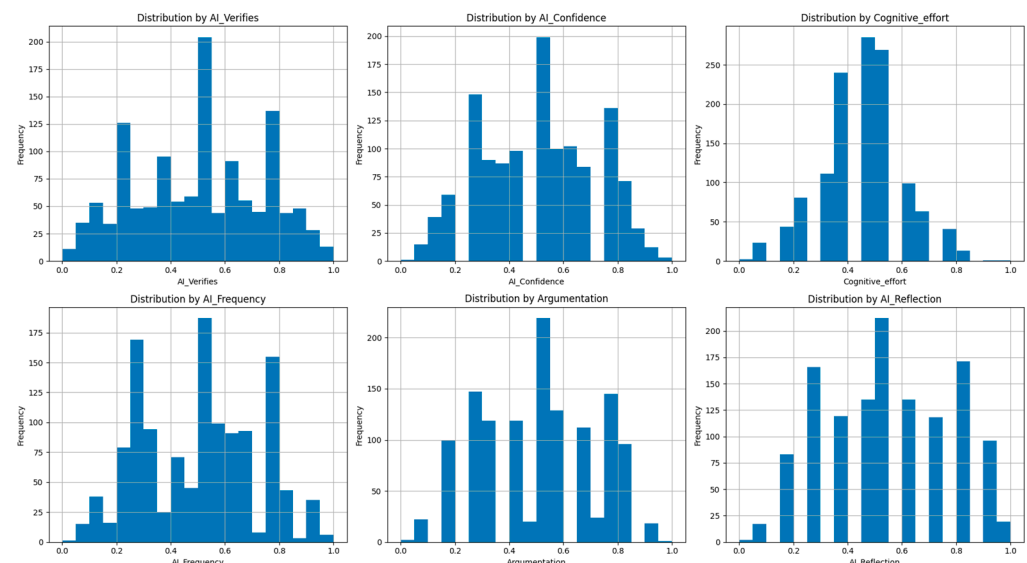


Figure 3. Distribution of the normalized scores (1273 users).

4.3. Analytical Pipeline

The analytical process followed a multi-layered pipeline that integrates both unsupervised and supervised learning methods, coupled with interpretable machine learning:

- Fuzzy C-Means clustering was employed to uncover latent user profiles with soft membership assignments, allowing users to partially belong to multiple clusters.
- XGBoost classification validated the clustering by predicting the most probable cluster for each user based on their raw feature values, demonstrating high classification accuracy.
- SHAP and LIME were used to enhance model transparency, providing both global importance rankings and local instance-level explanations of feature contributions.
- Strategic interpretation mapped the resulting clusters onto tailored pedagogical and institutional recommendations designed to reinforce critical thinking.

This methodology enabled the transformation of raw survey data into a robust and interpretable segmentation model.

4.3.1. Fuzzy C-Means Clustering

This step corresponds primarily to the fuzzy clustering phase described in Section 3.4, where users were segmented into soft membership profiles based on their cognitive and behavioral scores. Additionally, the representation of centroid values using 2-tuple linguistic descriptors draws on the fuzzy normalization process detailed in Section 3.3, enabling the interpretation of cluster characteristics in natural language terms.

To uncover latent profiles in the cognitive and behavioral dimensions of AI usage, the Fuzzy C-Means (FCM) clustering algorithm is applied. Unlike hard clustering methods (e.g., K-Means), FCM allows for soft membership: each user belongs to all clusters to a certain degree, represented by membership scores between 0 and 1. This is especially appropriate in educational contexts where behaviors often overlap and evolve over time.

To select the optimal number of clusters (c), the clustering solutions are evaluated for $c = 2$ to 7, using two complementary validation metrics:

- Partition Coefficient (PC): Measures the overall fuzziness of the partition. Higher values indicate clearer distinctions between clusters.

- Silhouette Index: Evaluates intra-cluster cohesion and inter-cluster separation. Higher values suggest better-defined clustering.

As shown in Figure 4, the PC gradually decreases with the number of clusters, while the Silhouette Index peaks at $c = 3$, suggesting that a three-cluster solution offers the best balance between clarity and representational richness. This solution was therefore adopted.

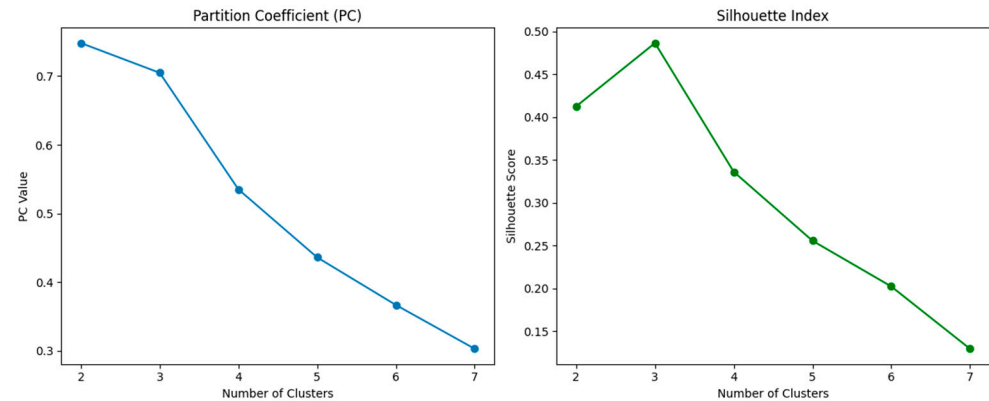


Figure 4. Partition Coefficient and Silhouette Index.

