big data and cognitive computing

MDPI

Article

Future-Ready Strategic Oversight of Multiple Artificial Superintelligence-Enabled Adaptive Learning Systems via Human-Centric Explainable AI-Empowered Predictive Optimizations of Educational Outcomes

Meng-Leong HOW

National Institute of Education, Nanyang Technological University Singapore, Singapore 639798, Singapore; mengleong.how@nie.edu.sg

Received: 31 May 2019; Accepted: 23 July 2019; Published: 31 July 2019

Abstract: Artificial intelligence-enabled adaptive learning systems (AI-ALS) have been increasingly utilized in education. Schools are usually afforded the freedom to deploy the AI-ALS that they prefer. However, even before artificial intelligence autonomously develops into artificial superintelligence in the future, it would be remiss to entirely leave the students to the AI-ALS without any independent oversight of the potential issues. For example, if the students score well in formative assessments within the AI-ALS but subsequently perform badly in paper-based post-tests, or if the relentless algorithm of a particular AI-ALS is suspected of causing undue stress for the students, they should be addressed by educational stakeholders. Policy makers and educational stakeholders should collaborate to analyze the data from multiple AI-ALS deployed in different schools to achieve strategic oversight. The current paper provides exemplars to illustrate how this future-ready strategic oversight could be implemented using an artificial intelligence-based Bayesian network software to analyze the data from five dissimilar AI-ALS, each deployed in a different school. Besides using descriptive analytics to reveal potential issues experienced by students within each AI-ALS, this human-centric AI-empowered approach also enables explainable predictive analytics of the students' learning outcomes in paper-based summative assessments after training is completed in each AI-ALS.

Keywords: future-ready; strategic oversight; artificial superintelligence; artificial intelligence; forecasting AI behavior; predictive optimization; simulations; Bayesian networks; adaptive learning systems; pedagogical motif; explainable AI; AI Thinking; human-in-the-loop; human-centric reasoning; policy making on AI

1. Introduction

Artificial intelligence (AI) [1] refers to the ability of human-made systems to mimic rudimentary human thought. The term "artificial superintelligence" [2] goes beyond this primary ability of AI; it refers to the capability of human-made systems that can surpass humans. For example, they might even be able to rapidly discover hidden motifs or patterns in the data and then make predictions, while humans might find it very challenging to apperceive these hidden patterns within the mind, or perform similar feats at the speeds and performance levels that these systems can. To be clear, it could be argued that an AI system does not care about the need to prove to humans that it has achieved human-like consciousness (also referred to as the state of "singularity" or "artificial general intelligence") in order to be validated, certified, or given the stamp of approval by humans, so that it can properly be accorded a definitional label of its level of AI. There would probably be no notifications from AI systems the day they autonomously become self-aware, regardless of whether humans like it or not.

Meanwhile, in lieu of that fateful day, researchers have observed in studies that we already have artificial superintelligence working inconspicuously and tirelessly in our midst [3–5]. In the field of education, since the 1950s, AI deployed in the form of adaptive learning systems (ALS) [6,7], which are contemporary forms of intelligent tutoring systems (ITS) [8], have been utilized to assist teachers in the training of students [9]. Great strides have been made by researchers and commercial companies toward creating ALS that are powered by artificial intelligence, and perhaps, even superintelligence [2], in the sense that some of them have—dare I say—already surpassed the human teacher in terms of the ability to relentlessly perform the task of one-to-one tutoring, initiate progress checks, and conduct remediation. They can concurrently perform these tasks, perpetually to an unlimited number of students, round the clock, whenever and wherever the students choose to learn [10]. The developers of ALS and the researchers who field-test them have often lauded improvements in learning gains, and efficiencies of learning similar amounts of subject content in reduced amounts of time [11]. The primary function of an ALS is to educe (draw out) the learning abilities of the students by making them solve problems [12].

The advent of AI has enabled advanced developments of ALS. In recent years, an artificial intelligence-enabled adaptive learning system (AI-ALS) might utilize, for example, a variant of the AI-based Bayesian Knowledge Tracing (BKT) [13] algorithm, or some other proprietary algorithms formed from an ensemble of multiple AI-based methods to make "adjustments in an educational environment in order to accommodate individual differences" to provide a personalized learning experience for each student [14]. An example of a procedure that an AI-ALS might use to interact with the student is: (1) present the student with a topic or sub-topic to learn, (2) present the student with learning material that illustrate the concepts, (3) initiate a short progress check quiz of each sub-topic for the student. If the student could consecutively correctly answer a few questions, the AI-ALS would deem that the student has "passed" the learning objective for that topic or sub-topic (which will be indicated as "topic_passed" in the dataset). Otherwise, the student would be remediated by the AI-ALS until the learning outcome is achieved, and (4) finally, after the student has passed the progress check quiz, the AI-ALS would unlock more topics or sub-topics that are considered to be "ready for learning" by the student (that will be indicated as "topic_ready_for_learning" in the dataset). The AI-ALS is often used in conjunction with the flipped learning pedagogy [15], where the students are expected to log into the AI-ALS and learn as much as they can on their own at home. Subsequently, when they are in the classroom, the teacher can spend the precious class time more effectively by helping students to address any learning issues that they might have.

The current paper does not purport to be an empirical study of the effectiveness of any current AI-ALS. Rather, it proffers a future-ready human-in-the-loop [16] analytical framework that is based upon intuitive human-centric probabilistic reasoning, which could be used to characterize the "pedagogical motifs" [17] of any number of AI-ALS that may be deployed in the future. So long as the data from those systems are available to human analysts, this framework would still be useful for education stakeholders to gain an oversight of the "timbre" of multiple AI-ALS that are deployed in schools, even if those AI-ALS in the future are artificially superintelligent.

2. Research Problem and Initial Hypothetical Conjecture

2.1. Research Problem

In reality, the Department of Education of a city or a state or a country might choose not to implement a policy that compels all of the schools to use one single AI-ALS that is provided by one vendor. Presumably, the schools would also rather have the freedom to choose the AI-ALS that they prefer to deploy for their students. However, it would be remiss if the students were entirely left to the AI-ALS. For example, if the students do very well in the formative assessment tests in the AI-ALS, but perform badly in the paper-based post-test, or if the relentless testing-checking-remediating-testing algorithm of a particular AI-ALS is suspected to be causing too much stress for the students, it would

be of concern to educational stakeholders. Currently, the AI-ALS products available in the educational industry have the ability to autonomously strive to make the student achieve mastery of the topics that they are required to learn. However, they are not yet fully equipped (e.g., with sensors or by other means) to take noncognitive factors (e.g., ability to manage stress, psychological well-being, motivation, level of engagement, etc.) of the students into consideration [18]. This is where a human-in-the-loop approach that is proffered by the current paper would play a vital role in bridging the gaps. It can be used to inform educational stakeholders in areas where the developers of the AI-ALS might have overlooked.

Coordination efforts between the educational stakeholders, such as policy makers, school leaders, and teachers, to assess the risks and safeguard the safety of students who are using the AI-ALS (in terms of noncognitive factors [19–23], such as, for example, the psychological well-being, or emotional intelligence to manage stress) are, undeniably, of paramount importance. Researchers, such as Manheim [24], Perry and Uuk [25], Turchin, Denkenberger, and Green [26], Umbrello [27], Watson [28], and by Ziesche and Yampolskiy [29], have made efforts to analyze the issues, values, and benefits of strategies and coordination in artificial superintelligence. Yet, in the field of education, there is still a dearth in the extant literature regarding the area of coordination and safety in artificial superintelligence [30]. From the perspective of education policy makers, it would be interesting to help to coordinate the analysis of data from multiple AI-ALS deployed in different schools, so they would be able to "see the big picture" and assess the potential issues to know whether each AI-ALS in the respective school is helping (or not helping) the students, and take further steps to address problems if necessary. Human teachers would be able to address the gaps in the students' learning process where the AI-ALS could not, and help to alleviate stressful situations for the students if they are uncomfortable using the AI-ALS.

