

2026-02-16

## Educators' Evaluation of Explanation Types in XAI for Higher Education Dropout

Henry Maathuis

*Hu University of Applied Sciences*

Jan Glazenborg

*Hu University of Applied Sciences*

Meike Grol

*Hu University of Applied Sciences*

*See next page for additional authors*

Follow this and additional works at: <https://arrow.tudublin.ie/faccommetro>

---

### Recommended Citation

Maathuis, H., Glazenborg, J., Grol, M., & Quille, K. (2026) 'Educators' Evaluation of Explanation Types in XAI for Higher Education Dropout', *HCAI-ep 2026 - Proceedings of the 2026 Conference on Human Centered Artificial Intelligence - Education and Practice*, , pp. 61-67. Association for Computing Machinery (ACM): Available at: [10.1145/3777490.3777499](https://doi.org/10.1145/3777490.3777499)

This Article is brought to you for free and open access by the School of Enterprise Computing and Digital Transformation at ARROW@TU Dublin. It has been accepted for inclusion in Research Outputs: 2025-Present by an authorized administrator of ARROW@TU Dublin. For more information, please contact [arrow.admin@tudublin.ie](mailto:arrow.admin@tudublin.ie).



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

---

## Authors

Henry Maathuis, Jan Glazenborg, Meike Grol, and Keith Quille

# Educators' Evaluation of Explanation Types in XAI for Higher Education Dropout

Henry Maathuis\*

HU University of Applied Sciences Utrecht  
Utrecht, Netherlands  
henry.maathuis@hu.nl

Meike Grol

HU University of Applied Sciences Utrecht  
Utrecht, Netherlands  
meike.grol@hu.nl

Jan Glazenborg\*

HU University of Applied Sciences Utrecht  
Utrecht, Netherlands  
jan.glazenborg@student.hu.nl

Keith Quille

Technological University Dublin  
Dublin, Ireland  
keith.quille@TUDublin.ie

## Abstract

First-year student dropout rates represent a global challenge, particularly in Computer Science programs. In response, the Predict Student Success (PreSS) tool was developed over two decades using Naïve Bayes, a black box machine learning model, to identify students at risk of early withdrawal or academic failure. Given growing concerns about the use of black box models in high-stakes domains such as education, this study builds on former research by comparing the Naïve Bayes model to three glass box AI models: Decision Tree, Explainable Boosting Machine, and Automatic Piecewise Linear Regression. These models were used to generate four types of explainable AI visualisations: Feature Importance, Similar and Contrastive Examples, Decision Rules, and Counterfactuals. Educators evaluated these visualisations through a qualitative survey. Results indicate that glass box models can perform competitively with Naïve Bayes, and that contrastive explanation types were most preferred among educators. Open-ended responses underscore the importance of clear user guidance when implementing such tools.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in visualization**; • **Computing methodologies** → **Machine learning algorithms**.

## Keywords

Student Success Prediction, Human-Centered Explainable AI, User Perception, Higher Education Dropout

## ACM Reference Format:

Henry Maathuis, Jan Glazenborg, Meike Grol, and Keith Quille. 2026. Educators' Evaluation of Explanation Types in XAI for Higher Education Dropout. In *Human Centred Artificial Intelligence - Education and Practice (HCAI-EP '26)*, January 21–22, 2026, Kildare, Ireland. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3777490.3777499>

\*Both authors contributed equally to this research and share first authorship.



This work is licensed under a Creative Commons Attribution 4.0 International License. *HCAI-EP '26, Kildare, Ireland*

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2153-3/26/01  
<https://doi.org/10.1145/3777490.3777499>

## 1 Introduction

The first year of Higher Education (HE) is a pivotal period for students, who must adjust to a new subject, environment, and academic culture while often still navigating adolescent development. It is therefore not surprising that first-year dropout rates are high across the globe. In the Netherlands, 12.4% of HE students discontinue their studies after the first year [5], while in Ireland the figure is approximately 15% [7]. Germany, Spain, and the United States all report dropout rates above 10% [11, 12, 28]. Such outcomes are costly, both for students and society: in the Netherlands, for instance, those with only a secondary school diploma earn on average €14,000 less per year than graduates with a HE degree.

The challenge is particularly acute in Computer Science (CS) programs. Compared to other disciplines, CS exhibits consistently higher dropout rates. In the United Kingdom and United States, CS has the highest dropout rates among all degree programs, at 9.8% and 10.7% respectively [8, 32]. In the Netherlands, after five years, only 41% of HBO-ICT (college-level CS) students graduate, compared to the national average of 52% [8, 32]. This is especially concerning given the European Commission's warnings of an expanding shortage of IT specialists across the EU [3]. Contributing factors include limited prior exposure to programming, misaligned expectations, and time constraints [27].