Once the fuzzy segmentation was established, the centroid values of each cluster in the six main variables were examined: “*AI_Verifies*, *AI_Confidence*, *Cognitive_effort*, *AI_Frequency*, *Argumentation*, *AI_Reflection*”.

As visualized in Figure 5, each cluster presents a unique behavioral–cognitive pattern:

- Cluster 0 exhibits average values across most dimensions, serving as a moderate baseline.
- Cluster 1 is characterized by high confidence and frequent use of AI, but lower scores in reflection and verification.
- Cluster 2 shows the opposite pattern, with high verification, argumentation, and reflection, and limited trust and frequency of AI usage.

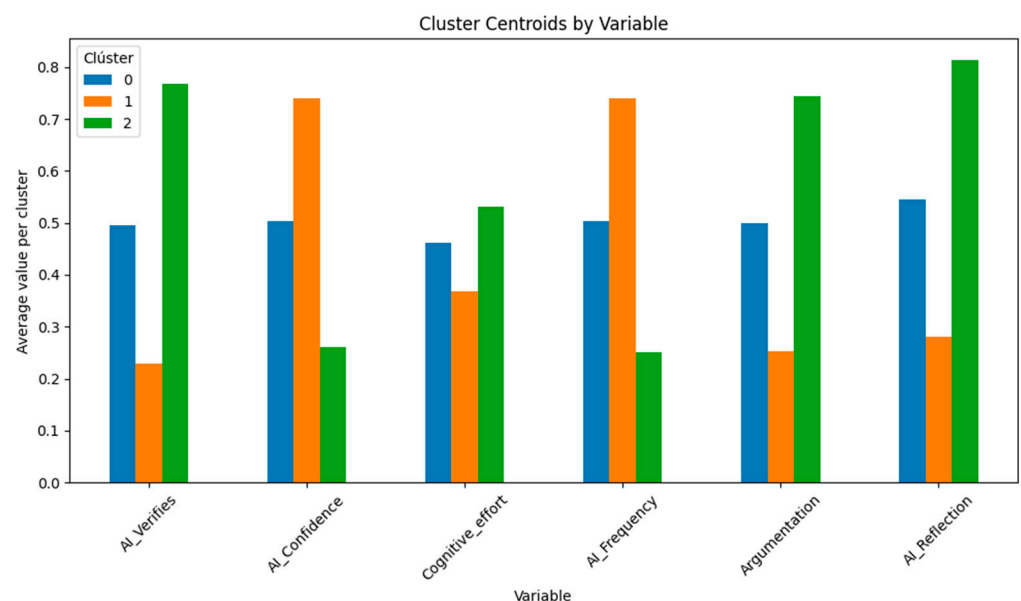


Figure 5. Cluster centroids by variable.

These centroid profiles offer a first layer of semantic interpretation, mapping the clusters to relevant pedagogical typologies.

For centroid representation, a simplified three-level scale (L, M, H) was used for clarity, derived from the original five-level term set. Each instance was described by a pair, as shown in Table 6:

- A linguistic tag: Low (L), Medium (M), or High (H).
- A semantic adjustment value: indicating the deviation from the center of the category.

Table 6. Linguistic 2-tuple representation of cluster centroids.

Cluster	AI_Verifies	AI_Confidence	Cognitive_Effort	AI_Frequency	Argumentation	AI_Reflection
0	(M, −0.004)	(M, 0.003)	(M, −0.039)	(M, 0.002)	(M, −0.0)	(M, 0.045)
1	(L, −0.021)	(H, −0.01)	(L, 0.119)	(H, −0.011)	(L, 0.002)	(L, 0.031)
2	(H, 0.017)	(L, 0.01)	(M, 0.031)	(L, 0.001)	(H, −0.007)	(H, 0.064)

This normalization supports the generation of natural language explanations and cluster-aligned recommendations based on interpretative logic.

4.3.2. XGBoost Classification

To validate the internal consistency of the fuzzy clustering and assess the discriminative power of the input features, an XGBoost classifier was trained to predict cluster membership from the original (non-clustered) feature set. This step corresponds to the supervised learning validation phase described in Section 3.5, where classification accuracy was used to confirm the robustness of the fuzzy segmentation. The dataset was split into a 70% training and 30% testing configuration, ensuring stratified representation of all clusters.

The performance of the classifier was evaluated using standard metrics: precision, recall, and F1-score, as well as overall accuracy. The results are summarized in Table 7.

Table 7. Performance metrics of the XGBoost classifier by cluster.

	Precision	Recall	F1-Score	Support
Cluster 0	0.9932	0.9932	0.9932	148.0000
Cluster 1	1.0000	1.0000	1.0000	112.0000
Cluster 2	0.9918	0.9918	0.9918	122.0000
Accuracy	0.9948	0.9948	0.99478	0.994764
Macro Avg	0.9950	0.9950	0.9950	382.0000
Weighted avg	0.9948	0.9948	0.9948	382.0000

The confusion matrix (Figure 6) shows near-perfect classification, with only two misclassifications out of 382 test instances. Such high accuracy confirms the robustness and internal validity of the fuzzy segmentation model, as the clusters derived from unsupervised learning can be reliably reproduced by a high-performing supervised model.

Given the high classification performance and the balanced distribution of instances across clusters, as evidenced by the confusion matrix, no additional techniques such as cross-validation or class balancing were applied at this stage.

These results, although exceptionally high, are consistent with the controlled nature of the dataset and the clear structure of the behavioral dimensions. Future applications to noisier, real-world settings will further test the model's robustness.