In the field of education, the question "would it be possible to predict the conditions during the use of an educational intervention (e.g., an AI system) to enhance optimal student performance in the paper-based summative assessments?" might intrigue educational stakeholders, such as policy makers, parents, students, and educational researchers [31,32]. However, to the authors' knowledge, it is beyond the scope of consideration by the developers of the AI-ALS to predict how the students' scores within the AI-ALS could influence their learning outcomes in a summative assessment (e.g., a paper-based standardized test that all the students are required to take in the school) after their training has been completed in the AI-ALS. To achieve this predictive capability, it is imperative for the pedagogical "motif" or "timbre" or "disposition" of the AI-ALS to be known, as each of them would interact with students in different ways. Although educational stakeholders need to examine the pedagogical characteristics of the AI-ALS, the vendors of the systems would understandably be reticent about divulging the exact algorithm to the customers, as they are closely-held trade secrets. Instead of believing all of the information provided by the vendors who are inclined to assure that everything will be excellent, it would be prudent for educational stakeholders to independently investigate the pedagogical characteristics that underlie these AI-ALS. Frameworks have been created by researchers for the evaluation of ALS [33]. Nevertheless, those laudable techniques were often formally presented as mathematical equations, which could prove to be difficult for educational stakeholders who might not have the necessary computer programming human resources or enough time to implement them. There remains a need for a more intuitive and practical way for educational stakeholders—rather than computer scientists—to apply human-in-the-loop AI-Thinking [34,35] and quickly achieve a strategic oversight of the multiple AI-ALS, which is crucial for informing educational policy and advancing pedagogical practice.

2.2. Initial Hypothetical Conjecture

The initial hypothetical conjecture assumes that the developers of an AI-ALS might have designed it to push the higher-performing students a little harder, and conversely, to go easy on the relatively lower-performing students. Therefore, it would not be unreasonable to imagine that a student who

had performed poorly in the AI-ALS might have experienced having his or her weaknesses being educed (drawn out) by the system. Subsequently, after a personal reflection of those problems via vicarious trial and error (VTE) [36], the student could become cognizant of those weaknesses and could avoid similar predicaments during problem-solving in the paper-based post-test. Conversely, a student who had performed well in the AI-ALS might not have experienced having his or her weaknesses educed, and hence might lack the personal reflections or the VTE to learn from those experiences. Consequently, he or she might perform poorly in the post-test. The approach being proffered in the current paper would purely characterize its informational pattern (its motif), regardless of whether a student scored high or low within the AI-ALS. In other words, it does not affect the calculation of the "gains" that are attributed to the prowess of the AI-ALS, as it will not simply be a subtraction of the results of the paper-based post-test from the paper-based pre-test.

Nevertheless, it would be contrived to only measure the "gains" in terms of cognitive dimensions while using the pre-test and post-test, as there might be noncognitive benefits for the students too. Hence, a survey that could be used to understand more about the noncognitive aspects of their learning experiences could also be administered to the students upon the completion of their learning process in the AI-ALS. Some of the possible noncognitive instruments that could be utilized by educational stakeholders include those that are offered by researchers such as Al-Mutawah and Fateel [37], Chamberlin, Moore, and Parks [38], Egalite, Mills, and Greene [39], Lipnevich, MacCann, and Roberts [40], and Mantzicopoulos, Patrick, Strati, and Watson [41].

2.3. Potential Issues that Education Researchers Might Encounter

When a school decides to let a class of students use an AI-ALS to assist the teachers, it might not occur to the school leaders or teachers to make any arrangements for the formation of a control group. Understandably, the school may have concerns that parents might be unwilling to give permission for their children to participate in a control group, merely to form a baseline group for comparison with the treatment group, with no assistive benefits from any educational technology. Moreover, it will not be easy to perform direct comparisons between the treatment and control group even if a control group could be formed by the school, as the teaching experiences and skills of the teachers between the control and the treatment group might be unevenly matched. Further, it might not be surprising if some students from the treatment group or control group have the advantage of receiving extra help from tuition lessons outside of school. In effect, the myriad potential confounding factors would be difficult to account for, if fair comparisons must be performed between the treatment group that attended lessons where the teacher had been assisted by the AI-ALS to learn mathematics, and the control group that attended lessons where the teacher had not been assisted by the AI-ALS. Last but not least, a major problem that is faced by analysts who are considering the use of null hypothesis significance testing (NHST) frequentist approaches is that there might not be results that yield any meaningful statistical significant difference, due to the low number of participants in real-world situations (e.g., 20 students per class in each school) and the corresponding non-parametric data distributions [42].

Practical examples will be provided in the current paper to overcome these constraints. They will be used to illustrate how strategic oversight could be implemented using an artificial intelligence-based analytical tool by educational stakeholders to analyze data from five dissimilar AI-ALS deployed in small-scale pilot studies, each in a different school, and how conditions in those different AI-ALS could be used for predictive optimizations of educational outcomes in the paper-based summative assessments.

3. Methods

3.1. Rationale for Using the Bayesian Approach for Human-Centric Probabilistic Reasoning

Bayesian approaches for analyzing statistical data [43] have gained traction in behavioral science research in recent years [44]. The Bayesian network (BN) [45–47] approach is suitable for analyzing non-parametric data from a small number of participants, because it does not require the underlying

variables of a model to assume or have a normal parametric distribution [42,48,49]. The Bayesian paradigm enables researchers to perform hypothesis testing by including prior knowledge into the analyses. Due to this capability, it becomes unnecessary to repeatedly perform multiple rounds of null hypothesis testing [50–52] when using Bayesian data analytical techniques.

Researchers in education, such as Kaplan [53], Levy [54], Mathys [55], and Muthén and Asparouhov [56], have employed the Bayesian approach to model the behavior of pedagogical systems operating under conditions with uncertainties, as the information about entropy in these systems could be harnessed to understand more about the factors that contribute (either positively or not) to their robustness and resiliency [57]. In educational technology, Bekele and McPherson [58] and Millán, Agosta, and Cruz [59] have also utilized the Bayesian approach, because it enables them to measure information gain, as depicted in Claude Shannon's Information Theory [60], which could be likened to the notion of learning by the students.

The primary advantage of BN is that its strong probabilistic theory empowers users to gain an intuitive understanding of the processes involved. It also enables predictive reasoning because, given observations of evidence, questions can be posed to find the posterior probability of any variable or set of variables. However, the current paper does not purport to perform comparisons between the use of BN and other AI-based techniques, such as artificial neural networks (ANN), as that has already been well-documented by Correa, Bielza, and Pamies-Teixeira [61]. They observe that BN can illustrate the relationships that exist between the nodes in a model to provide more information than an ANN, which has been likened to a black box.

3.2. The Bayesian Theorem

A succinct introduction to the Bayesian theorem and BN will be presented here. However, readers who are interested to learn more about BN are encouraged to peruse the works of Cowell, Dawid, Lauritzen, and Spiegelhalter [62]; Jensen [63]; and, Korb & Nicholson [64].

The mathematical theorem (see Equation (1)) for human-centric probabilistic reasoning was developed by the mathematician and theologian, Reverend Thomas Bayes, but he passed away and the notes were left unpublished in his drawer. They were later found and published posthumously by his friend Richard Price in 1763 [43].

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \quad (1)$$

According to Equation (1), H represents a hypothesis and E represents a piece of evidence. $P(H|E)$ is referred to as the conditional probability of the hypothesis H, which means the likelihood of H occurring given the condition that the evidence E is true. It is also referred to as the posterior probability, which means the probability of the hypothesis H being true after calculating how the evidence E influences the verity of the hypothesis H.

$P(H)$ and $P(E)$ represent the probabilities of the likelihood of the hypothesis H being true, and of the likelihood of the evidence E being true, independent of each other, and it is referred to as the prior or marginal probability—$P(H)$ and $P(E)$, respectively. $P(E|H)$ represents the conditional probability of the evidence E, that is, the likelihood of E being true, given the condition that the hypothesis H is true. Hence, the quotient $P(E|H)/P(E)$ represents the support that the evidence E provides for the hypothesis H.

3.3. The Research Model

The primary goal of the current paper is to offer one out of myriad possible ways that analytical collaboration between educational stakeholders could be performed for evaluation of potential issues by simulating how much (or how little) the learning of mathematics can be improved for the students in five different schools, which used five dissimilar AI-ALS that were provided by five vendors. The probabilistic reasoning techniques used are based on BN. Within the BN, the concept of the Markov Blanket [65], in conjunction with Response Surface Methodology (RSM) [66–69], are utilized, as they are

proven techniques for examining the optimization of the relations between the variables of theoretical constructs, even if they are not physically related.

The Bayesian approach has been chosen, because it is a methodology that has been used for modeling the performances and knowledge of students; in particular, by the developers of adaptive learning software applications, such as Collins, Greer, and Huang [70]; Conati, Gertner, VanLehn, and Druzdze [71]; Jameson [72]; and, VanLehn, Niu, Siler, and Gertner [73]. However, these published works were focused on the vantage points of the developers who were describing the advantages of their respective products.

