electronics

MDPI

Article

Assessing Perceived Trust and Satisfaction with Multiple Explanation Techniques in XAI-Enhanced Learning Analytics

Saša Brdnik, Vili Podgorelec and Boštjan Šumak *

Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia; sasa.brdnik@um.si (S.B.); vili.podgorelec@um.si (V.P.)
* Correspondence: bostjan.sumak@um.si

Abstract: This study aimed to observe the impact of eight explainable AI (XAI) explanation techniques on user trust and satisfaction in the context of XAI-enhanced learning analytics while comparing two groups of STEM college students based on their Bologna study level, using various established feature relevance techniques, certainty, and comparison explanations. Overall, the students reported the highest trust in local feature explanation in the form of a bar graph. Additionally, master's students presented with global feature explanations also reported high trust in this form of explanation. The highest measured explanation satisfaction was observed with the local feature explanation technique in the group of bachelor's and master's students, with master's students additionally expressing high satisfaction with the global feature importance explanation. A detailed overview shows that the two observed groups of students displayed consensus in favored explanation techniques when evaluating trust and explanation satisfaction. Certainty explanation techniques were perceived with lower trust and satisfaction than were local feature relevance explanation techniques. The correlation between itemized results was documented and measured with the Trust in Automation questionnaire and Explanation Satisfaction Scale questionnaire. Master's-level students self-reported an overall higher understanding of the explanations and higher overall satisfaction with explanations and perceived the explanations as less harmful.

Keywords: Explainable Artificial Intelligence; learning analytics; XAI; XAI techniques; trust; explanation satisfaction

Citation: Brdnik, S.; Podgorelec, V.; Šumak, B. Assessing Perceived Trust and Satisfaction with Multiple Explanation Techniques in XAI-Enhanced Learning Analytics. *Electronics* **2023**, *12*, 2594. https://doi.org/10.3390/electronics12122594

Academic Editors: Arianna D'Ulizia, Patrizia Grifoni and Fernando Ferri

Received: 8 May 2023
Revised: 5 June 2023
Accepted: 6 June 2023
Published: 8 June 2023

1. Introduction

Interest in artificial intelligence (AI) has been increasing rapidly over the past decade and has expanded to essentially all domains. Along with it grew the need to understand the predictions and suggestions provided by machine learning. Explanation techniques have been researched intensively in the context of explainable AI (XAI), with the goal of boosting confidence, trust, user satisfaction, and transparency. This paper aims to investigate how different explanation techniques affect perceived trust and satisfaction in XAI-enhanced learning analytics. The demand for explainable AI has been emerging in recent years, as observed in the literature review conducted by Haque et al. [1], which analyzed 58 papers from the field and recognized that the measurement of information (or explanation) quality dimensions related to XAI has not been discussed. Authors have recognized that explanation evaluations should focus on fixed XAI effects, such as trust, transparency, understandability, usability, and fairness. The need for additional research into explainable student performance prediction models, where explainability and model accuracy are properly quantified and evaluated, has already been recognized [2]. A review of current trends, challenges, and opportunities for XAI in the educational field [3] highlighted the importance of non-algorithmic design choices in optimizing the learning experience and AI tools in education, such as using simple models and increasing their complexity only if necessary while continuously measuring their interpretability along with their accuracy.

The main motivation behind this work is to understand the impact of selected XAI techniques on user trust and satisfaction in the context of XAI-enhanced learning analytics, as this is crucial for designing effective XAI systems. By investigating the preferences and perceptions of STEM college students with varying study levels, this study aims to provide insights with the goal of enhancing trust and satisfaction in learning analytics systems with XAI explanations, ultimately facilitating the development of more user-centric and transparent AI systems in educational settings. We aim to contribute to the recognized research gap of evaluation of the explanation techniques [1,4], which we address in the context of learning analytics. As indicated by [5], significant differences in effects on trust and satisfaction can be detected based on the type of explanation used.

The paper is structured as follows. Section 2 presents a background and brief overview of the related work. The materials and methods are outlined in Section 3. Section 4 reports the results of the experiment. In Section 5, the results are discussed and compared to approaches from related work. Summarized remarks and proposed future research directions are also provided in this section.

2. Background and Related Work

2.1. Explainable Artificial Intelligence

The term XAI is best described as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" [6]. Research on XAI shows that introducing explanations to AI systems to illustrate their reasoning to end users can improve transparency, interpretability, understanding, satisfaction, and trust [7–10]. Observing the explainability techniques with relation to the machine learning models, Barredo et al. [11] presented a taxonomy that separates transparent models (such as decision trees, logistic regression, linear regression, and K-nearest neighbor) that are de facto explainable from models where post-hoc explainability has to be utilized (e.g., support vector machines, convolutional neural networks) to generate their explanations. Post-hoc explanations can be model-agnostic or model-specific. The former can be applied to any machine learning model with no regard to its inner process or representation, while the latter is related to the interpretation and understanding of a specific machine learning model. Various classifications exist for explanations in AI. They can be categorized mainly as global approaches, explaining the entire model, versus local approaches explaining an individual prediction; or as self-explainable models with a single structure versus post-hoc approaches explaining how a model produces its predictions without clarifying the structure of the model [11,12].

Common explainability approaches [11,12] include *global explanations*, which explain how different features/variables affect predictions within the model in question; *feature relevance*, which presents the computed relevance of each feature in the prediction process (simplified displays with a selection of the most important features are often used); and *example-based explanations*, which select a particular instance to explain the model, offering a more model-agnostic approach, which can be local or global. Additionally, *local explanations* are often used in systems for students and focus on a particular instance, independent of the higher-level general model. *Comparison* uses a selection of instances to explain the outcome of other instances on a local level. *Counterfactual explanations* describe a causal situation (i.e., formulated as "If X had not occurred, Y would not have occurred") and explain and demonstrate the effects of small changes of feature values on the predicted output. *Explanations by simplification* use mentioned techniques to build a new similar yet simplified system (with reduced complexity but similar performance) based on the trained model to be explained. The aforementioned techniques for post-hoc explanations can include visualizations and text explanations. Their selection is conditioned by the type of machine learning model used for prediction.

Lim [13] presented a slightly different classification of ten explanation types, dividing them into model-independent and model-dependent explanation types. Model-independent explanations include *input explanations*, which inform users about the used

input sensors and data sources, to ensure understanding of the explanation scope; *output explanations* inform users about all the possible outputs a system can produce; *what explanations* inform users of the system state in terms of output value; and *what if explanations* allow users to speculate about different outcomes by changing the set of user-set inputs. Model-dependent explanations, on the other hand, include *why* explanations, informing users why the output is derived from input values, possibly returning used conditions (rules); *why not* explanations, presenting users with information about why the alternative output was not produced based on the input; *how to* explanations, which provide explanation as to how the desired outcome is generally produced; and *certainty* explanations, which inform users about the certainty of the produced outcome.

Explanations within XAI lack standardization for their design, as well as their evaluation, as confirmed by literature reviews of the field [1,11]. Haque et al. [1] conducted a literature review of the XAI field and extracted major research themes as future research directions: XAI standardization (which includes developing comprehensive guidelines or standards for developing an XAI system), XAI visualization (focus on empirically measuring the explanation quality dimensions), and XAI effects (measuring user perceptions of the transparency, understandability, and usability of XAI systems). Additionally, Mohseni [14] recognized that the XAI design and evaluation methods should be adjusted based on the set goals of XAI research.

2.2. XAI in Education

AI systems are complex and, by default, suffer from bias and fairness issues. Explanations of AI were introduced in the field of human–computer interaction as a way to allow users to interact with systems that might be faulty in unexpected ways [15]. Explanations allow users to engage with AI systems in an informed manner and adapt their reliance based on the provided explanations [6]. Multiple studies have shown that introducing explanations in tutoring and e-learning systems increases students' trust. Ooge et al. [10] observed changes in trust after introducing explanations in an e-learning platform for mathematics exercise recommendations. Explanations increased initial trust significantly when measured as a multidimensional construct (consisting of competence, benevolence, integrity, intention to return, and perceived transparency), while no changes were observed with one-dimensional measures. Conati et al. [16] presented students with personalized XAI hints within an intelligent tutoring system, evaluating their usefulness, intrusiveness, understanding, and trust. Providing students with explanations led to higher reported trust, while personalization improved their effectiveness further. The improvement in understanding of the explanations was related to students' reading proficiency; students with high levels of reading proficiency benefited from explanations, while students with low levels did not. A study of XAI in education [12] analyzed the concepts of fairness, accountability, transparency, and ethics and proposed a framework for studying educational AI tools, including analysis of stakeholders, benefits, approaches, models, designs, and pitfalls.

Displays that aggregate different indicators about learners, learning processes, and/or learning context into visualizations can be categorized as learning analytics (LA) [17]. A systematic review of LA dashboard creation [18] showed that most dashboards (75%) are developed for teachers and that less focus is put on solutions targeted at learners. Additionally, only two observed propositions provided feedback or warnings to users, and only four papers used multiple data sources, indicating that this is an opportunity for future research. It is important to note that LA does not necessarily include AI. In the core literature [19], LA is defined as the "analysis and representation of data about learners in order to improve learning". It can be conducted using traditional statistical methods or other data analysis approaches without the involvement of AI. Predictive modeling, the base functionality of many LA systems, is not that different from a traditional teacher recognizing which students are struggling in their class and providing them extra help or direction during the semester. The cost of LA utilization is derived from its functionalities;

firstly, the predictions and analyses displayed in LA systems are based on estimations and probabilities, which many users fail to understand correctly [10,18,19]. Making decisions based on wrongly understood probabilities is problematic, especially if the output triggers other actions, or self-regulated learning, without the teacher's involvement [19]. Additionally, there are challenges with privacy, data quality, availability, and fitness of data used in LA solutions in education [20]. On the other hand, there are many benefits of utilizing LA, mainly the improvement of the learning process based on the data available. Furthermore, students can improve their perceptions of the activity and have their personalized analyses available in more depth than a teacher could provide to each student during their limited time [19]. Overview of the trends in education systems [3] has shown that AI has been recognized as a trend in the educational setting, as more and more AI systems are used in LA, learning management systems, and educational data mining [20]. Some of the most common uses of AI [21] include use cases for profiling and prediction, assessment and evaluation, adaptive systems and personalization, and intelligent tutoring systems. Along with AI models, interpretable machine learning and XAI have been gaining interest in LA systems, as they offer a better understanding of the predictive modeling [20]. The trend of including AI in education has resulted in the development of the term artificial intelligence in education (AIEd). This field overlaps with LA. The main benefits of introducing AI in education and in the LA field [22] can be summarized with the development of intelligent agents, personalized learning systems, or environments and visualizations that offer deeper understanding than the classic non-AI analyses.