To mitigate these risks, predictive analytics tools have been developed to help identify at-risk students early in their studies. One such example is the Predict Student Success (PreSS) tool, developed in Ireland since 2006 [23–25]. PreSS employs artificial intelligence (AI) to predict whether students in introductory programming courses (CS1) will eventually pass or fail, enabling educators to target support towards students who need it most. However, while predictive accuracy has been the primary focus of PreSS and similar tools, questions remain regarding the interpretability of their outputs and the extent to which educators can meaningfully integrate such predictions into their teaching practices.

This paper addresses that gap. We investigate how different types of explainable machine learning models can support educators in understanding and acting upon predictive insights. To this end, we ask: (1) how does the existing PreSS model compare with several inherently interpretable “glass box” models in terms of predictive performance, and (2) how do computer science educators perceive and evaluate the quality of explanations generated by these models? By doing so, we aim to contribute not only to the technical

development of predictive models in education, but also to the broader discussion on the role of explainability in high-stakes AI applications.

## 2 Related Work

Artificial intelligence and machine learning have long been applied in education, from adaptive tutoring systems to early warning tools. Given their potential impact on learners, the EU AI Act designates educational algorithms as high risk [4], reinforcing the importance of transparency from the outset. Yet the opacity of many models poses challenges when educators must rely on their outputs to guide interventions. A growing body of research stresses that the starting point should always be interpretable or “glass-box” models rather than black-box approaches such as Random Forests, Support Vector Machines, or XGBoost [10]. Rudin [26] argues that in high-stakes domains, interpretable models should be the default, since in many structured tasks they can perform comparably to black-box methods. In either case, interpretable approaches such as decision trees or simple regression models provide transparency that enables users to scrutinize and trust predictions. Only when interpretable models are shown to be inadequate should black-box methods be considered.

In the context of black-box models, explainable AI (XAI) methods provide a way to achieve some degree of interpretability. These approaches are commonly divided into *ante-hoc* methods, which are interpretable by design, and *post-hoc* techniques, which aim to explain black-box models after training. Ante-hoc methods include decision trees, linear models, and scoring systems, whose structure makes them inherently understandable [20, 26]. More recent innovations extend this family, such as Explainable Boosting Machines (EBMs)[22], which build on Generalized Additive Models [9] to achieve state-of-the-art accuracy on tabular data [13], and Automatic Piecewise Linear Regression (APLR) [30], which captures nonlinear relationships while remaining interpretable. Post-hoc methods, in contrast, seek to approximate and explain complex models retrospectively. Examples include SHAP, which attributes feature importance using Shapley values[18]; counterfactual explanations, which show how small changes to inputs might alter predictions [29, 31]; and rule-based explanations, which approximate behavior with logical rules [20]. These techniques can be informative, but because they are approximations, their reliability and clarity — particularly for non-experts such as educators — remain limited [17].

In the educational domain, most research has approached explanations from a primarily technical perspective, focusing on objective and quantitative measures such as completeness, stability, or compactness of explanations [21]. Far less attention has been devoted to the human-centered dimension: how the intended users — educators — actually perceive, interpret, and act upon explanations in practice [1, 14]. This distinction is crucial. Technical interpretability ensures that an explanation is faithful to the model, but without careful consideration of user experience, such explanations may fail to be meaningful or actionable. Poorly designed interfaces can overwhelm or mislead users, whereas thoughtfully designed visualizations can translate algorithmic reasoning into insights that educators can readily use [15, 20]. Recent research

emphasises that the effectiveness of explanations depends not only on their technical rigour but equally on their accessibility, usability, and alignment with human sense-making [6, 16]. In this work, we focus on the human-centered dimension of explainability, developing explanation designs through iterative prototyping and educator evaluation. We emphasise qualities such as ease of understanding, user satisfaction, and practical actionability.

## 3 Methodology

This study consists of two complementary parts. First, a technical comparison was conducted between the current PreSS model (Naïve Bayes) and three glass box machine learning models. The outputs of these models were then used to generate four types of explanations. Second, these explanations were evaluated in a survey among computer science educators, with the aim of assessing their perceived quality.

### 3.1 Dataset and Technical Comparison

For the technical comparison, we employed the PreSS dataset collected by Quille et al. [23], which comprises survey responses from 692 students across 11 institutions (10 Irish and 1 Danish). The survey captured information on students’ programming and mathematics backgrounds, gaming habits, and included a short programming test administered 4–6 hours (approximately 10%) into their CS1 course.