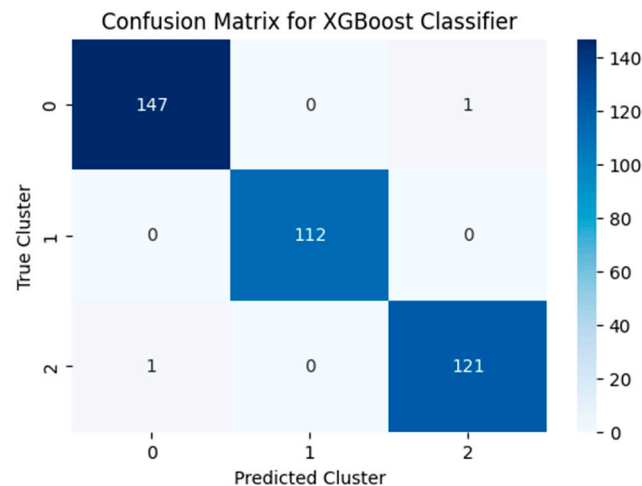


Figure 6. Confusion matrix for XGboost classifier.

4.3.3. Explainable AI (XAI)

As outlined in Section 3.6, SHAP and LIME were applied to enhance the transparency of the predictive model and to identify the most influential features in the assignment of users to clusters.

1. SHAP

To gain insight into the internal logic of the XGBoost classifier and understand which features most strongly influenced the cluster predictions, SHAP (SHapley Additive exPlanations), a model-agnostic interpretability technique grounded in cooperative game theory, was used. SHAP assigns each feature a contribution value representing its impact on the model's output.

Figure 7 shows the importance of global features across all predictions, as well as the contribution of each feature to each predicted class (cluster):

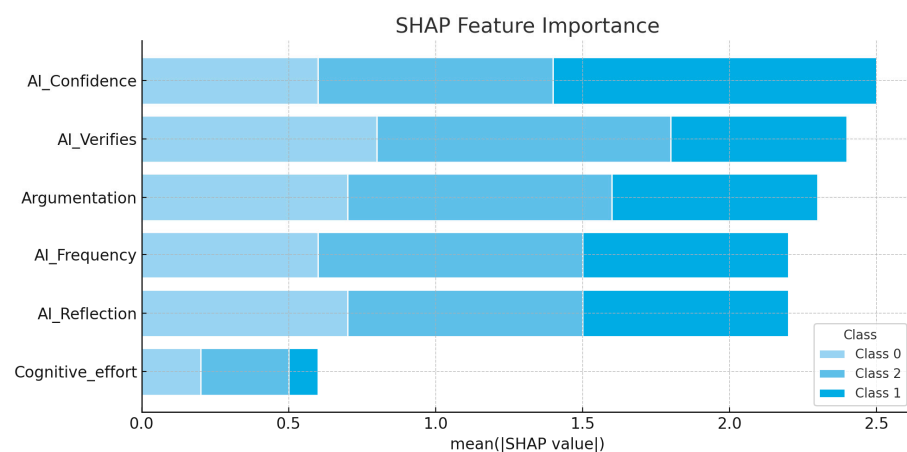


Figure 7. Global feature importance (SHAP values).

- *AI_Confidence* is the most influential feature overall, particularly in predicting Cluster 1 (users highly reliant on AI).
- *AI_Verifies*, *Argumentation*, and *AI_Reflection* are consistently relevant across all clusters, reinforcing their role in defining critical thinking profiles.
- *AI_Frequency* shows moderate influence, especially in discriminating between low- and high-automation users.

- *Cognitive_effort*, while conceptually relevant, showed the least impact on model predictions, suggesting potential redundancy or overlap with other variables.

These insights validate the interpretability and coherence of the feature set used for clustering. They also guide future refinements of the model by highlighting which dimensions carry more semantic weight in distinguishing user profiles.

2. LIME

While SHAP provided global insights into the average impact of each feature across all predictions, LIME (Local Interpretable Model-agnostic Explanations) was employed to understand the reasoning behind individual classifications. LIME perturbs the input data locally and fits an interpretable model to approximate the behavior of the black-box classifier for a specific instance.

To illustrate local explanations, three representative instances, one from each predicted cluster, were selected, as shown in Tables 8 and 9. The results, visualized in Figures 8–10, show the most influential features for each prediction.

Table 8. Feature values for representative instances from each cluster.

Instance	1193	43	128
Cluster	0	1	2
AI_Confidence	0.44	0.81	0.25
AI_Reflection	0.55	0.36	0.82
Argumentation	0.58	0.25	0.58
AI_Verifies	0.42	0.25	0.77
AI_Frequency	0.55	0.90	0.30
Cognitive_effort	0.46	0.80	0.62

Table 9. Cluster membership probabilities for representative instances.

Instance	Cluster_0_Prob	Cluster_1_Prob	Cluster_2_Prob
1193	0.899565	0.049658	0.050777
43	0.170850	0.771337	0.057813
128	0.106443	0.025738	0.867819

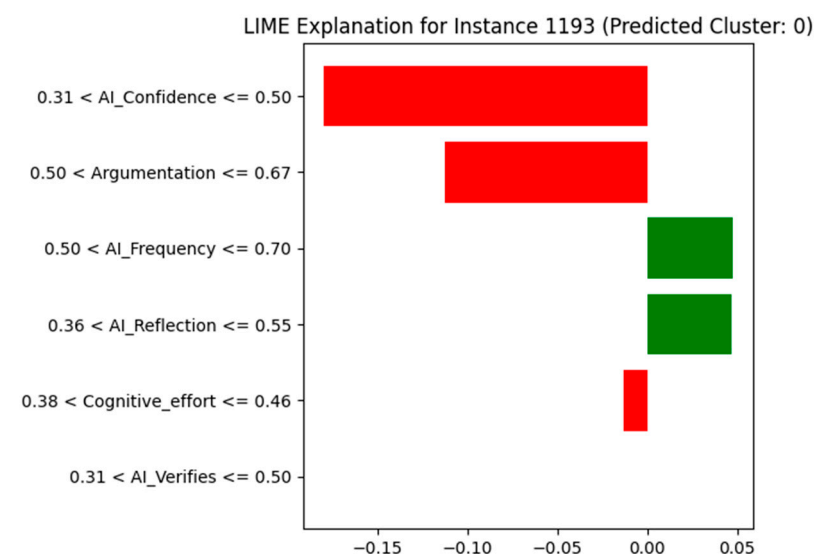


Figure 8. LIME explanation for instance 1193 (Predicted Cluster 0).