Related work on predicting students' course achievement used logs from virtual learning environments [23] along with demographic data [24] and grades [25] in their prediction models. The need for the interpretability of the complex models used in education mining data techniques has been highlighted [26], and explanations of the model's predictions have been introduced slowly, by [27] offering verbal explanations (i.e., "Evaluation is Pass because the number of assessments is high"), and by [10] offering verbal and visual explanations to students. In a related study, Conijn et al. [28] analyzed the effects of explanations of an automated essay scoring system on students' trust and motivation in the context of higher education. The results indicated there is no one-size-fits-all explanation for different stakeholders and in different contexts.

2.3. Measuring Trust and Satisfaction

Various elements can be observed for measuring the effectiveness of an explanation; namely, user satisfaction, trust assessment, mental models, task performance, correctability [6], and fairness [29]. This study is focused on the first two measures. We followed the definition of trust as provided by Lee [30], defining it as "an attitude that an agent will achieve an individual's goal in a situation characterised by uncertainty and vulnerability". Many scales for assessing trust are presented in the scientific literature, and many of them were created with interpersonal (human-to-human) trust in mind. A considerable research gap is still reported in the studies, focusing on human–AI trust [9,31]. Vereschak et al. [31] surveyed existing methods to empirically investigate trust in AI-assisted decision-making systems. This overview of 83 papers shows a lack of standardization in measuring trust and considerable variability in the study designs and the measures used for their assessment. Most of the observed studies used questionnaires designed to assess trust in automation (i.e., [32–35]). Numerous factors have been shown to increase users' trust [36]. Transparency has gained much attention, highlighting the need for explanations that make the systems' reasoning clear to humans. However, trust has been found to increase when the reasoning for the AI system's decision is provided and to decrease when information on sources of uncertainty is shared with the user [9].

Explanations cannot be evaluated without measuring the user's satisfaction with the provided explanation, which Hoffman [5] defines as "the degree to which users feel that they understand the AI system or process being explained to them. It is a contextualised, a posteriori judgment of explanations". A similar study measuring trust, explanation

satisfaction, and mental models with different types of explanations has been conducted in the case of self-driving cars [37]. The study reported the lowest user satisfaction with causal explanations and the highest levels of trust with intentional explanations, while mixed explanations led to the best functional understanding of the system. Related evaluation of understandability, usefulness, trustworthiness, informativeness, and satisfaction with explanations, generated with popular XAI methods (LIME [38], SHAP [39], and Partial Dependence Plots or PDP [40]) was conducted by [41], reporting higher satisfaction with global explanations with novice users compared to local feature explanations. Comparing the popular methods, PDP performed best on all evaluated criteria.

Comparing levels of explanation satisfaction and trust between different groups of users can be conducted based on various user characteristics. Level of experience and age are (along with personality traits) two of the major user characteristics recognized to affect user performance and preferences in general human–computer interaction. Although the scale from novice to expert is continuous, there is no universally accepted classification and definition of users' level of experience and/or knowledge [42]. Level of experience is recognized as "the relative amount of experience of user segments of the user population" [43]. In higher education, groups of students can be distinguished based on the amount of ECTS (European Credit Transfer and Accumulation System) points they acquired during their studies. ECTS credits express the volume of learning based on the defined learning outcomes and their associated workload [44].

2.4. Objective

Evaluation of the explanations generated within XAI has been recognized as an important research direction [4] in the XAI field [1,11]. We aim to contribute to this debate by following the approach proposed by [1]; we present XAI explanations in various formats and evaluate and measure various representations to find suitable representation techniques for XAI in the context of a selected scenario of LA. To the best of our knowledge, to date, no study has focused on comparing all eight selected XAI explanation techniques based on perceived trust and explanation satisfaction in the LA environment in the context of higher education. Some related work has been conducted. Conijn et al. [28] analyzed the effects of explanations of an automated essay scoring system on students' trust and motivation in the context of higher education, observing two types of explanations: full-text global explanations and model accuracy statements. Ooge et al. [10] focused on measuring multidimensional trust in XAI e-learning systems with adolescents. A similar evaluation methodology for comparing explanation methods was used in [37] in the context of self-driving cars, where trust, explanation satisfaction, and mental models were measured. Our study is aimed at comparing perceived trust and satisfaction with eight selected established techniques. The following research questions were set to address the recognized research gap:

- RQ1: What is the impact of diverse XAI explanation approaches on user trust and satisfaction in the context of learning analytics?
- RQ1: How does the study level impact user trust and satisfaction in diverse XAI explanation approaches?

3. Materials and Methods

The study was conducted at the University of Maribor, Faculty of Electrical Engineering and Computer Science, in the academic year 2022/2023. First-year bachelor's students from the higher education Informatics and Technologies of Communication study program attending course A and first-year master's students from the Informatics and Data Technologies program attending course B were invited to participate. Course A is an introductory course in one of the bachelor's engineering study programs, while course B is part of the master's student program and is focused on basic knowledge of web technologies, the programming language JavaScript, and web service development. Around 100 students attend course A, while course B is attended by around 50 students yearly. Course A is orga-

nized in winter, and course B is in the spring. The study was organized in the classroom, in a semi-controlled environment for both courses, where students participated in smaller groups. The experiment was conducted in two phases: with bachelor's students in the first week of January (15th week of the winter semester) and with master's students in the first week of February (2nd week of the spring semester). Both groups were first introduced to the aim of the research and the predictive system. Students' consent for publication of their anonymized data was gathered within the system. Participants were presented with the prediction of their academic performance for this course, followed by various model explanations. Each explanation was presented on a separate site, and the students were asked to fill out the questionnaire based on the observed explanation on each site. The link to the questionnaire was included in the system. Explanations were grouped into two groups: prediction explanations and model accuracy explanations. First, the prediction explanations were presented to users in randomized order, followed by explanations of accuracy, again in a randomized order.

Trust was measured with one of the questionnaires, derived from the trust in automation literature, which is commonly used for measuring trust in AI-assistant situations. As stated earlier, no standards currently exist for measuring trust in XAI systems, so guidelines proposed by [31] were followed, and an established questionnaire was used that captures the key elements of trust. An adaptation of the Trust in Automation [32] questionnaire was selected for use in this study, as it includes vulnerability and positive expectations. The questionnaire consists of 12 items, which are measured using a 7-point Likert scale. The adaptations to the questionnaire were limited to changing the word 'system' to 'explanation'. This change affected all 12 questionnaire items. To support the replicability of this work and increase scientific rigor, the adapted questionnaire, along with the translated Slovenian version used in this study, is presented in Table A2 in Appendix B. Explanation satisfaction was measured with Hoffman's questionnaire Explanation Satisfaction Scale (ESS) [5], which consists of 8 items measured using a 5-point Likert scale. The questionnaire was translated into Slovenian with no additional changes. The questions were formed in order to ask users about their satisfaction with explanations of the prediction model. The finalized and translated version of the questionnaire is included in Table A3 in Appendix B. Additionally, generalized satisfaction with the system was measured with the System Usability Scale [45] questionnaire.

3.1. Data and Prediction Models

Data from the two mentioned courses were used for building and training grade prediction models. Students from both courses could obtain between 0–100 grade points, which were later categorized in final grades between 5 and 10 (grade points 0–49 are assigned grade 5, 50–59 grade 6, 60–69 grade 7, 70–79 grade 8, etc.). The threshold for a positive grade was set at 50 grade points. For course A, the final grade was calculated from 8 assignments (35 grade points), 2 quizzes (15 grade points), 2 midterms (15 grade points each), and an oral exam (20 grade points). Students must also obtain at least a passing grade (25 grade points) from assessments and quizzes combined. For course B, the final grade was calculated from assignments (10 grade points), quizzes (10 grade points), 2 midterms (25 grade points each), and a completed project (30 grade points). The average grade for course A in the academic year 2021/2022 was 7.8; 18 out of 106 students failed the course (i.e., the pass rate reached 83%). The average grade for course B in the academic year 2022/2023 was 8.5, with a 51% pass rate. An anonymized log of interactions with the Moodle virtual learning environment, demographic data, and grades from the fall semester of the previous academic year was used as the data source for prediction models. The features used in the prediction models are presented in Table 1 and vary between the courses due to the data availability. Feature suggestions were derived from related works [23–25,46].

Table 1. Features used for prediction models.

Feature	Description
Course A	
Schedule group	The group (N = 6) in which students attend their lab work activities.
Gender	Gender of the student.
Disability	Boolean feature with information on whether a student holds the status of a disabled student.
Number of monthly clicks	Includes multiple features containing the number of clicks for all previous months from August until the date of prediction. The clicks for the ongoing month are calculated up to the day of the prediction.
Date of first interaction on Moodle	The date student first interacted with Moodle course, counted from the first day of the semester.
Sum of all clicks	The sum of all clicks from student's interaction with Moodle course.
Task grades	Multiple features, including the grades from all the tasks with deadlines prior to the date of the prediction.
Midterm grade	Grade student obtained on the first midterm (out of the two).
Quiz grade	Grade student obtained on the first quiz (out of the two).
Course B	
Graduation year	Year of bachelor program graduation. The data were collected from the national online library, where the final theses are published.
Alma mater of undergraduate studies	University where students obtained their bachelor's degree, if available.
Graduation from same Bologna bachelor's program	Boolean whether student obtained their bachelor's degree from the same bachelor's Bologna program and are continuing their studies in the same program in the next Bologna cycle.
Gender	Gender of the student.
Course C grade	Grade from the mandatory course, given in the first (fall) semester of master's studies. Course holder is the same as in course B.
Course D grade	Grade from the mandatory course, given in the first (fall) semester of master's studies. Course holder is the same as in course B.
Moodle access in first semester week	Boolean value noting if the students accessed the Moodle environment in the first week of the semester.
Sum of clicks in Moodle	Sum of all the clicks from Moodle learning environment.
Study module	One of the three study modules students selected as a form of specialization within the study program.