Demographic variables include institution type (university, institute of technology, or community college), age (17–60 years), gender<sup>1</sup> and mature student status (classified as 23 years or older). Behavioural measures capture average daily use of social media, weekly hours in part-time employment, and gaming activity both before and during the course. The academic background is represented by the student’s prior mathematics grade (normalised across exam levels) and their self-reported expected end-of-year result.

The dataset also incorporates several motivational and psychological constructs based on established instruments. Intrinsic Goal Orientation (IGO) [2] reflects a learner’s curiosity, desire for challenge, and ambition to master a task for its own sake, rather than simply aiming to complete it or get by. Intrinsic Questioning categorises students into three levels based on their IGO scores. Control of Learning Beliefs measures the perceived link between effort and outcome. Test Anxiety reflects self-reported stress levels in assessment situations. Finally, performance on the short programming test is included as an ordinal variable. These factors are summarised in Table 1, which lists each variable alongside its type and observed range.

We then compared four machine learning models — Explainable Boosting Machines (EBM), Naïve Bayes (NB), Automatic Piecewise Linear Regression (APLR), and Decision Trees (DT) — as a means of generating explanations. All models were trained using identical preprocessing, feature selection, and evaluation procedures. Hyperparameters were tuned through grid search, with all other settings left at their defaults (see Table 2).

<sup>1</sup>We acknowledge that gender should not typically be modelled as a binary construct; here we report it as such because the original dataset only recorded male/female responses.

| Factor                    | Type        | Range        |
|---------------------------|-------------|--------------|
| Institution Type          | Categorical | 3 categories |
| Survey Time (s)           | Continuous  | 132–1235     |
| Age (years)               | Discrete    | 17–60        |
| Mature Student            | Binary      | 0–1          |
| Gender                    | Binary      | 0–1          |
| Social Media (h/day)      | Discrete    | 0–23         |
| Part-Time Job (h/week)    | Discrete    | 0–23         |
| Expected Result (%)       | Discrete    | 0–100        |
| Programming Task (Likert) | Ordinal     | 1–5          |
| Goal Orientation (IGO)    | Ordinal     | 1–5          |
| Intrinsic Questioning     | Categorical | 3 categories |
| Learning Beliefs          | Ordinal     | 1–5          |
| Test Anxiety              | Ordinal     | 1–5          |
| Math Grade                | Discrete    | 1–12         |
| Gaming During (h/day)     | Discrete    | 0–10         |
| Gaming Before (h/day)     | Discrete    | 0–23         |

**Table 1: Student factors in the dataset with data types and observed ranges.**

#### APLR

Tuning: boosting steps {1000, 3000}, learning rate {0.05, 0.1, 0.2}, interaction depth {0, 1}, ridge penalty {0,  $10^{-4}$ }, linear-only steps {0, 1000}, non-linearity penalty {0, 0.5}, interaction penalty {0, 0.5}

Best: 1000 boosting steps, learning rate 0.1, interaction depth 1, ridge penalty 0, no linear-only steps, non-linearity penalty 0, interaction penalty 0.5

#### EBM

Tuning: feature bins {128, 256}, interaction bins {32, 64}, interactions {0, 5}, validation split {0.1, 0.15, 0.2}, early stopping {50, 100}, learning rate {0.01, 0.015, 0.02}, boosting rounds {1000, 5000}, min leaf size {2, 4, 10}, regularisation ( $\alpha \in \{0, 0.1, 1\}$ ,  $\lambda \in \{0, 1, 10\}$ ), gain scaling {1, 5, 10}

Best: 128 bins, 32 interaction bins, no interactions, validation split 0.2, early stopping 100, learning rate 0.015, 5000 rounds, min leaf size 2,  $\alpha = 0.1$ ,  $\lambda = 1.0$ , gain scale 5

#### Decision Tree

Tuning: criterion {gini, entropy}, max depth {2, 4, 6, 8}, min split {5, 10, 20}, min leaf {2, 4, 8}, class weights {balanced}

Best: criterion = gini, max depth 2, min split 5, min leaf 2, balanced weights

#### Naïve Bayes

Tuning: smoothing { $10^{-9}$ ,  $10^{-8}$ ,  $10^{-7}$ ,  $10^{-6}$ }

Best: smoothing  $10^{-9}$

**Table 2: Hyperparameter grids and selected configurations. Only the varied hyperparameters are reported; others remained at default values.**

### 3.2 Surveying Educators on Explanation Quality

To evaluate how educators perceive the quality of XAI visualisations within the PreSS context, we conducted a survey among computer science teachers. The four models from the technical comparison were used to generate explanations of the same prediction for a fictional student, Alice (21 years old, mathematics score 2/10, expected grade 85/100, programming self-efficacy 4/10). All models predicted that Alice would fail the course.