In contrast, it would be quite difficult for end-users of any AI-ALS to understand more about the inner workings of the proprietary algorithms that power the interactions with the students. The current paper proffers an approach that enables educational stakeholders to use descriptive analytics as well as predictive simulations to analyze the data that could be procured from the learners' performance reports in the server of an AI-ALS. This allows for analyses which could include comparisons and evaluations of multiple AI-ALS. The intention is to inform the educational stakeholders in each respective school, so that their teachers can remediate and bridge the gaps for the students, in whichever topics that the AI-ALS could not do so.

In Sections 4.5 and 4.6, the detailed BN model of the students' knowledge will be presented. It can inform educational stakeholders about the specific mathematics topics that the students are ready to learn, and the topics that they have already passed. Due to the coordination efforts between educational stakeholders in the five schools, they may use the vital information depicted by the relations between the nodes/variables in the BN to provide remediation for the students who are struggling in their studies. Hence, they could achieve better learning outcomes and decrease the probability of the potential risks that usage of an AI-ALS might entail (e.g., the students experiencing undue stress).

The BN model in the current paper is machine-learned from data procured from the scores of a paper-based pre-test, the learning progress scores while the students were using the AI-ALS, the Likert-scale scores from a survey, as well as the scores from a paper-based post-test. The current paper analyzes the relations using the generated BN. The theoretical constructs within the BN include the paper-based pre-test, the mediator (which is the AI-ALS), the paper-based post-test, and the noncognitive constructs (e.g., motivation, engagement, interest, self-regulation, etc.) in the survey. When researchers and educational stakeholders evaluate an AI-ALS, an understanding of these relations is essential for determining whether the interventions would be beneficial to the students. Therefore, the current paper proposes a practical Bayesian approach to demonstrate how educational stakeholders—rather than computer scientists—could analyze data from a small number of students. In order to explore the pedagogical motif of the AI-ALS, the following two types of analytics will be subsequently presented in Sections 4 and 5:

Descriptive analytics of "what has already happened?" in Section 4:

> Purpose: to use descriptive analytics to discover the pedagogical motifs of the five AI-ALS deployed in five different schools. For descriptive analytics, BN modeling in Section 4.7 will first utilize the parameter estimation algorithm to automatically detect the data distribution of each column in the dataset. Further descriptive statistical techniques that will be employed to understand more about the current baseline conditions of the students include quadrant analysis, curves analysis, and Pearson correlation analysis.

"What-if?" predictive analytics in Section 5:

> Purpose: to use predictive analytics to perform in-silico experiments with fully controllable parameters from the pre-test to the mediating intervention to the post-test for prediction of future outcomes. Instead of just simply measuring gains by subtracting the students' post-test scores from the pre-test scores, a probabilistic Bayesian approach

will be used to simulate counterfactual scenarios to better inform educators and policy makers about the pedagogical characteristics of the five AI-ALS that are being deployed in five different schools. For predictive analytics, counterfactual simulations in Section 5 will be employed to explore the pedagogical motif of the AI-ALS. In Section 6, the predictive performance of the BN model will be evaluated using tools that include the gains curve, the lift curve, the Receiver Operating Characteristic (ROC) curve, as well as by statistical bootstrapping of the data inside each column of the dataset (which is also the data distribution in each node of the BN model) by 100,000 times to generate a larger dataset to measure its precision, reliability, Gini index, lift index, calibration index, the binary log-loss, the correlation coefficient R, the coefficient of determination R2, root mean square error (RSME), and normalized root mean square error (NRSME).

4. Descriptive Analytics of "What has Already Happened?"

In this section, the procedures that were carried out in descriptive analytics to make sense of "what has already happened?" in the collected dataset will be presented. The dataset comprising 100 students (20 students from each school, from five different schools, all of whom were about 13–14 years old) who had used the AI-ALS, was imported into Bayesialab to deliberately illustrate the capabilities of BN in handling nonparametric statistical data from a small number of participants [74]. The purpose is to discover the informational "pedagogical motif" of the learning intervention generated by each AI-ALS. In the context of this study, the notion of "pedagogical motif" is conceptually defined as the pattern, timbre, disposition, and the unique characteristics with which each AI-ALS pedagogically interacts with the students.

4.1. The Dataset Procured from the Reports Generated by AI-ALS

The zip file containing the following datasets can be downloaded from https://doi.org/10.6084/m9.figshare.8206976.

The file "data_five_classes_AI_ALS.csv" contains the combined data of the five datasets from five different groups of students in different schools. For the convenience of the reader who may wish to import the data files from each group of students in each of the respective school into Bayesialab when prompted to do so in this paper, these files "data_ai_als_class_1.csv", "data_ai_als_class_2.csv", "data_ai_als_class_3.csv", "data_ai_als_class_4.csv", and "data_ai_als_class_5.csv" are also separately available in the zip file. The codebook describing the data, "ai-als-data_codebook.txt" is also included.

4.2. Codebook of the Dataset

The dataset could be procured from the reports that were generated by the server of each AI-ALS. Even though the variables from different datasets of the various AI-ALS would presumably be dissimilar, they could be aggregated to a form that is based on the mathematics topics and sub-topics (see Table A1 in Appendix A) that the students are required to learn in their curriculum. Each column in the dataset is presented as a node in the BN. It can be assumed that higher values in the data of both "math_topic_passed" (appended with the letter "P") and "math_topic_ready_for_learning" (appended with the letters "RL") are considered to be indicators of better performance, and vice-versa.

4.3. Software Used: Bayesialab

The software which will be utilized is Bayesialab [75]. A suggested pre-requisite activity for the reader is to peruse the free user-guide from http://www.bayesia.com/book/ before proceeding with the exemplars illustrated in the following sections, as it documents the tools and functionalities of the Bayesialab software.

4.4. Pre-Processing: Checking for Missing Values or Errors in the Data

It would be prudent to check the data (using the file "data_five_classes_AI_ALS.csv") for any anomalies or missing values before using Bayesialab to construct the BN. In the dataset used in this study, there were no anomalies or missing values. However, should other analysts encounter missing values in their datasets, they could use Bayesialab to predict and fill in those missing values, rather than discarding the row of data with a missing value. Bayesialab would be able to perform this by machine-learning the overall structural characteristics of that entire dataset being studied, before producing the predicted values. Bayesialab uses the Structural Expectation Maximization (EM) algorithms and Dynamic Imputation algorithms to calculate any missing values [76].

4.5. Overview of the BN Model

BN, which is also referred to as Belief Networks, Causal Probabilistic Networks, and Probabilistic Influence Diagrams are graphical models, which consist of nodes (variables) and arcs or arrows. Each node contains the data distribution of the respective variable. The arcs or arrows between the nodes represent the probabilities of the correlations between the variables [77].

Using BN, it becomes possible to use descriptive analytics to analyze the relations between the nodes (variables) and the manner in which initial probabilities, such as the number of hours spent in the AI-ALS and/or topics passed/ready to learn, and/or noncognitive factors, might influence the probabilities of future outcomes, such as the predicted learning performance of the students in the paper-based post-test.

Further, BN can also be used to perform counterfactual speculations regarding the initial states of the data distribution in the nodes (variables), given the final outcome. In the context of the current paper, exemplars will be presented in the predictive analytics segment (in Section 5) to illustrate how counterfactual simulations can be implemented while using BN. For example, we can simulate these hypothetical scenarios in the BN if we wish to find out the conditions of the initial states in the nodes (variables) that would lead to high probability of attaining high-level scores in the post-test, or if we wish to find out how to prevent students from attaining low scores or failing in the paper-based post-test.

The relation between each pair of connected nodes (variables) is determined by their respective Conditional Probability Table (CPT), which represents the probabilities of correlations between the data distributions of the parent node and the child node [78]. In the current paper, the values in the CPT are automatically machine-learned by Bayesialab, according to the data distribution of each column/variable/node in the dataset. Nevertheless, it is possible, but optional, for the user to manually enter the probability values into the CPT, if the human user wishes to override the machine learning software. In Bayesialab, the CPT of any node can be seen by double-clicking on it.