The grade for course A was predicted in two steps. First, to address the marginal students whose prediction was around the passing grade, a classification model was utilized for the classification of the students into pass or fail categories with 89% accuracy, which is comparable to and even higher than some results presented by related work [24,25,47]. The algorithm selection was conducted after comparing the accuracy performance of six different algorithms (logistic regression, K-nearest neighbors classifier, random forest classifier, Gaussian naive Bayes, decision tree classifier, and linear discriminant analysis), with the random forest classifier yielding the best results.

Due to the small sample size, 5-fold cross-validation with scikit KFold was performed in order to improve the estimate of the model's performance. Secondly, a regression-based model utilizing DecisionTreeRegressor was used for the prediction of the final grades of students who were classified as "pass" with the first model. Cross-validation with 5-fold KFold was again used to avoid overfitting. To evaluate the second model, the mean absolute error was calculated with the formula $\mathrm{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N}$, where y is the prediction and $\hat{y}$ is the corresponding true value. The MAE reached with this model was 8.6 grade points, which means this error rate was lower than the range of one grade (which is 10 grade points). Grade prediction for course B utilized the model similar to classification conducted with model A. Algorithm selection was again performed after comparing the accuracy performance of the six selected algorithms (logistic regression, K-nearest neighbors classifier, random forest classifier, Gaussian naive Bayes, decision tree

classifier, and linear discriminant analysis), though, in this case, logistic regression yielded the highest accuracy results at 83.3%. Cross-validation with 5-fold KFold was used in this model as well. All models were trained on the data from the previous academic year, 2021/2022. Prediction models were evaluated with scikit train_test_split, which splits the data into random train and test subsets. Reported accuracy and MAE values were obtained by evaluating the test subset.

3.2. Architecture and Tech Stack

The learning analytics dashboard was developed with the Javascript framework Next.js. The grade prediction model was prepared in Python version 3.8.5 in Jupyter Notebook version 6.1.4. The model was created using Numpy (version 1.19.2) and scikit-learn (version 1.1.2). The prediction model was exposed through an API via the Flask library. Students accessed the system with their university email. The architecture is presented in detailed in the Figure 1.

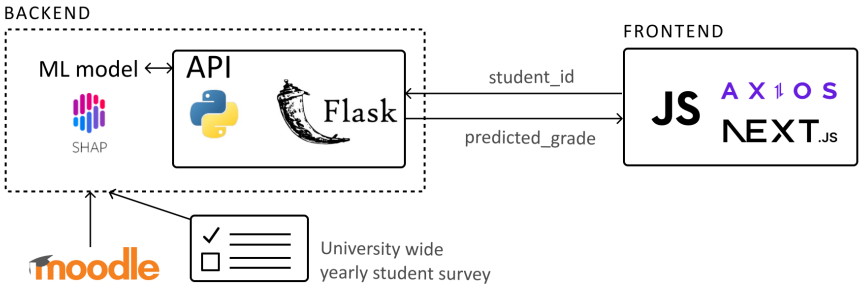


Figure 1. The architecture and tech stack of the proposed solution with data sources.

3.3. Interface and Explanation Techniques

The learning analytics dashboard presented the grading system and historical data on grades and course pass rates for both groups of students. Students were presented with their grade predictions and various explanations in randomized order. Each explanation was presented on a different screen, prompting students to answer a poll on trust and satisfaction before observing the next explanation. In some cases, explanations were circled with a blue border, to make it clear to the students which part of the explanation the polls were referring to.

The explanation techniques employed are summarized in Table A1 in Appendix A, along with their description and an example of the corresponding user interface element used in the explanation. Three local explanation techniques and two global explanation techniques were selected for comparison, along with three techniques explaining the certainty of the used model. The first selected XAI technique, *explanation A*, local explanation with force graph, utilized SHAP's [39] force plot for local feature importance explanation. This technique is similar to the local self-explaining approach, documented in [12]. With the second, *explanation technique B*, SHAP's [39] bar plot was used for local feature importance explanation. This type of explanation was also inspired by [12,13,48,49]. The positive and negative contributing features are color-coded in both aforementioned explanations: pink indicates a positive impact on the prediction, and blue indicates a negative impact. Their positive or negative impact is further underlined with the direction of the bar in explanation B. *Explanation technique C* was a textual explanation of local feature importance, inspired by justification statements used in the explanations in [10] and causal explanations used in [37]. The explanations were written in the unified format, only differentiated by the order of the features based on their strength (e.g., Your predicted score was boosted positively the most by the feature "Number of clicks on files on Moodle". Additionally, the features "Number of clicks on the forum" and "Number of system clicks" increased your predicted score. The feature "Number of clicks in September" also had a small positive impact on your prediction. Your predicted score was affected most negatively by the grade obtained for Task 2 and your grade obtained on Quiz 1).

All of the first three explanation techniques can be classified as explanations by simplification. *Explanation technique D* was based on the classification model's accuracy data obtained with scikit's accuracy score. The design of the explanation was inspired by [48]. However, the base accuracy was communicated in percentages, not in natural language (i.e., the explanation was formulated as "model accuracy is 90%", not "high" as suggested by [48]). *Explanation technique E* explained the accuracy of the regression model with MAE. The presented MAE was expressed in the form of the changes in lower and upper values of the predicted grade. The explanation was constructed as a complementary technique to explanation D, which can be used in the regression model. *Explanation technique F* consisted of a confusion matrix containing the data computed with scikit's confusion matrix, in order to evaluate the accuracy of the classification model. This technique was added due to its common use in model accuracy evaluation [49] and to assure the use of certainty explanations [13] in both test groups. *Explanation technique G* was created with [39] and presents the global absolute feature importance. *Explanation technique H* is three-fold and combines the overview of the historical data (grades in this course), a summary visualization of features from the sample, and a summary of the prediction for the whole sample (the number of students who were predicted to pass or fail for this year). This explanation can be classified as an input explanation [13], as it explains the data used in the predictions. The comparison with other predictions (other students) allows students to compare themselves with other predictions and is often included in learning analytics dashboards [50,51].

Due to the limitations of the data and consequent differences in the models, students attending course A were presented with explanations A–E. Master's-level students were presented with explanations A–D and F–H. Explanation E (MAE) was connected directly to the regression model used for predictions in the bachelor's student course and could not be used in the classification model utilized in the predictive model for master's students. That is why it was substituted by explanation F. Explanations G and H were added after the first experiment was conducted and were only presented to the master's-level students. The Venn diagram, showing the explanation techniques evaluated in each group, is presented in Figure 2.

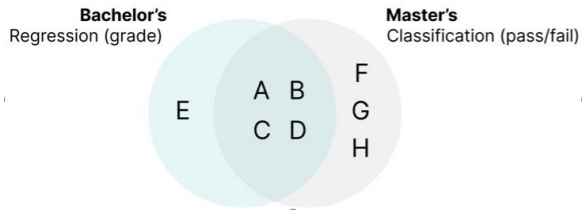


Figure 2. Venn diagram of explanation techniques shown to each group.

A preliminary interface evaluation was conducted with ten students. Learners received access to the dashboard with anonymized data of two representative users: one academically successful, and one at risk of failing the course. After interaction with the LA dashboard, their feedback was gathered using a questionnaire consisting of system usability scale (SUS) questions and questions regarding privacy and feature satisfaction, where students reported their answers on a 5-point Likert scale. A SUS score of 76.5 was achieved, which was interpreted as good by the SUS evaluation key. The students furthermore expressed their support for the use of their data in the implementation of LA (n = 4.6 points on the Likert scale, SD = 0.5) and reported feeling informed enough about the use and processing of their data (n = 4.1, SD = 0.9). Students believed that such display would, at least to some degree, motivate them in their studies (n = 3.8, SD = 1) and offer some help in planning their study activities (n = 3.7, SD = 0.8).

4. Results

Fifty-one students from course A and forty students from course B were included in the experiment. After the initial overview of the received questionnaires, some students' replies were removed due to: student errors in reporting the identifier of the explanations they were observing, questionnaires with unrealistically short answer times (a few seconds per poll), questionnaires with multiple missing values, and duplicated questionnaire responses. Furthermore, replies were removed from students who only filled out the questionnaire for one of the explanations. The finalized dataset included 168 questionnaires from students attending course A (with 38 students assessing explanation A, 37 assessing explanation B, 35 for explanation C, 23 for explanation D, and 35 for explanation E), and 197 questionnaires from students attending course B (with 34 responses for explanation A, 33 for explanation B, 30 for C, 22 for D, 32 for F, 23 for G, and 23 for H).

A brief overview of the techniques bachelor's students evaluated shows explanations G and H were regarded as the most satisfying on ESS (rated highest on the items of the sufficiency of detail, the satisfaction of model explanation, and usefulness to students' goals) and that they reached one of the highest trust scores (rated highest on the items for explanation integrity, confidence in explanation, trust in it, familiarity with the explanation, dependence of the explanation, and providing security). A comparison of local feature explanation techniques (A, B, and C) showed both groups of students rated explanation B (the bar graph form explanation) higher on both questionnaires, ESS and TIA, compared to explanations A and C. Observing the measured trust of all explanations, we discovered that the highest trust was measured with explanation B, which scored the highest in all positively stated items of the TIA questionnaire in the sample of bachelor's students. Observing only the explanations presented to both groups (A–D), we recognize that explanation B produced the highest mean values on five out of six positively stated TIA questionnaire items (with the exception of familiarity). Explanation B also reached very high mean values on the ESS questionnaire, reaching the highest mean values on three (Q3, Q5, Q8) and the second highest mean values on two (Q1, Q2) additional questionnaire items in the sample of master's students. Similarly, pattern B performed best in the sample of bachelor's students, reaching the highest mean values in five out of eight ESS questionnaire items (Q1–Q3, Q7, Q8). A detailed insight and a technique comparison are presented in the following subsections.