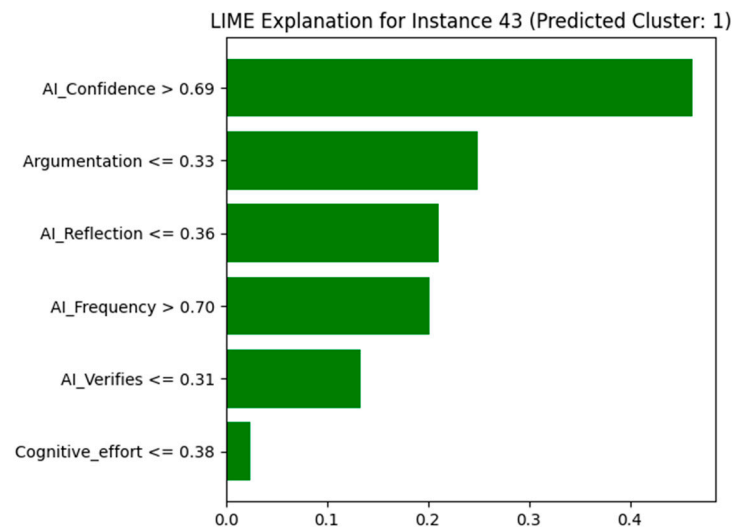


Figure 9. LIME explanation for instance 43 (Predicted Cluster 1).

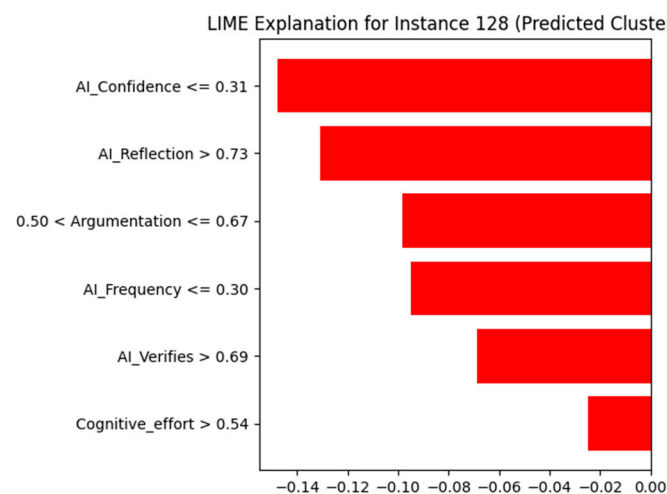


Figure 10. LIME explanation for instance 128 (Predicted Cluster 2).

These explainability results validate the fuzzy clustering structure and provide direct input for the personalized pedagogical strategies described in Section 4.4, as they highlight which traits are most salient and actionable for each user profile. In each LIME plot, green bars indicate features that contribute positively to the predicted cluster, while red bars indicate negative contributions.

1. Instance 1193 (Predicted Cluster 0):

Cluster 0 is characterized by medium-level scores across all dimensions, indicating users who interact with AI in a relatively balanced manner. These individuals neither over-rely on AI nor reject it; instead, they exhibit a moderate degree of confidence, verification behavior, reflection, and argumentation.

The LIME explanation reveals that the most influential factors for assigning instance 1193 to Cluster 0 are as follows:

- *Moderate AI_Confidence* (0.44): The value falls within the range $0.31 < x \leq 0.50$, which strongly contributes to its classification into this cluster.
- *Argumentation* (0.58): This moderate value also supports Cluster 0 membership, aligning with the centroid profile.
- *AI_Reflection* (0.55) and *AI_Frequency* (0.55): These contribute positively (green bars) to the assignment, as they match the expected moderate engagement level.

- *Cognitive_effort* (0.46) and *AI_Verifies* (0.42): These features, although contributing less, are still consistent with the balanced tendencies of Cluster 0.

2. Instance 43 (Predicted Cluster 1):

Cluster 1 represents users with high AI confidence and frequency of use, but low levels of reflection, verification, and critical reasoning. This profile represents individuals who tend to accept AI outputs quickly, with minimal cognitive engagement or questioning.

The LIME explanation shows the most influential features supporting the assignment of instance 43 to Cluster 1:

- *AI_Confidence* (0.81): This value is significantly high (>0.69) and is the most influential feature, clearly aligning with the over-reliant profile.
- *Argumentation* (0.25) and *AI_Reflection* (0.36): These low values indicate reduced critical thinking and reflective capacity, both of which contribute to the clustering.
- *AI_Frequency* (0.90): A high frequency of AI use reinforces the pattern of dependence on automation.
- *AI_Verifies* (0.25): A low score on verification implies limited cross-checking of AI outputs.
- *Cognitive_effort* (0.08): Extremely low effort suggests that the user tends to accept AI results without further elaboration.

3. Instance 128 (Predicted Cluster 2):

Cluster 2 groups individuals who show low confidence in AI outputs, but high levels of reflection, verification, and cognitive engagement. These users tend to approach AI as one of many tools, maintaining a critical stance and engaging more deeply in decision-making processes.