The BN model can be used to depict the data distribution of the students' score clusters (see Figure 1) in the AI-ALS in terms of the mathematics topics which include Arithmetic Readiness, Real Numbers, Linear Equations, Linear Inequalities, Functions and Lines, Exponents and Exponential Functions, Polynomials and Factoring, as well as Quadratic Functions and Equations. These score clusters were generated via machine-learning by the Bayesialab software. By generating this model from the data that contained varying levels of performance of the students (even if it was just 20 students from each school, with a total of 100 students from five schools), we could obtain a "pedagogical motif" of each AI-ALS, which meant that we could then perform simulations in each computational model to study how it could behave under certain conditions. This will be elaborated and presented later in Section 5.

4.6. Detailed Descriptions of the BN in the Current Paper

Nodes (both the blue round dots, as well as the round cornered rectangles showing the data distribution histograms) represent the variables of interest, for example, the score of a particular mathematics topic (connected to nodes with scores from their corresponding sub-topics), the number of hours that are spent by a student in the AI-ALS, the percentage of mathematics topics which a

student had passed in the AI-ALS, or the rating of a particular noncognitive factor (e.g., motivation of a student). Such nodes can correspond to symbolic/categorical variables, numerical variables with discrete values, or discretized continuous variables. We exclusively discuss BN with discrete nodes in the current paper even though BN can handle continuous variables, as it is more relevant in helping educational stakeholders categorize students into high, mid, and low achievement groups, so that teachers can utilize differentiated methods to better address the students' learning needs.

Directed links (the arrows) could represent informational (statistical) or causal dependencies among the variables. The directions are used to define kinship relations, i.e., parent-child relationships. For example, X is the parent node of Y, and Y is the child node in a Bayesian network with a link from X to Y. In the current paper, it is important to note that the Bayesian network presented is the machine-learned result of probabilistic structural equation modeling (PSEM); the arrows represent the probabilistic structural relationships between the parent node and the child nodes. The first letter of the name of each node/data entity is presented in the upper case for better readability.

In the BN model used in the current paper (see Figure 1), the node representing the Pre-test results (from a paper-based math test) is connected to the "mediator" node representing the pedagogical motif of the AI-ALS, and subsequently the "mediator" node that represents the pedagogical motif of the AI-ALS is also connected to the node that represents the Post-test results (from another paper-based math test). This enables the probabilities of the AI-ALS as a mediator of the students' performance to be calculated, and subsequently it will be possible to simulate hypothetical scenarios (to be presented later in Section 5).

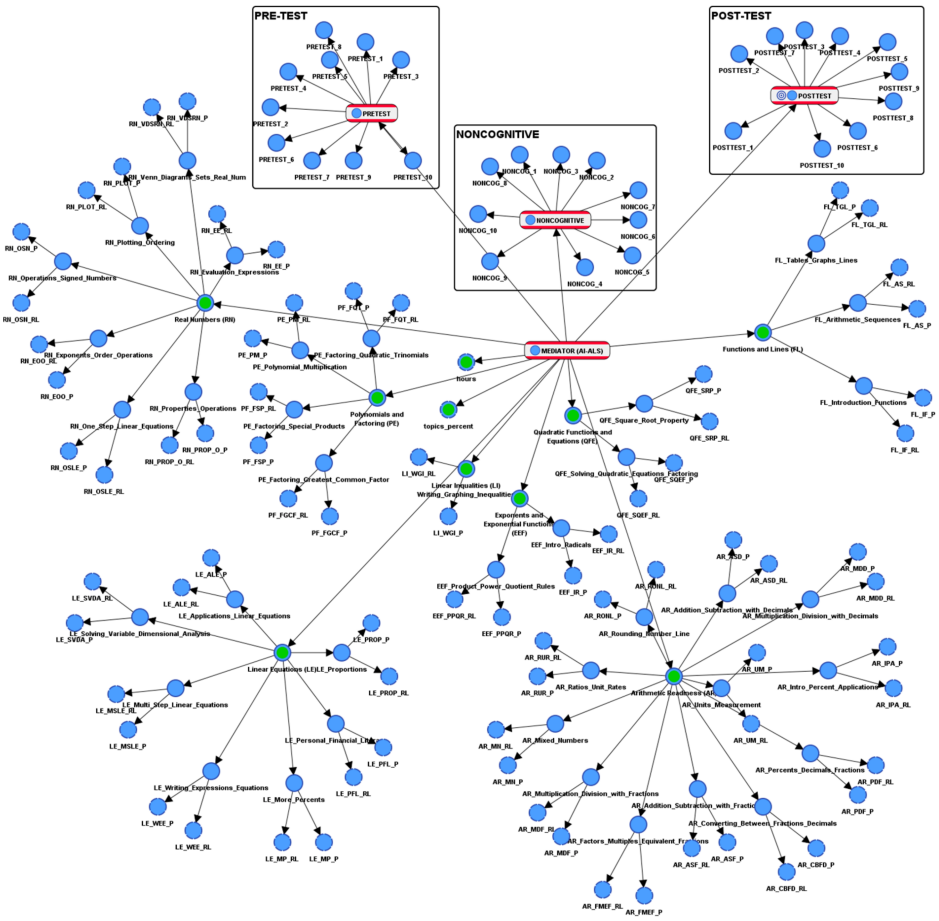

Figure 1. Full view of the Bayesian network: the component nodes (in blue) and the superordinate factor nodes (in green) were used for machine learning the overall performance of 100 students who had used the five different artificial intelligence-enabled adaptive learning systems (AI-ALS).

4.7. Descriptive Statistical Analysis of the Dataset

From the combined dataset of all the 100 students' performance who had used the five different AI-ALS (using the file "data_five_classes_AI_ALS.csv"), the following score-clusters machine-learned by Bayesialab were observed (see Figure 2):

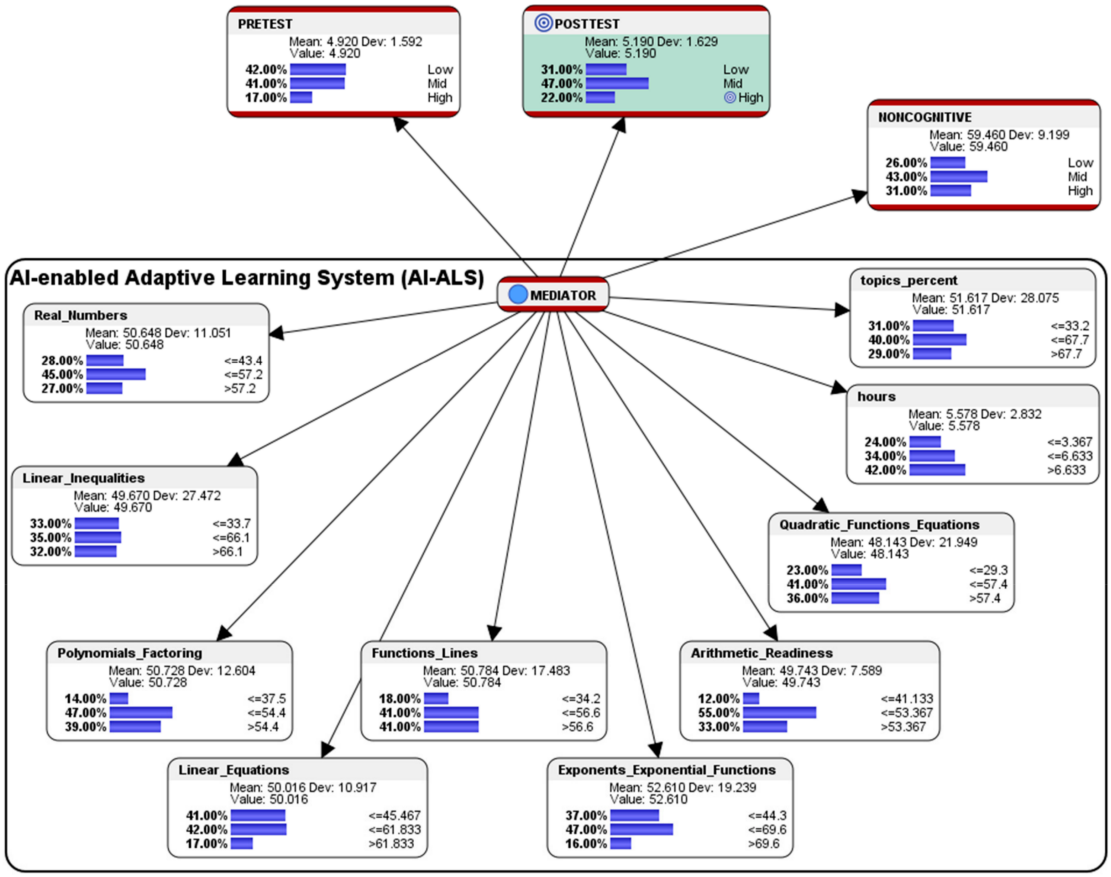

Figure 2. Simplified aggregated view of the Bayesian network previously shown in Figure 1, presenting only the superordinate factor nodes with their machine-learned score-clusters, depicting the overall performance levels of all 100 students who had used the five dissimilar AI-ALS from five different vendors.