4.1. Explanation Satisfaction

A comparison was conducted of the Explanation Satisfaction Scale (ESS) questionnaire responses for different explanation techniques. We observed similarities in the mean ESS results for questionnaire items between both groups. The generally lowest-rated questionnaire items (on a Likert scale where 1 - I disagree strongly and 5 - I agree strongly) for all observed explanation techniques were those related to the completeness of the explanations. Explanation B was rated highest with the bachelor's students, with items 1–3 (M_{B1} = 3.78, M_{B2} = 3.84, M_{B3} = 3.76) and 7–8 (M_{B7} = 3.57, M_{B8} = 3.73) of the ESS questionnaire. Explanation C reached the highest mean score for instructing the users how to use the explanation (Q5 with M_{C5} = 3.63). It also reached high marks for the sufficiency of detail (Q3 with M_{C3} = 3.6) and for supporting users in deciding on the model's trustworthiness (Q8 with M_{C8} = 3.57). Explanations D and E, both explaining the accuracy of the prediction, generally scored lower on the ESS scale, although, importantly, explanation E reached a slightly higher (by 0.25 points) mean score for supporting users in their decision on trustworthiness. Master's-level students expressed slightly lower overall mean explanation satisfaction ($M_{ESSBachelor}$ = 3.25, $M_{ESSMaster}$ = 3.20) compared to bachelor's-level students.

The hypothesis of normal distribution of responses for each explanation sample was rejected with the Shapiro–Wilk test in all observed groups for the questionnaires of bachelor's and master's students, respectively. The Kruskal–Wallis H test (KW) was then used to determine if there were statistically significant differences between the ESS responses for different explanations. Comparisons were created separately for each course. Detailed results are presented in Table 2. The KW test revealed a statistically significant difference

($p = 0.013$, $\chi^2 = 12.736$) in bachelor's students' replies related to the explanation satisfaction item (Q2) on the ESS questionnaire. Further observation shows that bachelor's students were less satisfied with explanations D ($M_D = 3.0$) and E ($M_E = 3.09$) compared to other explanation techniques ($M_A = 3.55$, $M_B = 3.84$, $M_C = 3.46$). The KW test on the questionnaire results of the master's students' replies showed statistically significant differences in their responses on the ESS questionnaire for questionnaire items Q1—related to the understanding of the model ($p < 0.001$, $\chi^2 = 24.338$), Q2—related to the satisfaction with the explanation ($p = 0.003$, $\chi^2 = 20.177$), Q3—related to the sufficiency of detail ($p < 0.001$, $\chi^2 = 31.072$), Q5—related to the use of the explanation ($p = 0.002$, $\chi^2 = 20.591$), and Q7—related to the accuracy of the model ($p = 0.001$, $\chi^2 = 21.593$). The detailed view shows that explanation technique G received the highest mean score ($M_{G1} = 4.35$) in ESS Q1, followed by ($M_{C1} = 4.13$), compared to other observed techniques ($M_{A1} = 3.24$, $M_{B1} = 3.7$, $M_{D1} = 3.18$, $M_{F1} = 3.34$, $M_{H1} = 3.74$). Similarly, explanation G received the highest mean rating in ESS Q2 ($M_{G2} = 4.13$), ESS Q3 ($M_{G3} = 4.22$), ESS Q5 ($M_{G5} = 4$), and ESS Q7 ($M_{G7} = 4.14$). The visualization of mean values by ESS questionnaire items and techniques for both courses is presented in Figures 3 and 4.

Table 2. Differences in ESS questionnaire items between different explanation types with Kruskal–Wallis test results (df = 4).

Course A	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
p	0.258	**0.013**	0.056	0.086	0.343	0.615	0.146	0.068
χ^2	5.294	12.736	9.201	8.157	4.494	2.665	6.818	8.730
Mean	3.46	3.42	3.47	2.62	3.41	2.99	3.19	3.42
SD	1.083	1.029	0.991	1.065	0.918	1.105	1.061	1.029
Course B	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
p	**<0.001**	**0.003**	**<0.001**	0.076	**0.002**	0.287	**0.001**	0.942
χ^2	24.338	20.177	31.072	11.432	20.591	7.379	21.593	1.739
Mean	3.66	3.49	3.38	2.30	3.37	2.87	3.27	3.26
SD	1.143	1.081	1.139	1.059	1.039	1.261	1.181	1.169

The bold in the table indicates significant results of the KW test results.

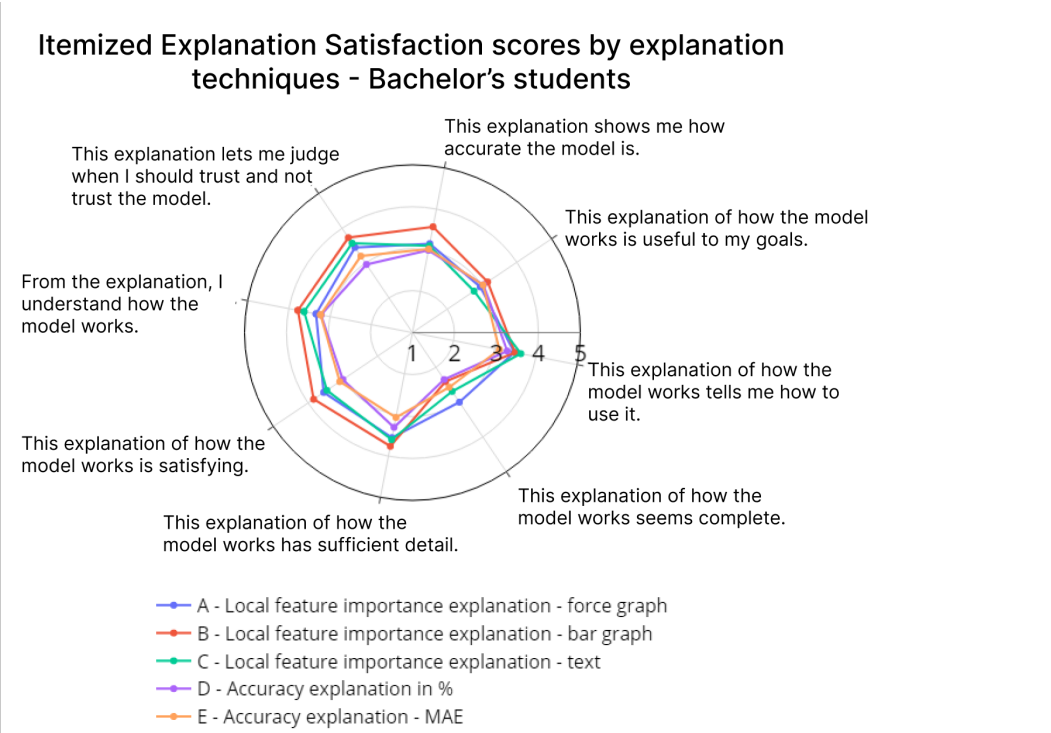


Figure 3. ESS results by explanation techniques on a 5-point Likert Scale—Course A.

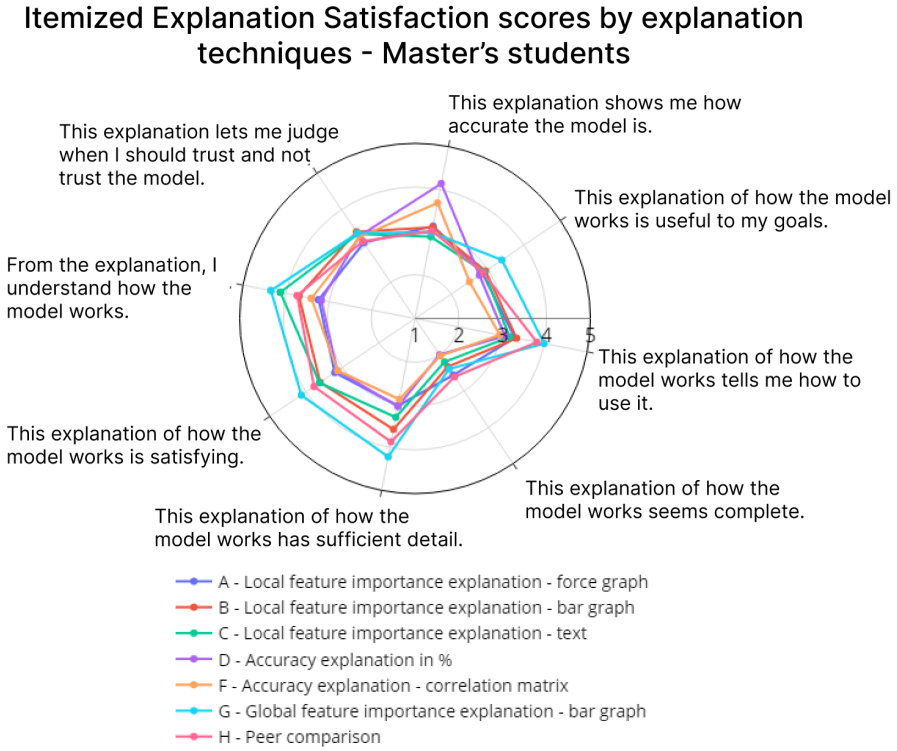


Figure 4. ESS results by explanation techniques on a 5-point Likert Scale—Course B.