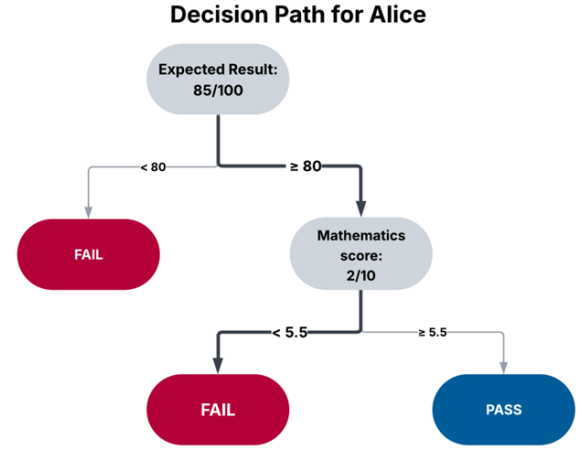


**Figure 1: Decision path visualisation used in the survey.**

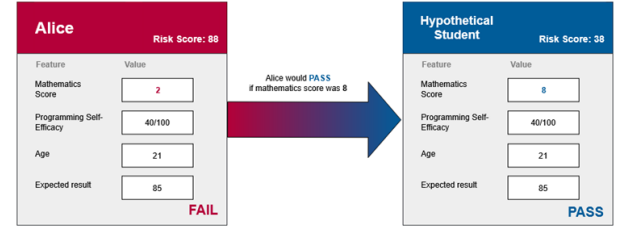


**Figure 2: Counterfactual visualisation used in the survey.**

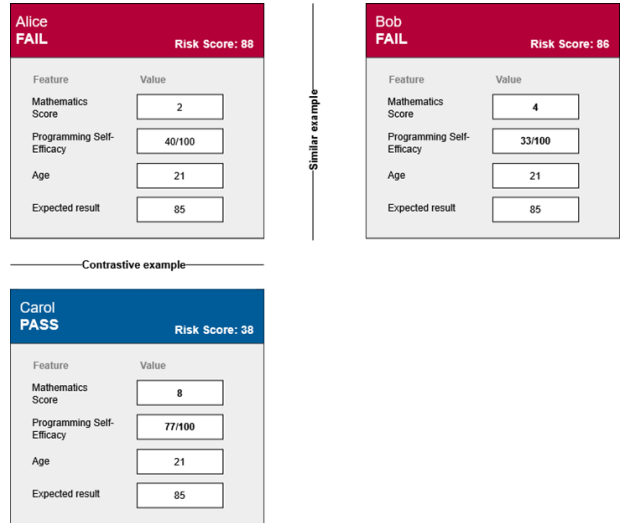


**Figure 3: Similar and contrastive example visualisation used in the survey.**

We considered four types of explanations — local feature importance, similar and contrastive examples, decision paths, and counterfactuals as depicted in Figures 1, 2, 3, and 4. To control for

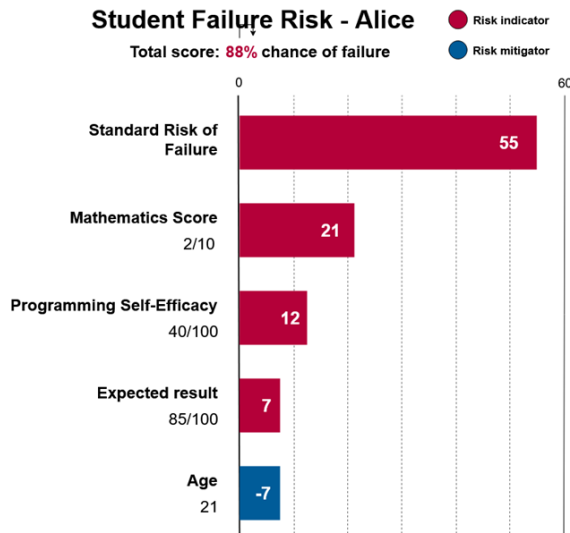


**Figure 4: Local feature importance visualisation used in the survey.**

confounding factors, all visualisations were designed to be visually consistent in terms of colours and layout, emphasising the effectiveness of the explanation type itself rather than peripheral details such as colour scheme, font, or stylistic variation.