In the paper-based Pre-test before the students used the AI-ALS, 42% of the students scored at the Low-level, 41% scored at the Mid-level, and 17% scored at the High-level. In the paper-based Post-test after the students had gone through the training within the AI-ALS, 31% scored at the Low-level, 47% scored at the Mid-level, and 22% scored at the High-level. Overall, in terms of conventional gains, there was an improvement of 11% of the students who had scored at the Low-level (a decrease from 42% in the Pre-test to 31% in the Post-test); there was an improvement of 6% in the students who had scored at the Mid-level (an increase from 41% in the Pre-test to 47% in the Post-test); and, there was an improvement of 5% in the students who had scored at the High-level (an increase from 17% in the Pre-test to 22% in the Post-test).

In the aggregated Noncognitive factor, 26% of the students were at the so-called Low-level, 43% were at the Mid-level, and 31% were at the High-level.

Within the AI-ALS, in the topic of Real Numbers, 28% of the students scored at the Low-level (<=43.4% of the total marks for Real Numbers), 45% scored at the Mid-level (>43.4 and <=57.2), and 27% scored at the High-level (>57.2).

In the topic of Linear Inequalities, 33% scored at the Low-level (<=33.7), 35% scored at the Mid-level (>33.7 and <=66.1), and 32% scored at the High-level (>66.1).

In the topic of Polynomials and Factoring, 14% of the students scored at the Low-level (<=37.5), 47% scored at the Mid-level (>37.5 and <=54.4), and 39% scored at the High-level (>54.4).

In the topic of Linear Equations, 41% of the students scored at the Low-level (<=45.467), 42% scored at the Mid-level (>45.467 and <=61.833), and 17% scored at the High-level (>61.833).

In the topic of Functions and Lines, 18% of the students scored at the Low-level (<=34.2), 41% scored at the Mid-level (>34.2 and <=56.5), and 41% scored at the High-level (>56.6).

In the topic of Exponents and Exponential Functions, 37% of the students scored at the Low-level (<=44.3), 47% scored at the Mid-level (>44.3 and <=69.6), and 16% scored at the High-level (>69.6).

In the topic of Arithmetic Readiness, 12% of the students scored at the Low-level (<=41.133), 55% scored at the Mid-level (>41.133 and <=53.367), and 33% scored at the High-level (>53.367).

In the topic of Quadratic Functions and Equations, 23% of the students scored at the Low-level (<=29.3), 41% scored at the Mid-level (>29.3 and <=57.4), and 36% scored at the High-level (>53.4).

Regarding the average number of hours spent by each student in the AI-ALS, 24% of the students were at the Low-level (<=3.367 h), 34% of the students were at the Mid-level (>3.367 and <=6.633 h), and 42% were at the High-level (>6.633 h).

In the percentage of the total number of topics that were mastered by the students in the AI-ALS, 31% of the students were at the Low-level (<=33.3%), 40% were at the Mid-level (>33.3% and <=67.7%), and 29% were at the High-level (>67.7%).

4.7.1. Descriptive Analytics: Profile Analysis of Each AI-ALS

A strategic overview of how the students performed (see Figures 3 and 4) could be accomplished via profile analysis. This tool can be activated in Bayesialab via these steps: *Bayesialab (validation mode) > Visual > Segment > Profile*.

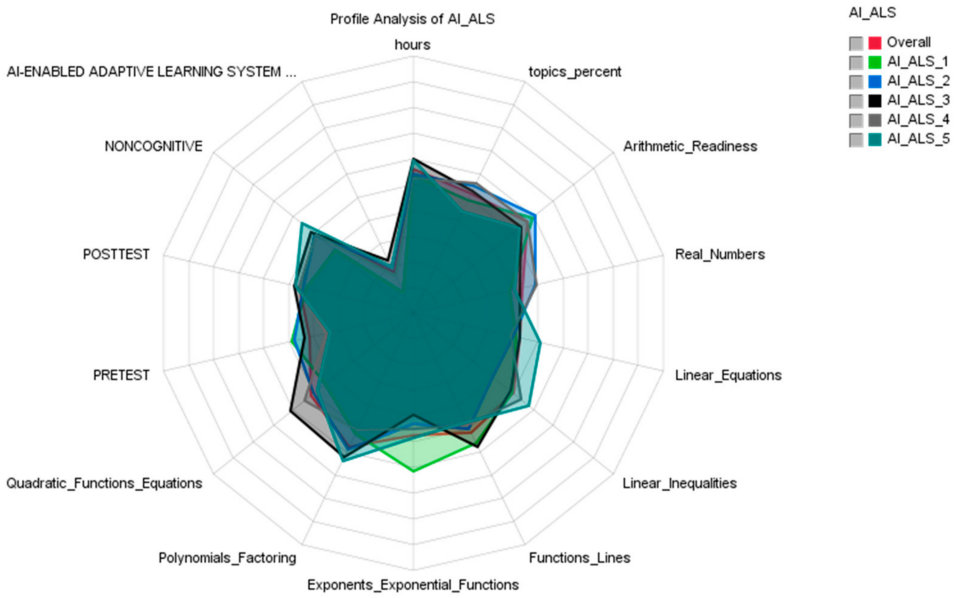

Figure 3. Profile analysis of the five groups of students, each of which had used a different AI-ALS.

Figure 4 is an alternative presentation of the profiles presenting the performance of the five groups of students in different schools, each of which had used a different AI-ALS.

Figure 4. Profiles of five different AI-ALS, each from a different vendor, superimposed on top of the overall profile.

4.7.2. Descriptive Analytics: Quadrant Analysis

Comparison of Total Effects of the five different AI-ALS on the paper-based Post-test can be performed while using quadrant analysis. This tool can be activated in Bayesialab via these steps: Bayesialab (validation mode) > Analysis > Report > Target > Total Effects on Target > Quadrants.

It would be contrived to measure the correlation between the scores achieved by the students in their respective AI-ALS against their scores in the hardcopy paper-based post-test, because some students could have scored poorly in the AI-ALS as their poor understanding of certain math concepts might have been "surfaced" by the systems, but subsequently, they might have scored well in the paper-based post-test. Conversely, some students might have scored high in the AI-ALS because the questions were easy, but they might have scored low in the paper-based post-test. Hence, it absolutely does not mean that an AI-ALS would be ranked higher in the quadrant analysis chart if the students' scores within the AI-ALS are higher.

Each chart of the quadrant analysis generated by Bayesialab (see Figures 5 and 6) is divided into four quadrants. The variables' means (of each mathematics topic) are represented along the x-axis. The mean of the standardized total effect on the target (the paper-based post-test) is represented along the y-axis. Quadrant analysis example 1 (see Figure 5) utilized the file "data_five_classes_AI_ALS.csv". As a suggestion, the quadrants could be interpreted, as follows:

Top Right Quadrant (high volume, high impact on target node): This group contains the important variables with greater total effect on the target than the mean value. These AI-ALS are effective in contributing to the success of the students in the paper-based post-test. The AI-ALS supplied by Vendor 1, Vendor 2, Vendor 4, and Vendor 5 are in this category.

Top Left Quadrant (low volume, high impact on target node): Any AI-ALS in this category might be beneficial to the high-performing students, but not so beneficial to the mid- or low-performing students. There is no AI-ALS from any vendor in this quadrant.

Bottom Right Quadrant (high volume, low impact on target node): The AI-ALS from Vendor 3 is in this category, so educational stakeholders should consider conducting further investigation to find out why this AI-ALS could not contribute to beneficial results in the paper-based post-test for the students.

Bottom Left Quadrant (low volume, low impact on target node): Any AI-ALS in this category has relatively lower impact on the target node (the paper-based post-test). There is no AI-ALS from any vendor in this quadrant.