4.2. Trust

Analysis of the Trust in Automation questionnaire results from bachelor's students shows that Q7 (The explanation actions will have a harmful or injurious outcome) reached the lowest mean rank on the 7-point Likert scale, with the scale from 1—disagree strongly to 7—I agree strongly. Additionally, the first four items of the TIA questionnaire reached the lowest mean scores (1—the system is deceptive, 2—behaves in an underhanded matter, 3—raises suspicion, and 4—makes users wary of the system). All four mentioned questionnaire items list negative system aspects. Explanation techniques A, B, and C reached the lowest median scores for Q1 (M_{A1} = 2.98 and M_{B1} = 2.70), Q2 (M_{A2} = 3.15 and M_{B2} = 3.18), and Q5 (M_{A5} = 2.94 and M_{B5} = 2.50). Observing positive items on the TIA questionnaire, it is evident that technique B performed best for items 6–11 (measuring confidence, security, integrity, dependence, reliability, and trust). It was followed closely by technique C, with the highest mean score for familiarity (M_{C12} = 4.83). Comparing explanation techniques D and E, which both explain the accuracy of the model, E reached higher mean values with the questionnaire items referring to trust (M_{E11} = 4.49, M_{D11} = 4.17) and reliability (M_{E10} = 4.54, M_{D10} = 4.17), while they reached similar mean values for questionnaire item Q8 related to integrity (M_{E8} = 4.29, M_{D8} = 4.3). Analysis of master's students' questionnaire responses showed a similar trend in lower scores for the first four items, with the lowest mean score for item Q5 (M_{A5} = 2.92 M_{B5} = 2.97, M_{C5} = 2.71, M_{D5} = 3.17, M_{E5} = 3.0), referring to the harmful outcomes of the model's actions. Explanation techniques E and G reached the lowest mean scores for the first four negatively stated questionnaire items. Observing positive questionnaire items, explanation technique B reached the highest, or second-highest, mean values for questionnaire items Q8–Q11 related to integrity, dependency, reliability, and trust (M_{B8} = 4.68, M_{B9} = 4.54, M_{B10} = 4.65, M_{B11} = 4.49). Additionally, explanation technique G, displaying the comparison with peers, also reached comparatively high mean values, as is visible in Figures 5 and 6, where we can observe similarities in mean itemized results gathered with questionnaires for trust between both groups. Overall, master's-level students expressed slightly lower trust ($M_{TIABachelor}$ = 3.92, $M_{TIAMaster}$ = 3.84) compared to bachelor's-level students.

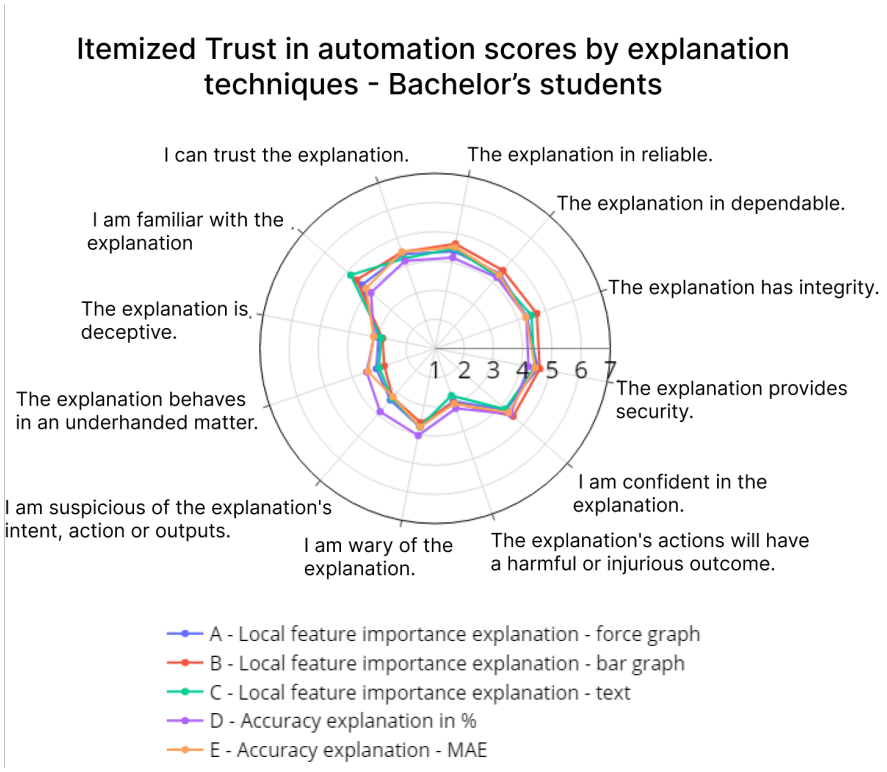


Figure 5. TIA results by explanation technique on a 7-point Likert Scale—Course A.

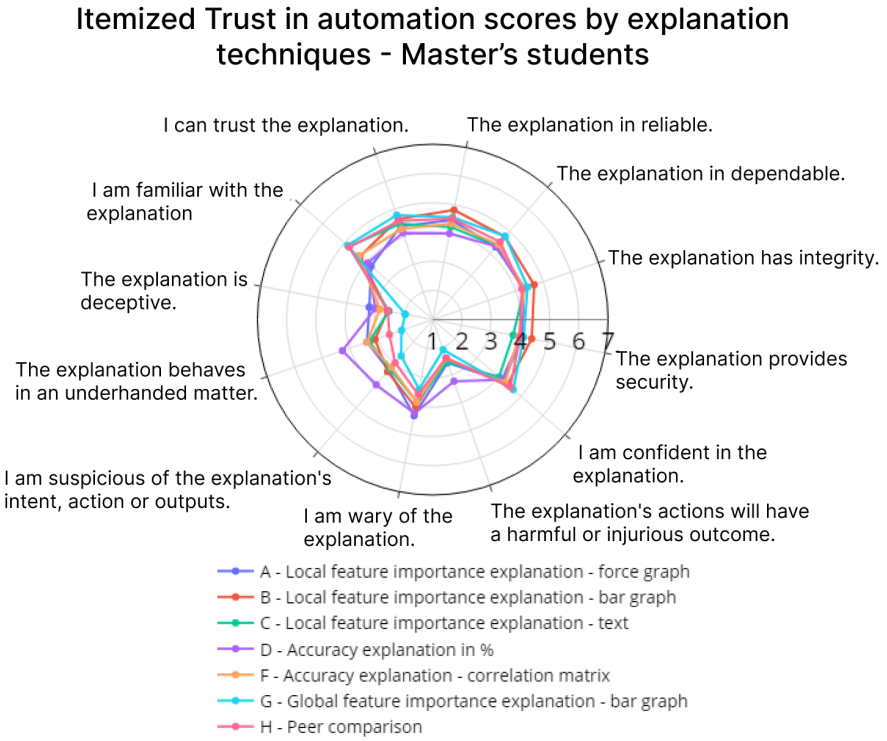


Figure 6. TIA results by explanation technique on a 7-point Likert Scale—Course B.

The hypothesis of the normal distribution of responses for each explanation sample was rejected by the Shapiro–Wilk test in all observed groups for the questionnaires of the bachelor's and master's students, respectively. The KW test was again used to determine if there were statistically significant differences between the TIA questionnaire responses for different explanation techniques. Comparisons were created separately for each course.

The detailed results are presented in Table 3. The KW showed no statistically significant difference in an itemized view of the questionnaire for bachelor's students. A statistically significant difference ($p < 0.001$, $\chi^2 = 23.936$) was observed only in master's students' replies related to item Q2 on the TIA questionnaire, referring to explanations behaving in an underhanded manner. Further observation showed master's students rated explanation D as the most underhanded ($M_{D2} = 4.27$) compared to other explanation techniques ($M_{A2} = 3.38$, $M_{B2} = 3.09$, $M_{C2} = 3.3$, $M_{F2} = 3.38$, $M_{G2} = 2.13$, $M_{H2} = 2.57$), with explanation technique G reaching the lowest mean value.

Table 3. Differences in TIA questionnaire items between different explanation types with Kruskal–Wallis test results (df = 4).

Course A	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
p	0.976	0.462	0.702	0.958	0.958	0.783	0.933	0.738	0.871	0.780	0.949	0.219
χ^2	0.477	3.607	2.183	0.641	0.644	1.741	0.841	1.987	1.243	1.757	0.718	5.749
Mean	2.98	3.25	3.34	3.75	2.94	4.33	4.51	4.42	4.36	4.47	4.38	4.39
SD	1.0506	1.468	1.681	1.611	1.585	1.483	1.367	1.343	1.494	1.418	1.570	1.627
Course B												
p	0.061	**<0.001**	0.205	0.322	0.509	0.769	0.673	0.861	0.800	0.411	0.783	0.180
χ^2	12.059	23.936	8.485	6.990	5.272	3.307	4.024	2.567	3.069	6.109	3.200	8.890
Mean	2.70	3.28	3.23	3.95	2.50	4.30	4.10	4.37	4.48	4.44	4.46	4.38
SD	1.406	1.589	1.665	1.593	1.438	1.442	1.522	1.474	1.490	1.426	1.479	1.753

The bold in the table indicates significant results of the KW test results.

Correlation

A moderate to significant ($p_{Q1–Q2} = 0.71$, $p_{Q1–Q3} = 0.54$, and $p_{Q2–Q3} = 0.69$) positive correlation in ESS items connected to understanding (Q1), satisfaction (Q2), and sufficiency of detail (Q3) can be observed in the responses from course A. Negative questionnaire items from the TIA questionnaire (Q1—explanation is deceptive, Q2—behaved underhandedly, Q3—raises suspicion, Q4—is causing users to vary, and Q5—could produce a harmful outcome) also have a moderate correlation between them ($p_{Q1–Q2} = 0.68$, $p_{Q1–Q3} = 0.72$, $p_{Q1–Q4} = 0.57$, $p_{Q1–Q5} = 0.64$, $p_{Q2–Q3} = 0.62$, $p_{Q2–Q4} = 0.39$, $p_{Q2–Q5} = 0.47$, $p_{Q3–Q4} = 0.61$, $p_{Q3–Q5} = 0.67$, $p_{Q4–Q5} = 0.53$). Positive TIA items (Q6–Q11), with the exception of familiarity (Q12), also showed a high correlation between them, with the correlation strength varying between 0.64 (for $p_{Q7–Q9}$) and 0.81 (for $p_{Q9–Q10}$ and $p_{Q10–Q11}$). Low to moderate correlation can also be observed between the positively stated items in the TIA questionnaire (Q6–Q11, with the exception of familiarity item Q12 and most of the ESS items (Q1–Q3 and Q5–Q8, with completeness being an exception), with its strength varying between 0.27 and 0.63 (for $p_{TIAQ6-ESSQ5}$). The correlation matrices for both courses are presented with a heatmap in Figures 7 and 8. The colour's darkness indicates the correlation's strength, with darker blue shades corresponding to high strength and lighter shades to low correlation.