The LIME explanation provides insights into why instance 128 was assigned to Cluster 2:

- *AI_Confidence* (0.25): A low value (≤ 0.31) strongly supports this instance's position in Cluster 2, indicating skepticism or caution toward AI-generated suggestions.
- *AI_Reflection* (0.82): A high score (>0.73) confirms a strong reflective attitude, reinforcing this user's critical approach to information.
- *Argumentation* (0.58): A mid-to-high value within the range $0.50 < x \leq 0.67$ suggests capacity for logical reasoning and autonomy.
- *AI_Frequency* (0.30): A lower frequency of use indicates the user does not rely heavily on AI tools.
- *AI_Verifies* (0.77): A high verification score (>0.69) reveals that the user consistently cross-checks AI output with other sources.
- *Cognitive_effort* (0.62): A value above 0.54 reflects considerable cognitive involvement in tasks, aligning with independent thinking.

Each explanation aligns with the fuzzy cluster centroids and validates the semantic coherence of the segmentation model. The use of LIME enhances trust in the model's decision-making process by highlighting which behavioral patterns are most representative at the individual level.

4.4. Cluster-Based Strategic Design for Pedagogical Action

This section reflects the final phase of the CRITIC-AI framework (Section 3.7), where the clustered profiles, validated and interpreted using SHAP and LIME, are translated into targeted pedagogical and institutional strategies aimed at fostering critical thinking.

Building on the differentiated cognitive-behavioral profiles uncovered through clustering, tailored strategies are proposed to enhance responsible AI engagement and critical

thinking development for each identified group. The goal is to promote adaptive interventions that align with the users' unique cognitive orientations and AI interaction patterns.

- **Cluster 0—Moderate AI Engagement with Balanced Scores:**

Users in this cluster showed intermediate values across confidence, verification, reflection, and argumentation. Their profile suggests a transitional or exploratory stance toward AI. Strategic interventions should focus on reinforcing foundational critical thinking through low-barrier activities, such as guided reflection prompts, collaborative reasoning tasks, and contextual discussions that encourage gradual engagement with complex reasoning without overwhelming cognitive demand.

- **Cluster 1—High Confidence, High Frequency, Low Reflection:**

This group is characterized by frequent and confident use of AI tools, often accepting results with limited scrutiny. The behavioral risk lies in automation bias and cognitive shortcutting, potentially undermining analytical depth. Strategic recommendations for this cluster include metacognitive prompts before and after AI usage, structured debate sessions, and scenario-based dilemmas that require justification and argumentation beyond AI outputs. Instructional design may also integrate contrasting viewpoints and Socratic questioning to stimulate reflection.

- **Cluster 2—High Reflection, Verification, and Effort:**

These users demonstrate strong cognitive autonomy and self-regulated engagement with AI, showing high reflection, argumentation, and verification behavior. This cluster aligns with critical thinkers and independent learners. It is suggested to take advantage of this profile through peer-to-peer mentoring programs, collaborative knowledge-building roles, and involvement in co-design of AI-integrated tasks. At the institutional level, these individuals can be invited to participate in critical literacy workshops and serve as student ambassadors for responsible AI practices.

These individualized strategies aim to foster an inclusive, data-driven pedagogical ecosystem in which students are empowered to use AI, as well as question it, reflect on it, and co-construct meanings in digitally mediated learning environments. While the current study offers a concrete example of this methodology in practice, future applications may yield different profiles and actions, depending on the educational context and population. The flexibility of the proposed framework makes it suitable for iterative refinement and transfer to other domains.

5. Discussion

This study presents a novel methodological framework for understanding and improving the interaction between university students and artificial intelligence (AI) tools through multidimensional, data-driven lens. By combining fuzzy clustering, 2-tuple logic modeling, and explainable AI (XAI) techniques such as SHAP and LIME, distinct cognitive-behavioral profiles of AI usage were identified. These profiles are characterized by measurable variables such as AI confidence, usage frequency, reflective thinking, verification behavior, argumentation, and cognitive effort, which are also contextualized by interpretable models that explain individual decisions and group patterns.

The integration of the Analytic Hierarchy Process (AHP) allowed for a principled weighting of subcriteria within each main dimension, ensuring that the final profiles reflect meaningful prioritization of reflective practices, verification behavior, and engagement with AI tools. Fuzzy C-Means clustering enabled the modeling of partial cluster membership, reflecting the nuanced and non-binary nature of user behavior. XGBoost classification models were trained and validated with high performance (accuracy: 99.47%), and explainability techniques further illuminated the rationale behind the clustering process.

One of the key contributions of this study lies in the strategic interpretation of clusters for decision-making. While most existing approaches analyze AI literacy at an individual or tool-specific level, this research introduces a group-based perspective that allows instructors and institutions to design collective interventions. In this sense, the clustering results function as a decision-support system, enabling targeted pedagogical actions that enhance both AI fluency and critical thinking.

In practical terms, the insights gained offer scalable and adaptable strategies for instructional planning. By identifying groups at risk of over-reliance on automation or lacking metacognitive engagement, educational stakeholders can implement personalized scaffolding mechanisms. These interventions do not require parallel curricula but instead operate within a shared learning environment, adapting support according to cognitive behavior rather than academic level alone.

Although the empirical data used in this study were obtained from a specific educational setting, the methodological design and analytical structure follow a replicable and generalizable approach. The proposed system can thus be applied beyond case-specific context, supporting scalable interventions and comparative studies across institutions.