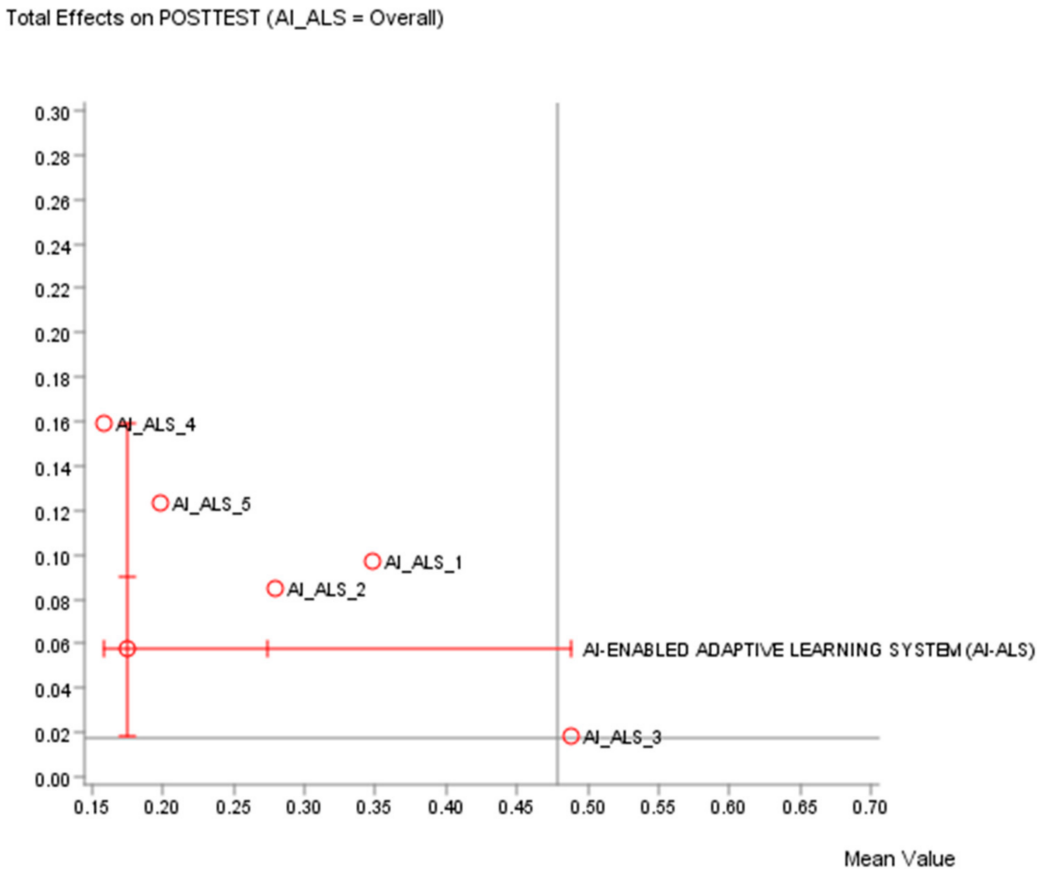

Figure 5. Comparison of Total Effects of the five different AI-ALS on the Post-test, which was machine-learned and generated by Bayesialab.

Quadrant analysis example 2 (see Figure 6) utilized the file "data_five_classes_AI_ALS.csv". As a suggestion, the quadrants could be interpreted, as follows:

Top Right Quadrant (high volume, high impact on target node): This quadrant contains the AI-ALS with greater total effect on the target than the mean value. Only the AI-ALS from Vendor 2 is in this quadrant. These noncognitive factors associated with this AI-ALS are important to the success of the students in the paper-based post-test, and the educational stakeholders should further explore how the noncognitive factors (e.g., motivation, stress management, psychological well-being, etc.) that are associated with the AI-ALS from Vendor 2 could be beneficial in helping the students to understand and learn the concepts well in these mathematics topics.

Top Left Quadrant (low volume, high impact on target node): Any AI-ALS in this category is associated with the noncognitive factors that might be beneficial for the high-performing students, but might not be so beneficial to the mid- or low-performing students. The AI-ALS supplied by Vendor 4 and Vendor 5 are in this quadrant.

Bottom Right Quadrant (high volume, low impact on target node): There is no AI-ALS from any vendor in this quadrant. If there is any AI-ALS in this category, educational stakeholders should consider conducting further investigation to find out why the noncognitive factors associated with this AI-ALS could not contribute to beneficial results in the paper-based post-test for the students.

Bottom Left Quadrant (low volume, low impact on target node): Any AI-ALS in this category has noncognitive factors that have relatively lower impact on the target node (the paper-based post-test). The AI-ALS from Vendor 1 and Vendor 3 are in this quadrant.

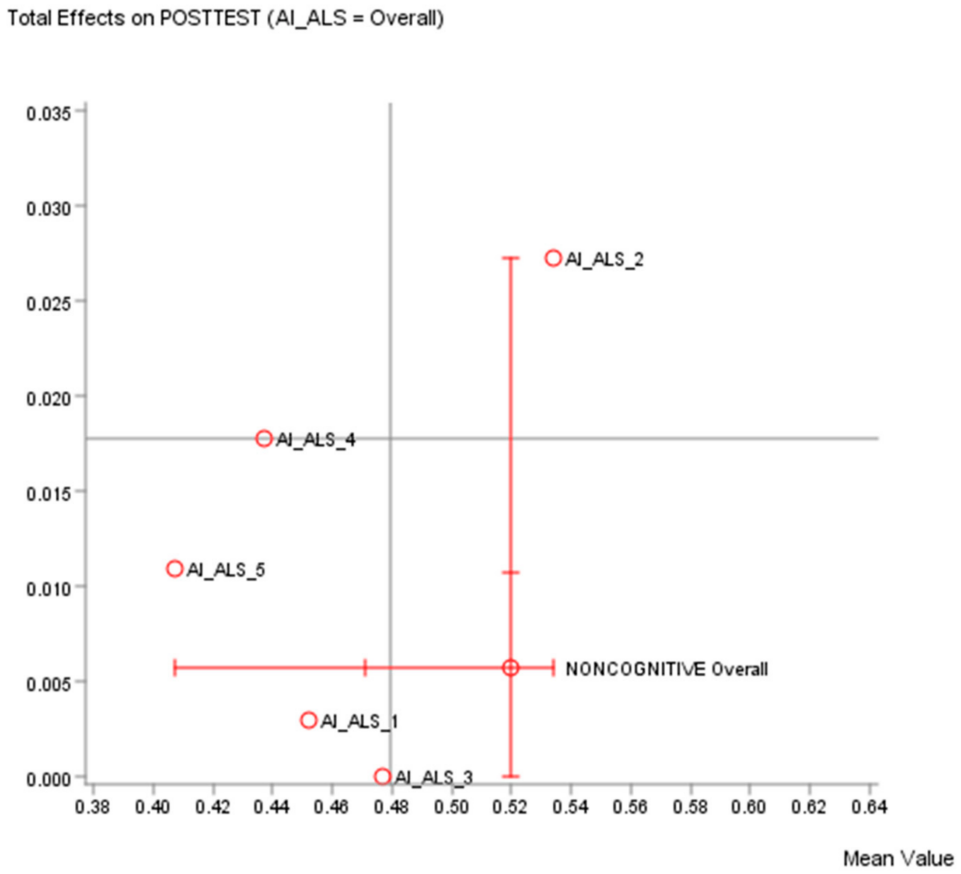

Figure 6. Comparison of Total Effects of the data in the Noncognitive node on the Post-test node, which was machine-learned from the data of the five different groups of students who had used five dissimilar AI-ALS.

4.7.3. Descriptive Analytics: Comparative Analysis of the Five AI-ALS

In this section, the performance results of the five classes of students who had used five dissimilar AI-ALS in five different schools will be presented.