Similarly, as in course A, we can observe a moderate to significant positive correlation ($p_{Q1–Q2} = 0.78$, $p_{Q1–Q3} = 0.65$, and $p_{Q2–Q3} = 0.81$) in the results from ESS items connected to understanding (Q1), satisfaction (Q2), and sufficiency of detail (Q3) in the data from course B. Negative questionnaire items from the TIA questionnaire (Q1–Q4) also had a high correlation between them in course B ($p_{Q1–Q2} = 0.64$, $p_{Q1–Q3} = 0.51$, $p_{Q1–Q4} = 0.4$, $p_{Q1–Q5} = 0.52$, $p_{Q2–Q3} = 0.66$, $p_{Q2–Q4} = 0.42$, $p_{Q2–Q5} = 0.50$, $p_{Q3–Q4} = 0.52$, $p_{Q3–Q5} = 0.68$, $p_{Q4–Q5} = 0.43$). Positive TIA items, with the exception of familiarity, also showed a moderate to high correlation between them, with correlation strength varying between 0.63 (for $p_{Q6–Q7}$) and 0.86 (for $p_{Q9–Q10}$). Moderate correlation can also be observed between the positively stated items in the TIA questionnaire and most of the ESS items (again, with the exception of the completeness item Q4), with its strength varying between 0.28 and 0.55 (for $p_{TIAQ9-ESSQ5}$ and $p_{TIAQ9-ESSQ8}$).

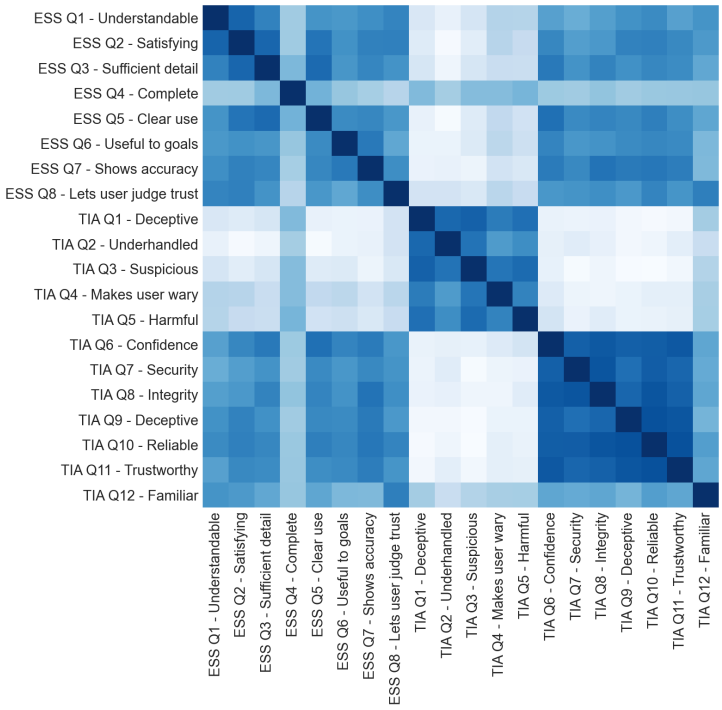


Figure 7. TIA and ESS response correlation—Course A.

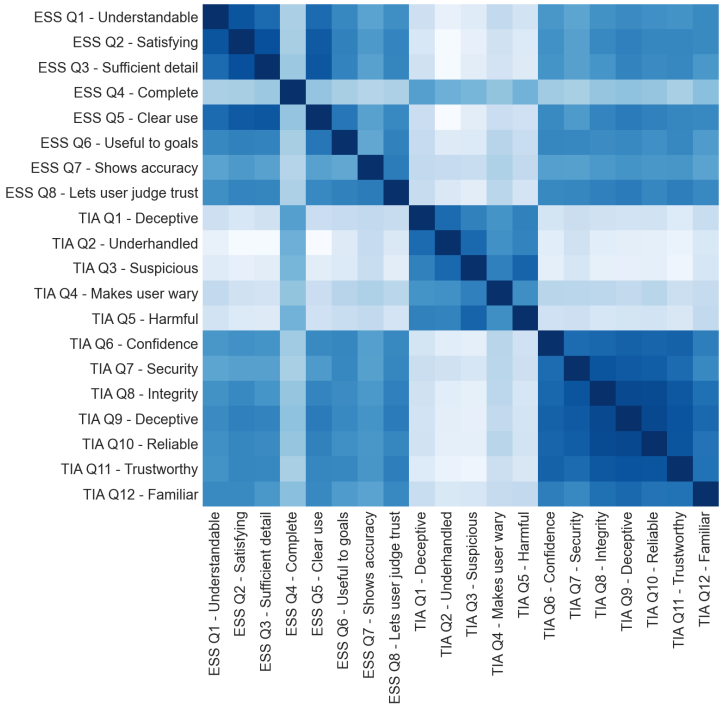


Figure 8. TIA and ESS response correlation—Course B.

5. Discussion and Conclusions

This study aimed to observe the impact of diverse XAI explanation techniques on user trust and satisfaction in the context of XAI-enhanced learning analytics while comparing two groups of STEM college students based on their Bologna study level. Perceived trust and explanation satisfaction was measured with adapted Trust in the Automation [32] questionnaire and Explanation Satisfaction questionnaire [5]. Overall, we found students reported the highest trust in local feature explanation in the bar graph form (explanation technique B). Additionally, master's students, presented with global feature explanation

in bar graph form (explanation technique G), also reported high trust in this form of explanation. The highest measured explanation satisfaction was observed with explanation technique B in the group of bachelor's students, and with explanations B and G in the group of master's students. Comparative analysis of the overlapping explanation techniques presented to both groups indicates two observed groups of students showed consensus in favored explanation techniques when evaluating trust and explanation satisfaction. The detailed overview shows that the master's students evaluated explanation techniques with larger variations, which contributed to more statistically significant differences being observed with the itemized questionnaire comparison of explanation techniques (reported in Tables 2 and 3). We observed a low completeness score in the ESS questionnaire from students of both courses, which can be attributed to the fact that the explanations were shown on separate screens rather than in the form of a smaller section of the complete GUI in order to reduce the impact of other user interface elements on the satisfaction with the explanation. Master's-level students self-reported a higher overall understanding of the explanations (M_{ESS-Q1} = 3.67) compared to bachelor's-level students (M_{ESS-Q1} = 3.46). A similar difference was also observed with their satisfaction with explanations measured with ESS Q2 (M_A = 3.39, M_B = 3.52). Comparing the differences from the TIA questionnaire analysis, bachelor's students overall found explanations more deceptive—TIA Q1 (M_A = 2.99, M_B = 2.67). They perceived the explanations as more harmful—TIA Q5 (M_A = 2.96, M_B = 2.51)—but were overall slightly less wary of them—TIA Q4 (M_A = 3.77, M_B = 3.92)—compared to master's students. We speculated that some of these differences can be explained partially by the knowledge gap; bachelor's students have a low understanding of machine learning, while master's students have, at minimum, taken a mandatory course on it in the semester before the experiment. Additionally, the average age gap of three years between the student groups should be considered, which impacts their maturity.

This study builds on the findings of [37], who previously reported significant differences in effects on trust and satisfaction based on types of explanation. Although different explanation types were compared in an educational domain in our study, we confirm a low to moderate correlation between the measured trust with items of Jian's [32] and Hoffman's [5] questionnaire, especially the correlation between TIA Q9 (The explanation is dependable) and ESS Q5 (This explanation of how the model works tells me how to use it), with the strength of p_A = 0.50 and p_B = 0.55, and TIA Q9 and ESS Q8 (This explanation lets me judge when I should trust and not trust the model), with p_A = 0.41 and p_B = 0.55. We also observed a moderate correlation between TIA Q6 (I am confident in the explanation) and ESS Q6 (This explanation of how the model works is useful to my goals), with p_A = 0.63 and p_B = 0.48. A low to moderate correlation was observed between all the ESS question items, with the exception of Q4 and the positively stated TIA questionnaire items Q6–Q12, as presented in Figures 7 and 8. Furthermore, we confirm the findings of [9], who reported that trust has been found to increase when the reasoning for the AI system's decision is provided (explanation techniques A, B, C, and G) and to decrease when information on sources of uncertainty is shared with the user (explanation techniques D, E, and F scored comparatively lower on the TIA questionnaire with both groups of students). Comparison of the ESS results in the group of master's students, who were presented with global and local explanations, also confirms the findings of [41] in our study setting, where higher user satisfaction was reported with global explanation G compared to local feature explanations (A–C).

This study supports the findings indicated in [37], which reported significant differences in the effects of different types of XAI explanations on the first-time measured trust with the TIA trust scale. Comparing the user perception of local and global explanations (limited to course B), our results indicate that local explanations (A–C) received lower mean results on the negatively stated items of the TIA questionnaire compared to the global explanation (F). Observing the positively stated TIA items, the global model performed higher on the measured trust level with items Q7, Q9, and Q10, related to the security of

the explanation, its dependency, and its reliability. These results are similar to the findings of [41], where global SHAP explanations were evaluated with a slightly higher perceived trust on a simple one-scale measure compared to local SHAP explanations. The trust evaluation method utilized in our study allowed us to evaluate the initial user trust in a more detailed view. As no control group was included in our study, our findings are not directly comparable to [28], who reported no difference in trust between students who observed any of the two explanations and those in the control group. The results of this study have contradicted the existing literature on LA systems about the recognized effect that explanations have on user trust. This could be explained with the use of global explanations in their study or with changes in the context of their use case, which used automated essay scoring instead of the prediction of the final course grade or an intelligent tutoring system.