5.1. Pedagogical Implications and Differentiated Strategies for Each Cluster

To ensure that the results remain aligned with the methodological framework presented, it is important to clarify how the observed clusters relate to the input variables and the weighting strategy applied. Each cluster was interpreted based on five core dimensions assessed through the questionnaire: AI confidence, verification behavior, reflective thinking, argumentation, and frequency of use. The AHP-based prioritization influenced the relative impact of each dimension, highlighting verification and reflection as the most decisive variables in the overall maturity scoring.

The use of fuzzy clustering allowed us to identify non-binary behavioral patterns, where users can exhibit partial membership to more than one group. Complementary explainability techniques (SHAP and LIME) confirmed the interpretability of the clustering outcomes: for example, AI confidence and verification behavior were among the most influential factors in distinguishing automation-heavy users from critical-autonomous ones, while reflective thinking helped separate procedural users from more self-regulated profiles.

Based on this structure, the following pedagogical interventions are proposed as contextual examples, aligned with the behavioral traits found in each group. These strategies are not intended as fixed prescriptions, but as illustrative outputs of the methodological pipeline applied in this case study. Different populations or datasets may yield different results, and accordingly, different interventions. This flexibility is one of the strengths of the proposed approach, which can be embedded into multicriteria, multi-expert decision-making systems using AHP and similar frameworks.

1. Cluster 0—Moderate AI Confidence, Moderate Reflection (Procedural Users)

This group exhibits a mechanical, non-critical use of AI, with mid-range scores in most dimensions and a tendency to rely on AI outputs without deeper analysis. Suggested strategies include the following:

- Reflective prompt interventions.
- Double-output comparison tasks.
- Visual critical thinking scaffolds.
- Error-awareness sessions.
- Micro-workshop: “Detecting AI Errors: A Critical Thinking Starter Kit”.

These users can benefit from prompts embedded in LMS platforms that require justification before accepting AI responses, thus promoting reflective reasoning. These prompts could be automatically implemented in formative quizzes or learning analytics dashboards.

2. Cluster 1—High Confidence, High Frequency, Low Reflection (Automation-Heavy Users)

This profile is marked by over-reliance on AI, strong trust in its outputs, and minimal verification or reasoning. Suggested strategies include the following:

- Cognitive disruption exercises.
- Forced deviation tasks.
- Bias simulation challenges.
- “AI Detox” activities.
- Micro-workshop: “Break the Pattern: Thinking Beyond Automation”.

These users could be detected at an early stage of academic programs and guided through mandatory ‘critical engagement with AI’ modules, where students are challenged to question AI-generated results and reflect on cognitive shortcuts.

3. Cluster 2—High Reflection, High Verification, Low AI Frequency (Critical-Autonomous Users)

This group demonstrates mature, responsible use of AI and a strong tendency to verify information and think critically. Suggested strategies include the following:

- Peer mentoring roles.
- Ethical prompt design.
- Bias detection tasks.
- Meta-analytical assignments.
- Micro-workshop: “Mentoring with AI: Leading Ethical and Critical Dialogue”.

Since these users show high autonomy and verifying behavior, they could be formally invited to act as peer mentors in AI-related classroom activities, such as assessing the quality of outputs from generative tools or leading group discussions on the ethical use of AI.

5.2. Future Directions and Limitations

Although this study offers a sound methodological approach and practical insights, several avenues for future research remain open.

First, the incorporation of temporal dynamics would allow us to explore how users may evolve over time, analyzing the temporal evolution of the development of critical thinking. Second, longitudinal intervention studies could be conducted to validate the proposed strategies and evaluate their sustained impact by carrying out practices that motivate reflection and criticality.

It is important to emphasize that the present work does not attempt to define a universal or prescriptive model. Instead, it proposes a flexible methodological framework that integrates fuzzy clustering, AHP-based prioritization, and explainable AI to identify cognitive-behavioral patterns and derive context-sensitive strategies. These strategies are derived from a specific sample and application context and should be understood as illustrative process outcomes, not normative prescriptions.

Applying the methodology to datasets from a variety of educational and professional settings would reinforce its practical relevance and reveal different user profiles and pedagogical priorities. In addition, the proposed process could be extended to compare the reflective and cognitive impact of different AI tools (e.g., generative, assistive, analytic), enabling more nuanced pedagogical decisions.

In practical terms, ideally, the AHP weighting process would incorporate input from expert panels (e.g., faculty, program coordinators, pedagogical teams), whose criteria will vary depending on institutional goals and values. This step, already considered for future implementations, would allow for context-specific prioritization, aligning the framework with local realities and improving its decision-making capacity.

In addition, a second-level AHP model could be introduced that would allow aggregate maturity scores to be calculated for benchmarking purposes, although this was excluded from the current design to have profile-based segmentation and develop interpretability.

As with any expert-based model, the AHP approach is subject to variability based on subjective input in pairwise comparisons. Although consistency coefficients provide internal checks, panel diversity and sensitivity analysis should be applied in future versions to increase reliability.

Although Fuzzy C-Means (FCM) was selected for its ability to model smooth cognitive transitions, future studies could explore alternative algorithms such as K-Means or Gaussian Mixture Models (GMMs) to assess robustness and consistency between methods. In the future, benchmarking could be performed.

Future adaptations of the methodology could also incorporate ethical and social dimensions, such as user awareness of fairness, bias, or responsible use of AI in decision-making contexts. Longitudinal studies and follow-up designs would make it possible to assess the long-term impact on the development of critical thinking.

Although this study focuses on higher education, the methodology is adaptable to other settings, such as professional development, healthcare, administrative decision-making, or even K12 education, provided the instruments are adjusted to developmental stages and supported by teacher mediation, where human–AI interactions require reflective judgment and ethical alignment.

Finally, extending the application to datasets from different institutions, countries, or academic disciplines will enable validation of this approach in heterogeneous contexts, contributing to the construction of more inclusive and adaptive pedagogical systems.