The survey instrument was based on the taxonomy of explanation quality proposed by Kim et al. [14], which identifies eleven properties of explanation quality. Each property was operationalised as a single Likert-scale item (1 = strongly disagree to 5 = strongly agree). The items are listed in Table 3. No personal data about respondents were collected.

**Table 3: Survey items and their corresponding construct components, adapted from Kim et al. [14].**

| #  | Item                                                | Construct              |
|----|-----------------------------------------------------|------------------------|
| 1  | I understand the explanation.                       | Understanding          |
| 2  | The explanation is easy to understand.              | Ease of understanding  |
| 3  | The explanation is easy to use.                     | Ease of use            |
| 4  | The explanation is satisfactory.                    | Satisfaction           |
| 5  | The explanation is helpful in achieving my goals.   | Usefulness             |
| 6  | The explanation is trustworthy.                     | Trustworthiness        |
| 7  | The explanation is typical for similar AI outputs.  | Typicality             |
| 8  | The explanation is sufficiently detailed.           | Sufficiency            |
| 9  | The explanation seems accurate.                     | Subjective correctness |
| 10 | The explanation is concise and not overly detailed. | Compactness            |
| 11 | The explanation tells me how to act next.           | Actionability          |

The survey began with a general overview and a description of the fictional student, Alice. Respondents were first asked to rank the four explanation types by likelihood of use, prior to seeing any further details. For each visualisation, a short description was

provided to accommodate respondents without prior exposure to XAI, followed by the eleven survey items. At the end, respondents could provide open-ended feedback.

The survey was distributed online to approximately 200 educators in Ireland and the Netherlands, resulting in 19 complete responses. Participants were recruited through professional education networks and institutional mailing lists. No demographic or personal data (e.g., gender, teaching phase, or AI-related experience) were collected in order to avoid the need for formal ethical review and to maintain participant anonymity.

All 19 respondents provided qualitative feedback. Given the limited volume of qualitative responses, thematic insights were distilled directly from the open-ended responses without formal coding. The first two authors reviewed and discussed the responses to reach an agreement on the emergent themes. While the small sample size falls below conventional thresholds for statistical power, the repeated-measures design enhances sensitivity to within-subject differences.

## 4 Results

The results are presented in two parts, reflecting the research questions. First, we compare the predictive performance of NB, currently used in PreSS, with three glass box models. This analysis addresses the technical feasibility of adopting alternative models. Second, we report the evaluation of four explanation types in a survey with computer science educators, focusing on their perceived quality and practical usefulness.

### 4.1 Technical Performance of Models

Naïve Bayes (NB), currently used in the PreSS tool, was compared against three inherently interpretable “glass box” models: Decision Tree (DT), Automatic Piecewise Linear Regression (APLR), and Explainable Boosting Machine (EBM). All models were evaluated on the same test set using accuracy, recall, precision, and F1-score. While accuracy provides an overall indication of performance, the models were specifically optimised to increase recall, as the primary objective of PreSS is to identify, and not miss, at-risk students.

Overall, the models performed comparably to NB. DT achieved the highest accuracy (0.69), closely followed by EBM (0.68) and NB (0.66). APLR achieved the highest recall (0.87), correctly identifying most at-risk students, but at the cost of lower accuracy (0.60). EBM and DT demonstrated balanced performance, achieving the highest F1 scores. Detailed results are reported in Table 4.

**Table 4: Best configurations were selected on the validation set, and predictive performance was reported on the independent test set.**

| Model | Accuracy | Recall | Precision | F1 Score |
|-------|----------|--------|-----------|----------|
| EBM   | 0.676    | 0.769  | 0.690     | 0.727    |
| APLR  | 0.597    | 0.867  | 0.615     | 0.720    |
| DT    | 0.691    | 0.808  | 0.692     | 0.746    |
| NB    | 0.662    | 0.731  | 0.687     | 0.708    |

## 4.2 Survey Results

Nineteen educators evaluated four types of XAI visualisations generated from the models: feature importance (FI), similar and contrastive examples (SCE), decision paths (DP), and counterfactuals (CF). The survey followed the eleven-item taxonomy of explanation quality by Kim et al. [14]. At the beginning of the survey, respondents ranked the explanation types based on their likelihood of use, before being shown any further information. FI was rated highest (mean rank 2.42), followed by SCE (2.47), with CF (2.74) and DP (2.79) ranked lower (Table 5).