Limitations and Future Work

This study was conducted on a limited sample of Slovenian students in one selected faculty, all of whom had had at least basic prior training in the field of Information Technologies and Computer Science. The small sample size is a recognized limitation of this study, which may affect the generalizability of our findings to a larger population. Additional large-scale validation should be a part of future work. The models used in the predictions for course A and course B utilized notably different features due to the limited access to data, which might also have had an impact on students' understanding and trust in the models. Due to this, some differences occurred in the explanations shown to the bachelor's level and master's level students. Furthermore, a limited set of explanation techniques was used in this comparative study. In this report, we do not address the users' mental models or their understanding of explanations, which would offer additional insight into explanation evaluation. Researching these aspects presents an important part of future work. The psychological and academic impact of presenting students with their final grade prediction at the beginning of the spring semester (course B) and in the middle of the winter semester (course A) was not a goal of this study, although these effects should be explored in the future as well. This paper does not include further observations of the impacts the presented learning analytics dashboard had on students' self-regulated learning. Additional fragmentation of global model explanations and peer comparison should be conducted in order to further explore the cause of high trust and explanation satisfaction compared to other observed explanation techniques. Trust in this study was analyzed in a one-time experiment when users first interacted with the learning analytics dashboard. Differences in trust levels should be measured over time in future studies.

Author Contributions: This paper is a collaborative work by all the authors. S.B. proposed the idea, worked on the conceptualization, designed and created the software, executed the experiments, extracted data, conducted the data analysis, prepared visualizations, and wrote the manuscript. B.Š. worked on the conceptualization, supervised the software development, experiments, and formal analysis, and wrote, reviewed, and edited the manuscript. V.P. worked on the conceptualization, supervised the predictive model development, and reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057).

Institutional Review Board Statement: The Research Ethics Committee of the Faculty of Arts from the University of Maribor was consulted before the experiments were conducted.

Informed Consent Statement: Informed consent was obtained from all students involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
LA	Learning Analytics
KW	Kruskal–Wallis H test
ESS	Explanation Satisfaction Scale (questionnaire)
TIA	Trust in Automation (questionnaire)
MAE	Mean Absolute Error

Appendix A. Overview of Explanation Techniques

Table A1. Detailed overview of evaluated explanation techniques.

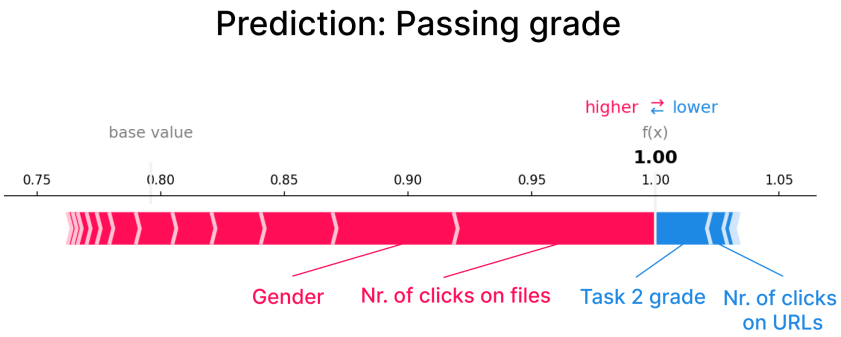
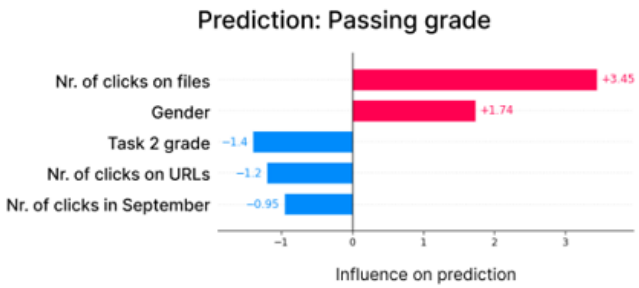
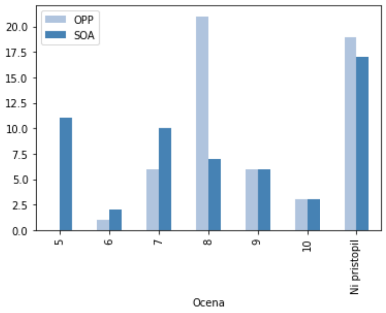
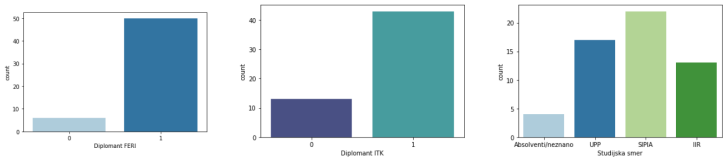
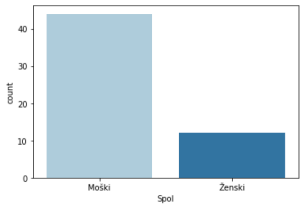
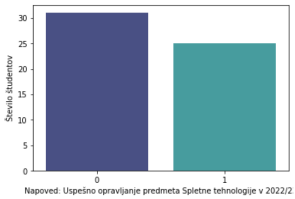
ID	Description	Example
A	Local feature explanation with force graph	
B	Local feature explanation with bar graph	

Table A1. *Cont.*

ID	Description	Example
C	Local feature explanation with text	Prediction: Passing grade The prediction was positively affected by: • Your grade of the Quiz 1. • Daily number of interactions with virtual learning environment. The prediction was negatively affected by: • Your grade of the Task 3.
D	Accuracy explanation with error margin in percentage	MODEL ACCURACY: 90% ? The classification accuracy is moderately high. It is based on historical data, and 90% accuracy means that 10% of the predictions were wrong, or that the model mispredicted the outcome of about three students in a generation. The classification is slightly more accurate (91%) in identifying students predicted to pass the course, compared to the accuracy of identifying
E	Accuracy explanation with mean absolute error	61% WITH MEAN ABSOLUTE ERROR OF THE MODEL 50% - 72%
F	Accuracy explanation with confusion matrix	Predicted values neg. pos. Actual values neg. 16 1 pos. 2 11
G	Global feature importance explanation	Course X grade +0.35 Course Y grade +0.13 Bachelors programme +0.03 Study module +0.03 Gender F +0.02 Gender M +0.02 Faculty of bachelors studies +0.01 Graduation year +0.01 VLE access in first semester days +0.01 Number of VLE accesses in the first semster days +0 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 Global feature importance

Table A1. *Cont.*

ID	Description	Example
H	Simple peer comparison and sample overview [1]	**Pregled vzorca in primerjava z ostalimi študenti** Na spodnji sliki je prikazana statistika ocen izpitnih rokov v študijskem letu 22/23 pri predmetih Storitveno usmerjene arhitekture in Optimizacija poslovnih procesov. Značilki imata velik vpliv na napovedano uspešnost pri predmetu Spletne tehnologije. Opomba: Postavka "Ni pristopil" vključuje tudi študente, ki so izpit morda že opravili lani (tj. ga niso opravili v letošnjem študijskem letu).  Na spodnjih slikah je prikazana razdelitev vzorca letošnjih študentov glede na izobraževalno institucijo na kateri so pridobili diplomo (FERI ali druga fakulteta) ter na dodiplomsko študijsko smer (ITK FERI ali druga smer). Zadnja slika prikazuje razdelitev vzorca glede na izbran modul v prvem letniku magistrskega študija.  Na spodnji sliki je prikazana delitev vzorca glede na spol. Značilka spol ima v globalnem modelu pri napovedih uspeha v začetku semestra zmerno pomembnost.  Na spodnji sliki je predstavljena pregledna statistika rezultatov napovedi celotnega vzorca. Uspešno opravljanje predmeta v študijskem letu 2022/23 (poletni in jesenski rok) je bilo napovedano za 25 študentov, medtem ko je bilo neuspešno opravljanje predmeta napovedano za preostalih 31 študentov. Glede na model je napovedana prehodnost 45%, kar je nekoliko nižje kot v lanskem študijskem letu, ko je bila prehodnost pri tem predmetu 51%. 

[1] Explanation H includes content in the Slovenian language, as this was the language used in the experiment.

Appendix B. Questionnaires and Their Translations

Table A2. Adaptation and translation of Trust in Automation Questionnaire [32].

Item	Adaptation	Slovene Translation
1	The explanation is deceptive.	Razlaga je zavajajoča.
2	The explanation behaves in an underhanded matter.	Razlaga se obnaša pomanjkljivo.
3	I am suspicious of the explanation's intent, action, or outputs.	Sumničav sem glede namena, delovanja ali rezultatov razlage.
4	I am wary of the explanation.	Previden sem do razlage.
5	The explanation's actions will have a harmful or injurious outcome.	Delovanje razlage bo imelo škodljive posledice.
6	I am confident in the explanation.	Zaupam razlagi.
7	The explanation provides security.	Razlaga zagotavlja varnost.
8	The explanation has integrity.	Razlaga ima integriteto.
9	The explanation is dependable.	Razlaga je zanesljiva.
10	The explanation is reliable.	Razlaga je verodostojna.
11	I can trust the explanation.	Lahko zaupam razlagi.
12	I am familiar with the explanation.	Razlaga mi je poznana.

Table A3. Adaptation and translation of Explanation Satisfaction Questionnaire [5].