Although the participants in this study were drawn from a convenience sample of higher education institutions known to the author, this choice served the purpose of illustrating the applicability of the methodology using real, context-bound data. No claim is made regarding the representativeness of the resulting profiles, which are intended solely to demonstrate the framework's capacity to produce actionable insights under defined conditions.

6. Conclusions

Although the present study offers a solid methodological foundation and actionable strategies for instructional adaptation, it is important to acknowledge its limitations. The data analyzed in this research were collected from users through a structured questionnaire, reflecting real behavioral patterns collected in educational settings, and aligned with validated theoretical constructs. Consequently, although the clustering profiles and resulting recommendations are theoretically sound and supported by machine learning evidence, their applicability across diverse educational populations requires further validation using larger and more heterogeneous real-world datasets.

Even so, the approach presented here demonstrates both flexibility and adaptability, making it suitable for implementation across varied educational settings once reliable data becomes available. What makes the model valuable goes beyond its technical rigor and is its conceptual coherence. The system provides institutions and educators with the capacity

to tailor their pedagogical responses based on students' cognitive engagement profiles with AI.

Although the intervention strategies proposed for each user profile may appear general, they are deliberately formulated to serve as a flexible pedagogical foundation. Their specificity and operationalization should be developed collaboratively with educators, considering course level, institutional goals, and learner needs. The aim is not to prescribe fixed solutions, but to support adaptive design processes that align with critical thinking development and responsible AI integration in each unique educational context.

However, a deeper challenge remains, extending beyond data precision or algorithmic optimization. It lies in our collective readiness, as educators and learners, to harness the true transformative power of artificial intelligence in education. We live in an era where immediate access to vast information and automated reasoning is the norm. But the real question is the following: how do we transform this abundance into meaningful cognitive development?

The answer lies in deliberately crafted educational approaches that position AI not as a shortcut to learning, but as a companion in cognitive growth. Students should be guided not merely on how to operate AI tools efficiently, but on how to interact with them reflectively, ethically, and critically. Likewise, the role of educators must evolve from knowledge providers to catalysts of intellectual curiosity and development, helping learners construct understanding, ask deeper questions, and build personal meaning.

From an institutional perspective, the approach can serve as a diagnostic tool to identify areas where students may require additional support in developing critical thinking and responsible AI usage. Academic departments could use the maturity profiles to guide targeted interventions, such as embedding reflective tasks in AI-supported coursework or designing training modules for faculty on how to scaffold AI literacy. At the policy level, the model offers a scalable foundation for monitoring cognitive engagement trends across programs, aligning digital transformation strategies with ethical and pedagogical goals.

Ultimately, what this study offers is not a fixed or generalizable model, but a scalable methodological framework that can be adapted to the specific needs of different educational contexts. By combining fuzzy clustering, AHP-based decision-making, and explainability techniques, the framework enables institutions to derive meaningful insights from their own data and translate them into tailored pedagogical actions. The strategies and profiles presented here are illustrative of what the method can produce in a particular context; future applications may yield different outcomes, depending on the criteria, population, and expert input involved. This adaptability is one of the key strengths of the approach.

In this context, this work contributes to a broader educational mission: shift from passive AI usage to a paradigm of constructive cognitive co-evolution, where artificial tools augment rather than replace human intelligence. The strategies proposed here offer a foundation for designing adaptive learning ecosystems that foster this evolution. In such environments, both students and educators thrive, cultivating digital competence while developing the capacity to generate value, insight, and critical thought in an age defined by technological transformation.

This study consolidates fuzzy logic, multicriteria decision-making, clustering, and explainable AI into a coherent model for behavioral–cognitive profiling.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to its anonymous and non-interventional nature, in accordance with Spanish national regulations on educational research and data protection (Organic Law 3/2018 on the Protection of Personal Data and Guarantee of Digital Rights).

Informed Consent Statement: Informed consent was not required as the study involved no personal or sensitive data and participation was fully anonymous and voluntary.

Data Availability Statement: The dataset used in this study was fully anonymized and collected for academic research purposes within an educational setting. Due to privacy considerations and the absence of explicit consent for public sharing, the data cannot be made publicly available. All data processing was conducted in accordance with GDPR-compliant practices. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The author declare no conflicts of interest.