**Table 5: Average first-impression rankings of explanation types (1 = most likely, 4 = least likely).**

| Explanation Type             | Average Ranking |
|------------------------------|-----------------|
| Feature Importance           | 2.42            |
| Similar/Contrastive Examples | 2.47            |
| Counterfactuals              | 2.74            |
| Decision Rules               | 2.79            |

When evaluating the explanations on perceived quality, CF received the highest overall average rating (3.54), followed by SCE (3.43), with DP (3.25) and FI (3.06) scoring lower. At the construct level, understanding (4.10) and ease of understanding (3.87) achieved the highest averages, whereas actionability was rated lowest (2.28). Tables 6 and 7 provide the detailed results across constructs and explanation types.

A repeated-measures ANOVA tested for differences between explanation types. Results showed a significant main effect on trust ratings,  $F(3, 57) = 2.93, p = .042$ , with a small effect size ( $\eta^2 = .042$ ). Pairwise comparisons indicated that CF was rated significantly higher than FI ( $p = .017, d = 0.58$ ), and SCE higher than FI ( $p = .025, d = 0.45$ ). However, these results did not remain significant after Bonferroni correction. Internal consistency of survey items was excellent, with Cronbach's  $\alpha$  ranging from 0.89 to 0.94.

## 4.3 Qualitative Analysis of Open-Ended Responses

Open-ended responses ( $n = 19$ ) from educators revealed four main themes. First, six educators expressed uncertainty about how the explanations should inform their teaching. One noted, *"I don't think I understand the purpose of the suggested tool. I don't know what to do next based on the data shown."* Second, five educators pointed to issues with the questionnaire itself, such as overlapping or unclear items: *"At the beginning I hardly knew what it was about, too much explanation for a questionnaire; the examples were better explained individually."* Another added, *"Many questions about the output, difficult to distinguish between them."* Third, two educators highlighted technical or content inconsistencies in the materials. Finally, two educators offered broader reflections on predictive models, cautioning against fatalism and emphasising the importance of student agency: *"It seemed too fatalistic, or Calvinistic if you prefer... We always need to give students hope."* In contrast, twelve educators stated that they encountered no issues. These comments underscore the importance

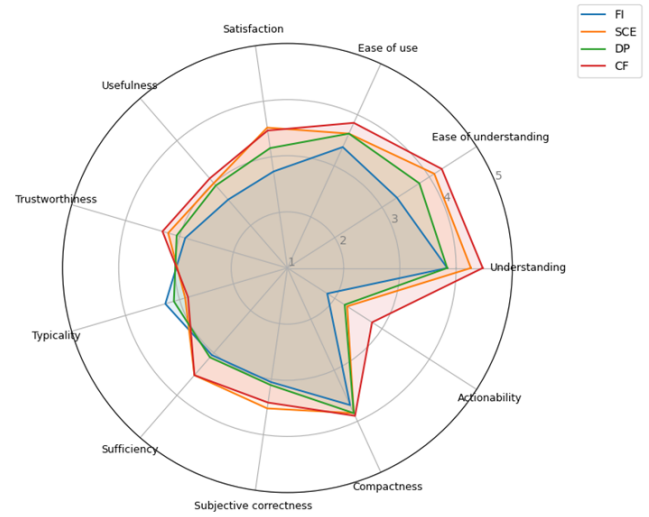


**Figure 5: Average scores across explanations and construct components.**

of aligning predictive tools and their explanations with educators' practical needs.

## 5 Discussion

The findings of this study offer several insights for model design, the PreSS application domain, and the broader use of XAI in education.

A first implication concerns the technical comparison central to our first research question. Both EBMs and Decision Trees performed on par with, or in some cases better than, the current NB model, particularly when recall was prioritised. This is important since the main objective of PreSS is to identify at-risk students without overlooking them. APLR achieved the highest recall, making it a strong candidate when the focus is solely on early detection, though at the cost of reduced accuracy. This reflects a deliberate shift in emphasis compared to previous work, which often prioritised accuracy above other metrics. Within this project, recall was optimised explicitly, meaning that results cannot be directly compared to earlier studies, but they do demonstrate that glass-box models can outperform NB under recall-driven evaluation.

Second, the survey results provide insights into educators' explanation preferences. Counterfactuals (3.54) and contrastive examples (3.43) received the highest ratings, consistent with Miller's [19] observation that humans naturally seek contrastive explanations, while decision paths (3.25) and feature importance (3.06) were rated lower. At the construct level, understanding (4.10) and ease of understanding (3.87) scored highest, whereas actionability (2.28), along with usefulness and typicality, lagged behind. This pattern suggests that while educators generally grasp what explanations depict, they struggle to apply them in practice. Notably, the most preferred explanation types are both post-hoc and model-independent, underscoring the need for human-centered design that pairs interpretability with clear, actionable guidance for classroom use.