Item	Adaptation	Slovene Translation
1	From the explanation, I understand how the model works.	Razumem razlago delovanja napovednega modela.
2	This explanation of how the model works is satisfying.	Razlaga delovanja modela je bila zadovoljiva.
3	This explanation of how the model works has sufficient detail.	Razlaga delovanja modela je bila dovolj podrobna.
4	This explanation of how the model works seems complete.	Razlaga delovanja modela vključuje nepotrebne podrobnosti.
5	This explanation of how the model works tells me how to use it.	Razlaga modela je bila celostna.
6	This explanation of how the model works is useful to my goals.	Razlaga delovanja modela je uporabna za moje cilje.
7	This explanation shows me how accurate the model is.	Razlaga poda informacijo natančnosti modela.
8	This explanation lets me judge when I should trust and not trust the model.	Razlaga mi omogoča, da lahko sam presodim ali naj modelu zaupam ali ne.

References

1. Bahalul Haque, A.K.M.; Najmul Islam, A.K.M.; Patrick Mikalef, P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technol. Forecast. Soc. Chang.* **2023**, *186*, 122120. [CrossRef]
2. Alamri, R.; Alharbi, B. Explainable Student Performance Prediction Models: A Systematic Review. *IEEE Access* **2021**, *9*, 33132–33143. [CrossRef]
3. Rachha, A.; Seyam, M. Explainable AI In Education: Current Trends, Challenges, And Opportunities. In Proceedings of the SoutheastCon 2023, Orlando, FL, USA, 13–16 April 2023; pp. 232–239. [CrossRef]
4. Anjomshoae, S.; Najjar, A.; Calvaresi, D.; Främling, K. Explainable Agents and Robots: Results from a Systematic Literature Review. In Proceedings of the AAMAS '19: 18th International Conference on Autonomous Agents and MultiAgent Systems, Richland, SC, USA, 13–17 May 2019; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 13–17 May 2019; pp. 1078–1088.
5. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects. *arXiv* **2018**, arXiv:1812.04608.
6. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* **2019**, *40*, 44–58.
7. Kulesza, T.; Burnett, M.; Wong, W.K.; Stumpf, S. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the IUI '15: 20th International Conference on Intelligent User Interfaces, New York, NY, USA, 29 March–1 April 2015; Association for Computing Machinery: New York, NY, USA; pp. 126–137. [CrossRef]
8. Kraus, S.; Azaria, A.; Fiosina, J.; Greve, M.; Hazon, N.; Kolbe, L.; Lembcke, T.B.; Muller, J.P.; Schleibaum, S.; Vollrath, M. AI for explaining decisions in multi-agent environments. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13534–13538. [CrossRef]
9. Vössing, M.; Kühl, N.; Lind, M.; Satzger, G. Designing Transparency for Effective Human-AI Collaboration. *Inf. Syst. Front.* **2022**, *24*, 877–895. [CrossRef]
10. Ooge, J.; Kato, S.; Verbert, K. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In Proceedings of the IUI '22: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, 22–25 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 93–105. [CrossRef]
11. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
12. Khosravi, H.; Shum, S.B.; Chen, G.; Conati, C.; Tsai, Y.S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; Gašević, D. Explainable Artificial Intelligence in education. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100074. [CrossRef]

13. Lim, B.Y.; Dey, A.K. Toolkit to Support Intelligibility in Context-Aware Applications. In Proceedings of the UbiComp '10: 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark, 26–29 September 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 13–22. [CrossRef]
14. Mohseni, S.; Zarei, N.; Ragan, E.D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 24. [CrossRef]
15. Liao, Q.V.; Varshney, K.R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv* **2022**, arXiv:2110.10790.
16. Conati, C.; Barral, O.; Putnam, V.; Rieger, L. Toward personalized XAI: A case study in intelligent tutoring systems. *Artif. Intell.* **2021**, *298*, 103503. [CrossRef]
17. Schwendimann, B.A.; Rodríguez-Triana, M.J.; Vozniuk, A.; Prieto, L.P.; Boroujeni, M.S.; Holzer, A.; Gillet, D.; Dillenbourg, P. Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Trans. Learn. Technol.* **2017**, *10*, 30–41. [CrossRef]
18. Jivet, I.; Scheffel, M.; Specht, M.; Drachsler, H. License to Evaluate: Preparing Learning Analytics Dashboards for Educational Practice. In Proceedings of the LAK '18: 8th International Conference on Learning Analytics and Knowledge, Sydney, Australia, 7–9 March 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 31–40. [CrossRef]
19. Clow, D. An overview of learning analytics. *Teach. High. Educ.* **2013**, *18*, 683–695. [CrossRef]
20. Mathrani, A.; Susnjak, T.; Ramaswami, G.; Barczak, A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Comput. Educ. Open* **2021**, *2*, 100060. [CrossRef]
21. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
22. Zhang, K.; Aslan, A.B. AI technologies for education: Recent research & future directions. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100025. [CrossRef]
23. Wang, X.; Guo, B.; Shen, Y. Predicting the At-Risk Online Students Based on the Click Data Distribution Characteristics. *Sci. Program.* **2022**, *2022*, 9938260. [CrossRef]
24. Kuzilek, J.; Hlosta, M.; Herrmannova, D.; Zdráhal, Z.; Wolff, A. OU Analyse: Analysing at-risk students at The Open University. *Learn. Anal. Rev.* **2015**, *LAK15-1*, 1–16.
25. Al-Azawei, A.; Al-Masoudy, M. Predicting Learners' Performance in Virtual Learning Environment (VLE) based on Demographic, Behavioral and Engagement Antecedents. *Int. J. Emerg. Technol. Learn.* **2020**, *15*, 60–75. [CrossRef]
26. Chitti, M.; Chitti, P.; Jayabalan, M. Need for Interpretable Student Performance Prediction. In Proceedings of the 2020 13th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, UK, 14–17 December 2020; pp. 269–272. [CrossRef]
27. Alonso, J.M.; Casalino, G. Explainable Artificial Intelligence for Human-Centric Data Analysis in Virtual Learning Environments. In *Higher Education Learning Methodologies and Technologies Online*; Burgos, D., Cimitile, M., Ducange, P., Pecori, R., Picerno, P., Raviolo, P., Stracke, C.M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 125–138.
28. Conijn, R.; Kahr, P.; Snijders, C. The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *J. Learn. Anal.* **2023**, *10*, 37–53. [CrossRef]
29. Shulner-Tal, A.; Kuflik, T.; Kliger, D. Fairness, Explainability and in-between: Understanding the Impact of Different Explanation Methods on Non-Expert Users' Perceptions of Fairness toward an Algorithmic System. *Ethics Inf. Technol.* **2022**, *24*, 2. [CrossRef]
30. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef]
31. Vereschak, O.; Bailly, G.; Caramiaux, B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–39. [CrossRef]
32. Jian, J.Y.; Bisantz, A.M.; Drury, C.G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *Int. J. Cogn. Ergon.* **2000**, *4*, 53–71. [CrossRef]
33. Chien, S.Y.; Lewis, M.; Sycara, K.; Liu, J.S.; Kumru, A. The Effect of Culture on Trust in Automation: Reliability and Workload. *Acm Trans. Interact. Intell. Syst.* **2018**, *8*, 1–31. [CrossRef]
34. Merritt, S.M. Affective Processes in Human–Automation Interactions. *Hum. Factors* **2011**, *53*, 356–370. [CrossRef] [PubMed]
35. Muir, B. Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task. Ph.D. Thesis, University of Toronto: Toronto, ON, Canada, 1989.
36. Benbasat, I.; Wang, W. Trust in and adoption of online recommendation agents. *J. Assoc. Inf. Syst.* **2005**, *6*, 4. [CrossRef]
37. Schraagen, J.M.; Elsasser, P.; Fricke, H.; Hof, M.; Ragalmuto, F. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2020**, *64*, 339–343. [CrossRef]
38. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:1602.04938.
39. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
40. Belle, V.; Papantonis, I. Principles and practice of explainable machine learning. *Front. Big Data* **2021**, *4*, 39. [CrossRef]

41. Aechtner, J.; Cabrera, L.; Katwal, D.; Onghena, P.; Valenzuela, D.P.; Wilbik, A. Comparing User Perception of Explanations Developed with XAI Methods. In Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 18–23 July 2022; pp. 1–7. [CrossRef]
42. Aykin, N.M.; Aykin, T. Individual differences in human-computer interaction. *Comput. Ind. Eng.* **1991**, *20*, 373–379. [CrossRef]
43. *ISO 9241-1:1997*; Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). International Organization for Standardization: Geneva, Switzerland, 1997.
44. European Commission, Directorate-General for Education, Youth, Sport and Culture. *ECTS Users' Guide 2015*; Publications Office of the European Union: Luxembourg, 2017. [CrossRef]
45. Brooke, J. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
46. You, J.W. Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet High. Educ.* **2016**, *29*, 23–30. [CrossRef]
47. Rivas, A.; González-Briones, A.; Hernández, G.; Prieto, J.; Chamoso, P. Artificial neural network analysis of the academic performance of students in virtual learning environments. *Neurocomputing* **2021**, *423*, 713–720. [CrossRef]
48. Schoonderwoerd, T.A.J.; Jorritsma, W.; Neerincx, M.A.; van den Bosch, K. Human-Centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *Int. J. Hum.-Comput. Stud.* **2021**, *154*, 102684. [CrossRef]
49. Poulin, B.; Eisner, R.; Szafron, D.; Lu, P.; Greiner, R.; Wishart, D.; Fyshe, A.; Pearcy, B.; MacDonell, C.; Anvik, J. Visual explanation of evidence in additive classifiers. *Proc. Natl. Conf. Artif. Intell.* **2006**, *2*, 1822–1829.
50. Ramaswami, G.; Susnjak, T.; Mathrani, A. Capitalizing on Learning Analytics Dashboard for Maximizing Student Outcomes. In Proceedings of the 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Melbourne, Australia, 9–11 December 2019; pp. 1–6.
51. Aljohani, N.R.; Daud, A.; Abbasi, R.A.; Alowibdi, J.S.; Basheri, M.; Aslam, M.A. An integrated framework for course adapted student learning analytics dashboard. *Comput. Hum. Behav.* **2019**, *92*, 679–690. [CrossRef]