References

- Abdulmunem, R. A. (2023). Artificial intelligence in education. In *Comparative research on diversity in virtual learning: Eastern vs. western perspectives* (pp. 241–255). IGI Global Scientific Publishing. [\[CrossRef\]](#)
- Adadi, A., & Berrada, M. (2018). Inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. [\[CrossRef\]](#)
- Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017, May 14–19). *Machine learning approaches to predict learning outcomes in Massive open online courses*. 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 713–720), Anchorage, AK, USA. [\[CrossRef\]](#)
- Andreas, C. (2012). UTAUT and UTAUT 2: A review and agenda for future research. *The Winners*, 13(2), 106–114.
- Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2), 191–203. [\[CrossRef\]](#)
- Boeren, E. (2019). Understanding Sustainable Development Goal (SDG) 4 on “quality education” from micro, meso and macro perspectives. *International Review of Education*, 65(2), 277–294. [\[CrossRef\]](#)
- Casalino, G., Castellano, G., Di Mitri, D., Kaczmarek-Majer, K., & Zaza, G. (2024). A Human-centric approach to explain evolving data: A case study on education. In J. A. I. Martinez, R. D. Baruah, D. Kangin, & P. V. D. Souza (Eds.), *IEEE conference on evolving and adaptive intelligent systems 2024, IEEE EAIS 2024* (pp. 208–215). IEEE. [\[CrossRef\]](#)
- Cernega, A., Nicolescu, D. N., Meleşcanu Imre, M., Ripszky Totan, A., Arsene, A. L., Şerban, R. S., Perpelea, A. C., Nedea, M. I., & Pişuru, S. M. (2024). Volatility, Uncertainty, Complexity, and Ambiguity (VUCA) in healthcare. *Healthcare*, 12(7), 773. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, J.-F., Hsieh, H.-N., & Do, Q. H. (2015). Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach. *Applied Soft Computing*, 28, 100–108. [\[CrossRef\]](#)
- Chimatapu, R., Hagraas, H., Kern, M., & Owusu, G. (2021, July 11–14). *Enhanced deep type-2 fuzzy logic system for global interpretability*. 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1–8), Luxembourg. [\[CrossRef\]](#)
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. L. A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. *Educational Psychologist*, 46(3), 141–167. [\[CrossRef\]](#)
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 22. [\[CrossRef\]](#)
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*, 13(3), 319–339. [\[CrossRef\]](#)
- Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991–16005. [\[CrossRef\]](#)
- Facione, P. A. (2011). *Critical thinking: What it is and why it counts* (pp. 1–28). Insight Assessment. ISBN 13: 978-1-891557-07-1.
- Favero, L., Pérez-Ortiz, J. A., Käser, T., & Oliver, N. (2024). Enhancing critical thinking in education by means of a socratic chatbot. *arXiv*. [\[CrossRef\]](#)
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911. [\[CrossRef\]](#)
- Herrera, F., & Martinez, L. (2000a). An approach for combining linguistic and numerical information based on the 2-tuple fuzzy linguistic representation model in decision-making. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(5), 539–562. [\[CrossRef\]](#)
- Herrera, F., & Martinez, L. (2000b). A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, 8(6), 746–752. [\[CrossRef\]](#)
- Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Conference on Human Factors in Computing Systems—Proceedings*. Association for Computing Machinery. [\[CrossRef\]](#)

- Hutson, J., Jeevanjee, T., Graaf, V. V., Lively, J., Weber, J., Weir, G., Arnone, K., Carnes, G., Vosevich, K., Plate, D., Leary, M., & Edele, S. (2022). Artificial Intelligence and the Disruption of Higher Education: Strategies for Integrations across Disciplines. *Creative Education*, 13(12), 3953–3980. [CrossRef]
- Karpouzis, K. (2024). *Explainable AI for Intelligent Tutoring Systems* (pp. 59–70). Springer. [CrossRef]
- Kember, D., Leung, D. Y. P., & Ma, R. S. F. (2007). Characterizing learning environments capable of nurturing generic capabilities in higher education. *Research in Higher Education*, 48(5), 609–632. [CrossRef]
- Lin, Y., & Yu, Z. (2023). Extending technology acceptance model to higher-education students' use of digital academic reading tools on computers. *International Journal of Educational Technology in Higher Education*, 20(1), 34. [CrossRef]
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates Inc. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (accessed on 30 April 2025).
- Marín Díaz, G., Galán Hernández, J. J., Gómez Medina, R., & Aijón Jiménez, J. A. (2024). *Understanding AI adoption in education: A TAM perspective on students' and teachers' perceptions*. Dykinson, S.L. [CrossRef]
- Marín Díaz, G., Galdón Salvador, J. L., & Galán Hernández, J. J. (2023). Smart cities and citizen adoption: Exploring tourist digital maturity for personalizing recommendations. *Electronics*, 12(16), 3395. [CrossRef]
- Mencar, C. (2019). Looking at the branches and roots. In R. Fullér, S. Giove, & F. Masulli (Eds.), *Fuzzy logic and applications* (pp. 249–252). Springer International Publishing.
- Mintz, J., Holmes, W., Liu, L., & Perez-Ortiz, M. (2023). Artificial intelligence and K-12 education: Possibilities, pedagogies and risks. *Computers in the Schools*, 40(4), 325–333. [CrossRef]
- Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable* (p. 247) Self-published. Available online: <https://christophm.github.io/interpretable-ml-book> (accessed on 21 April 2025).
- Qian, M., & Clark, K. R. (2016). Game-based Learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, 63, 50–58. [CrossRef]
- Queiroga, E. M., Lopes, J. L., Kappel, K., Aguiar, M., Araújo, R. M., Munoz, R., Villarroel, R., & Cechinel, C. (2020). A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences*, 10(11), 3998. [CrossRef]
- Rajki, Z., Dringó-Horváth, I., & Nagy, J. T. (2025). Artificial intelligence in higher education: Students' artificial intelligence use and its influencing factors. *Journal of University Teaching and Learning Practice*, 22(2). [CrossRef]
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv*. [CrossRef]
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. [CrossRef] [PubMed]
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601–618. [CrossRef]
- Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3), 234–281. [CrossRef]
- Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *Journal of Manufacturing Technology Management*, 26(6), 791–806. [CrossRef]
- Salomon, G., & Perkins, D. N. (1998). Individual and social aspects of learning. *Review of Research in Education*, 23, 1–24. [CrossRef]
- Sottolare, R. A., Goldberg, B., & Hu, X. (2014). *Design recommendations for intelligent tutoring systems—Volume 2: Instructional management* (Vol. 2, Issue July 2015). US Army Research Laboratory.
- UNESCO. (2023). *Digital learning and transformation of education*. Available online: <https://www.unesco.org/en/digital-education> (accessed on 30 April 2025).
- Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions subject areas: Design characteristics, interventions. *Decision Sciences*, 39(2), 273–315. [CrossRef]
- Venkatesh, V., & Davis, F. D. (2000). Theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186–204. [CrossRef]
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. [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.