Comparison between the AI-ALS from Vendor 1 and the Combined Average of the Five AI-ALS:

Using the file "data_ai_als_class_1.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 1 (see Figure 7):

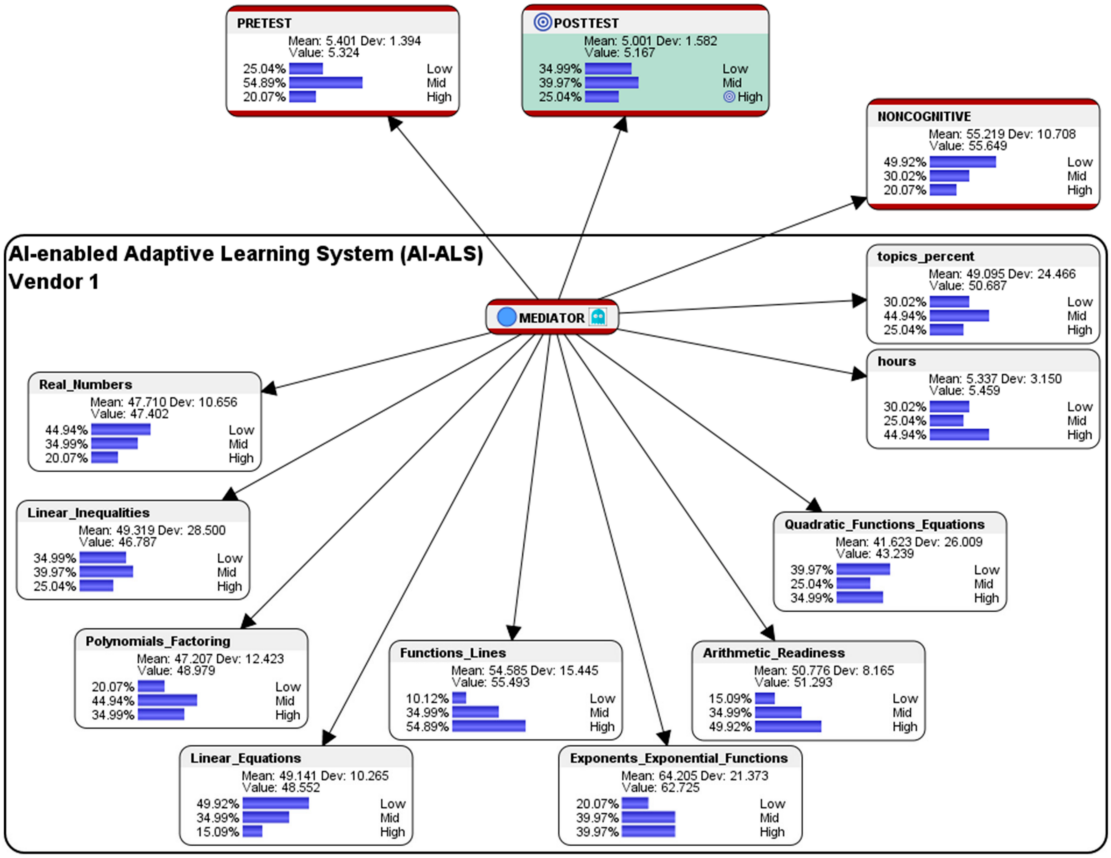

Figure 7. BN model of the students who had used the AI-ALS from Vendor 1 (N = 20 students).

In the paper-based Pre-test before the students used the AI-ALS from Vendor 1, 25.04% had scored at the Low-level (as compared to the combined average of 42% of the students who had scored at the Low-level), 54.89% had scored at the Mid-level (when compared to the combined average of 41% who had scored at the Mid-level), and 20.07% had scored at the High-level (as compared to the combined average of 17% scored at the High-level).

In the paper-based Post-test after the students had gone through the training within the AI-ALS from Vendor 1, 34.99% had scored at the Low-level (as compared to the combined average of 31% who had scored at the Low-level), 39.97% had scored at the Mid-level (when compared to the combined average of 47% who had scored at the Mid-level), and 25.04% had scored at the High-level (as compared to the combined average of 22% who had scored at the High-level). Overall, in terms of conventional gains by comparing the Pre-test vis-à-vis the Post-test, there was an unfavorable higher difference of 9.95% of the students who scored at the Low-level (from 25.04% in the Pre-test to 34.99% in the Post-test); there was a decline of 14.92% in the students who scored at the Mid-level (an decrease from 54.89% in the Pre-test to 39.97% in the Post-test); however, there was a favorable higher difference of 4.97% in the students who scored at the High-level (from 20.07% in the Pre-test to 25.04% in the Post-test).

In the aggregated Noncognitive factor, 49.92% of the students who had used the AI-ALS from Vendor 1 were at the so-called Low-level (a higher difference of 23.92% as compared to the combined average of 26% of the students who were at the Low-level), 30.02% were at the Mid-level (a lower difference of 12.98% as compared to the combined average of 43% of students who were at the Mid-level), and 20.07% were at the High-level (a lower difference of 10.93% when compared to the combined average of 31% of student who were at the High-level).

Within the AI-ALS from Vendor 1, in the topic of Real Numbers, 44.94% of the students scored at the Low-level (a higher difference of 16.94% as compared to the combined average of 28% of the

students who scored at the Low-level), 34.99% of the students scored at the Mid-level (a lower difference of 10.01% as compared to the combined average of 45% of the students who scored at the Mid-level, and 20.07% of the students scored at the High-level (a lower difference of 6.93% compared to the combined average of 27% of the students who scored at the High-level.

In the topic of Linear Inequalities, 34.99% of the students scored at the Low-level (a higher difference of 1.99% compared to the combined average of 33% of the students who scored at the Low-level), 39.97% of the students scored at the Mid-level (a higher difference of 4.97% when compared to the combined average of 35% of the students who scored at the Mid-level), and 25.04% of the students scored at the High-level (a lower difference of 6.96% as compared to the combined average of 32% of the students who scored at the High-level.

In the topic of Polynomials and Factoring, 49.92% of the students scored at the Low-level (a higher difference of 35.92% when compared to the combined average of 14% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 12.01% as compared to the combined average of 47% of the students who scored at the Mid-level), and 15.09% scored at the High-level (a lower difference of 23.91% when compared to the combined average of 39% of the students who scored at the High-level).

In the topic of Linear Equations, 49.92% scored at the Low-level (a higher difference of 8.92% when compared to the combined average of 41% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 7.01% when compared to the combined average of 42% of the students who scored at the Mid-level), and 15.09% scored at the High-level (a lower difference of 1.91% when compared to the combined average of 17% scored at the High-level).

In the topic of Functions and Lines, 10.12% scored at the Low-level (a lower difference of 7.88% compared to the combined average of 18% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 6.01% as compared to the combined average of 41% of the students who scored at the Mid-level), and 54.89% who scored at the High-level (a higher difference of 13.89% as compared to the combined average of 41% of the students who scored at the High-level).

In the topic of Exponents and Exponential Functions, 20.07% scored at the Low-level (a higher difference of 16.93% when compared to the combined average of 37% of the students who scored at the Low-level), 39.97% scored at the Mid-level (a lower difference of 7.03% as compared to the combined average of 47% of the students who scored at the Mid-level), and 39.97% scored at the High-level (a higher difference of 23.97% when compared to the combined average of 16% of the students who scored at the High-level).

In the topic of Arithmetic Readiness, 15.09% scored at the Low-level (a higher difference of 3.09% compared to the combined average of 12% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 20.01% compared to the combined average of 55% of the students who scored at the Mid-level), and 49.92% scored at the High-level (a higher difference of 16.92% s compared to the combined average of 33% scored at the High-level).

Regarding the topic of Quadratic Functions and Equations, 39.97% of the students scored at the Low-level (a higher difference of 16.97% as compared to the combined average of 23% of the students who scored at the Low-level), 25.04% scored at the Mid-level (a lower difference of 15.96% compared to the combined average of 41% scored at the Mid-level), and 34.99% scored at the High-level (a lower difference of 1.01% when compared to the combined average of 36% of the students who scored at the High-level).

Within the AI-ALS by Vendor 1, in the average number of hours spent by each student, 30.02% of the students were at the Low-level (a higher difference of 6.02% compared to the combined average of 24% of the students were at the Low-level), 25.04% were at the Mid-level (a lower difference of 8.96% as compared to the combined average of 34% of the students who were at the Mid-level), and 44.94% were at the High-level (a higher difference of 2.94% when compared to the combined average of 42% who were at the High-level).

In the percentage of the total number of topics that were mastered by the students in the AI-ALS by Vendor 1, 30.02% of the students were at the Low-level (a slightly lower difference of 0.98% compared to the combined average of 31% of the students who were at the Low-level), 44.94% were at the Mid-level (a higher difference of 4.94% compared to the combined average of 40% who were at the Mid-level), and 25.04% were at the High-level (a lower difference of 3.96% when compared to the combined average of 29% who were at the High-level).