**Table 6: Survey results across explanation types and construct components (Part 1).**

| Explanation | Understanding | Ease of Understanding | Ease of Use | Satisfaction | Usefulness  |
|-------------|---------------|-----------------------|-------------|--------------|-------------|
| FI          | 3.84          | 3.32                  | 3.37        | 2.74         | 2.61        |
| SCE         | 4.26          | 4.11                  | 3.63        | 3.53         | 3.00        |
| DP          | 3.84          | 3.79                  | 3.63        | 3.16         | 2.94        |
| CF          | <b>4.47</b>   | <b>4.26</b>           | <b>3.84</b> | <b>3.47</b>  | <b>3.11</b> |

**Table 7: Survey results across explanation types and construct components (Part 2).**

| Explanation | Trustworthiness | Typicality  | Sufficiency | Subjective Correctness | Compactness | Actionability |
|-------------|-----------------|-------------|-------------|------------------------|-------------|---------------|
| FI          | 2.89            | <b>3.26</b> | 3.05        | 3.05                   | 3.68        | 1.84          |
| SCE         | 3.21            | 2.89        | 3.52        | <b>3.53</b>            | 3.84        | 2.26          |
| DP          | 3.05            | 3.11        | 3.11        | 3.11                   | 3.84        | 2.21          |
| CF          | <b>3.32</b>     | 2.84        | <b>3.53</b> | 3.42                   | <b>3.89</b> | <b>2.79</b>   |

Third, the open-ended survey responses highlight broader concerns about AI-powered decision-support tools as quoted in Section 4.3. Several educators expressed uncertainty about how to act upon the explanations, questioning whether they should adapt their teaching, advise students, or simply interpret the model output. Others raised concerns that predictive tools risk overlooking individual or contextual factors, such as students who initially perform poorly on a task but later improve, or those whose strengths (e.g., creativity or perseverance) are not captured by the selected variables. These reflections emphasise the importance of aligning tools such as PreSS with the realities of educational practice. For example, if mathematics proficiency is highlighted as a key risk factor, teachers could respond with targeted support sessions. Conversely, indicators such as gaming or time spent online might serve as starting points for discussions on time management. At the same time, educators cautioned against potential harms, including undermining student confidence through negative predictions or inadvertently reinforcing biases.

Beyond these insights, two important limitations of the study should be noted. First, while Cronbach’s alpha values (0.89–0.94) indicate high internal consistency, the survey instrument was not validated through expert review, pilot testing, or cognitive interviews. As such, the present survey should be considered an exploratory pilot. Future work should improve robustness by validating the scale, refining items, and employing probability-based sampling from a more complete frame of computer science educators. Respondent feedback also suggests that clearer instructions are needed, both within the survey and regarding the purpose of the PreSS tool. Second, the balance between sufficiency and compactness warrants consideration. On average, explanations were rated higher on compactness (3.81) than sufficiency (3.30), indicating that explanations tended to be concise but perhaps lacking in detail. The relatively small difference suggests scope for improvement, but future designs should strive to balance brevity with informativeness.

In summary, this study addressed two research questions. The technical comparison showed that glass-box models can serve as viable alternatives to NB in the PreSS context, thereby answering

the first research question. The survey results indicated that educators value contrastive forms of explanation, but also highlighted challenges in designing explanations that are not only understandable but also actionable, directly informing the second research question. Taken together, these findings suggest that before such systems are implemented in educational settings, careful attention must be given to explanation design and survey validation.

## 6 Conclusion

This research investigated whether inherently explainable machine learning models can serve as effective alternatives to the current NB model within the PreSS tool, and how educators perceive the quality of explanations generated by these models. The technical comparison showed that glass box models not only match NB in accuracy, but can also outperform it when optimised for recall, making them well-suited for identifying at-risk students.

The survey results revealed that while all explanation types were judged sufficient, contrast-based explanations — counterfactuals and similar/contrastive examples — were consistently rated highest by educators. This aligns with Miller’s [19] argument that humans have a natural preference for contrastive reasoning, and suggests that such explanation formats may be especially effective in educational contexts.

Future work should expand the range of explanation formats, and conduct more scientifically robust surveys with validated instruments and representative samples. Open-ended feedback also highlighted the importance of clear instructions and guidance for educators when using PreSS or similar systems. Addressing these concerns is critical to ensuring that predictive tools support, rather than hinder, teaching practice.

If these issues are resolved, tools such as PreSS have the potential to play a significant role in reducing student failure rates. By helping educators identify and support at-risk students, such systems could enable more efficient allocation of attention and resources. While this study focused on computer science education, the approach could be extended to other disciplines where early dropout is a pressing challenge.

## References

- [1] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.
- [2] Seung Youn Chyung, Amy J Moll, and Shelley A Berg. 2010. The role of intrinsic goal orientation, self-efficacy, and e-learning practice in engineering education. *Journal of Effective Teaching* 10, 1 (2010), 22–37.
- [3] European Commission. 2025. *2025 State of the Digital Decade package | Shaping Europe's digital future*. <https://digital-strategy.ec.europa.eu/en/policies/2025-state-digital-decade-package>
- [4] European Commission. 2025. *AI Act | Shaping Europe's digital future*. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- [5] DUO. 2024. *Trends in het ho - Publicaties - DUO*. <https://www.duo.nl/open Onderwijsdata/hoger-onderwijs/publicaties/trends-in-het-ho.jsp>
- [6] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.
- [7] Erudera. 2024. *Rise in College Dropouts in Ireland: About 7,000 Students Don't Progress to Second Year*. <https://erudera.com/news/rise-in-college-dropouts-in-ireland-about-7000-students-dont-progress-to-second-year/>
- [8] Karl Flinders. 2019. *Computer science undergraduates most likely to drop out | Computer Weekly*. <https://www.computerweekly.com/news/252467745/Computer-science-undergraduates-most-likely-to-drop-out>
- [9] Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statistical science* 1, 3 (1986), 297–310.
- [10] Johan Huysmans, Bart Baesens, and Jan Vanthienen. 2006. Using rule extraction to improve the comprehensibility of predictive models. (2006).
- [11] Studying in Germany. 2024. *Bachelor Students in Germany More Likely to Drop Out Within First Academic Year*. <https://www.studying-in-germany.org/bachelor-students-in-germany-more-likely-to-drop-out-within-first-academic-year/>
- [12] Education Data Initiative. [n. d.]. *College Dropout Rate [2023]: by Year + Demographics*. <https://educationdata.org/college-dropout-rates>
- [13] InterpretML. [n. d.]. *Explainable Boosting Machine — InterpretML documentation*. <https://interpret.ml/docs/ebm.html>
- [14] Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable AI applications: a systematic review. 7 (2024), 1456486. doi:10.3389/frai.2024.1456486 Publisher: Frontiers.
- [15] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5686–5697.
- [16] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.
- [17] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
- [18] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [19] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [20] Christoph Molnar. 2025. *Interpretable Machine Learning* (3 ed.). <https://christophm.github.io/interpretable-ml-book/>
- [21] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s, Article 295 (July 2023), 42 pages. doi:10.1145/3583558
- [22] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).
- [23] Keith Quille and Susan Bergin. 2018. Programming: predicting student success early in CS1. a re-validation and replication study. In *Proceedings of the 23rd annual ACM conference on innovation and technology in computer science education*. 15–20.
- [24] Keith Quille and Susan Bergin. 2019. CS1: how will they do? How can we help? A decade of research and practice. *Computer Science Education* 29, 2-3 (2019), 254–282.
- [25] Keith Quille, Natalie Culligan, and Susan Bergin. 2017. Insights on Gender Differences in CS1: A Multi-institutional, Multi-variate Study.. In *Proceedings of the 2017 acm conference on innovation and technology in computer science education*. 263–268.
- [26] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
- [27] Sigrid Schefer-Wenzl, Igor Miladinovic, Sandra Bachinger-Raithofer, and Christina Muckenhuber. 2023. A study on reasons for student dropouts in a computer science bachelor's degree program. In *International Conference on Interactive Collaborative Learning*. Springer, 391–400.
- [28] Statista. [n. d.]. *Spain: dropout rate by region 2023*. <https://www.statista.com/statistics/1295163/spain-dropout-rate-by-region/>
- [29] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual explanations and algorithmic recourses for machine learning: A review. *Comput. Surveys* 56, 12 (2024), 1–42.
- [30] Mathias von Ottenbreit and Riccardo De Bin. 2024. Automatic piecewise linear regression. *Computational Statistics* 39, 4 (2024), 1867–1907.
- [31] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
- [32] George Winograd. 2025. *College Dropout Rates Statistics 2025 (By Majors)*. <https://missiongraduatennm.org/college-dropout-statistics/>. Published May 28, 2025.