Visualization of the Performance of the Students Who had Used Vendor 2's AI-ALS:

Using the file "data_ai_als_class_2.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 2 (see Figure 8):

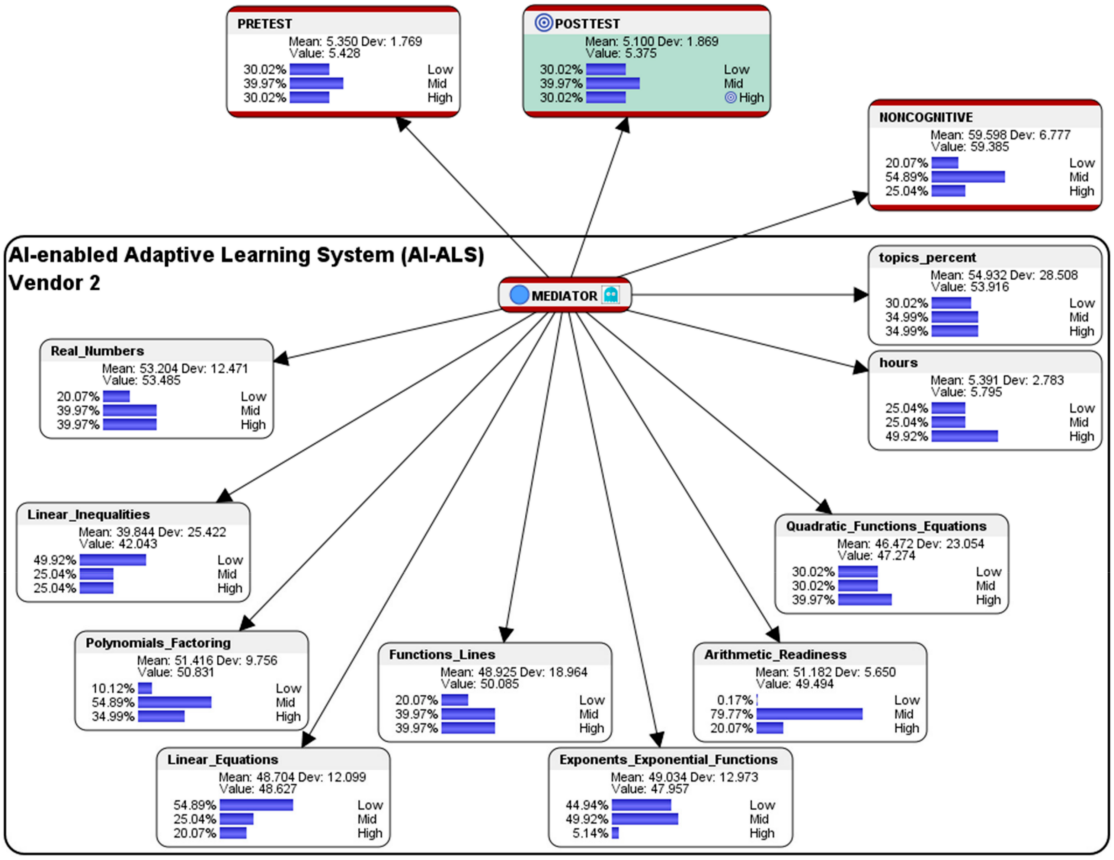

Figure 8. BN model of the students who had used the AI-ALS from Vendor 2 (N = 20 students).

Visualization of the Performance of the Students Who Had Used Vendor 3's AI-ALS:

Using the file "data_ai_als_class_3.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 3, (see Figure 9):

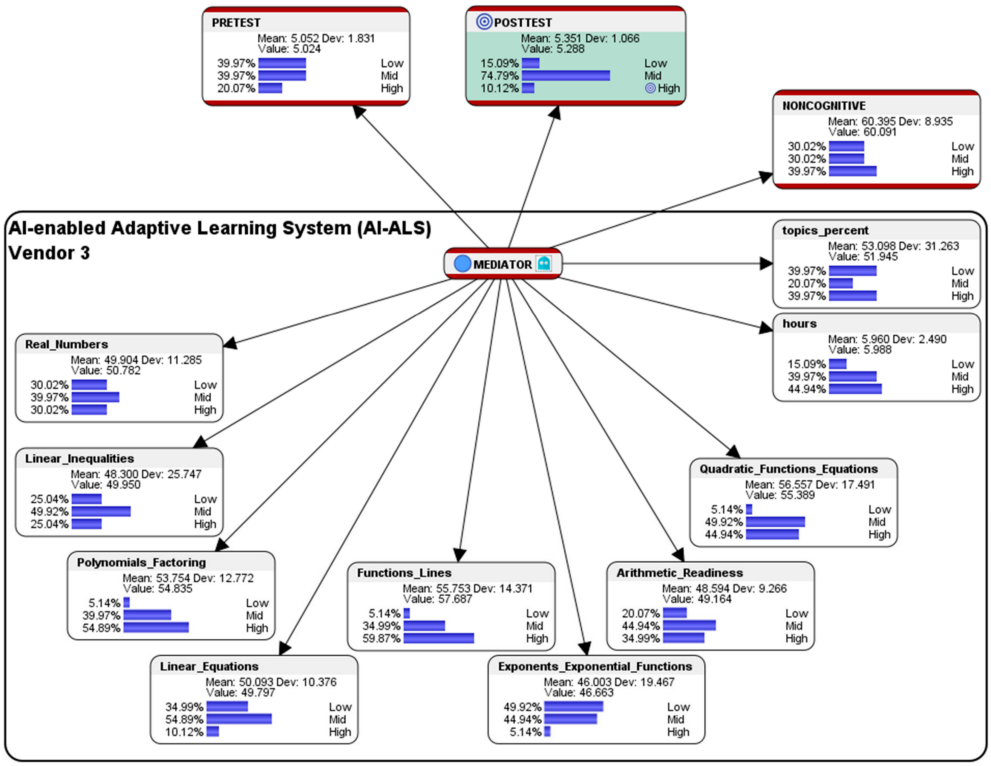

Figure 9. BN model of the students who had used the AI-ALS from Vendor 3 (N = 20 students).

Visualization of the Performance of the Students Who Had Used Vendor 4's AI-ALS:

Using the file "data_ai_als_class_4.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 4, (see Figure 10):

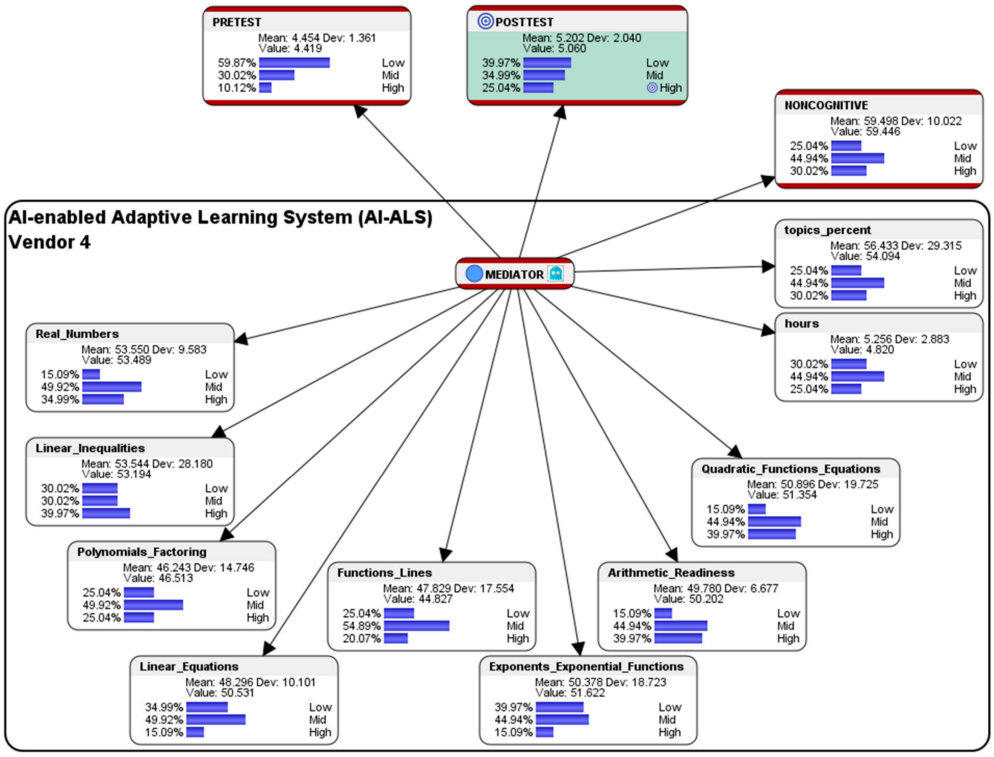

Figure 10. BN model of the students who had used the AI-ALS from Vendor 4 (N = 20 students).

Visualization of the Performance of the Students Who had Used Vendor 5's AI-ALS:

Using the file "data_ai_als_class_5.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 5 (see Figure 11):

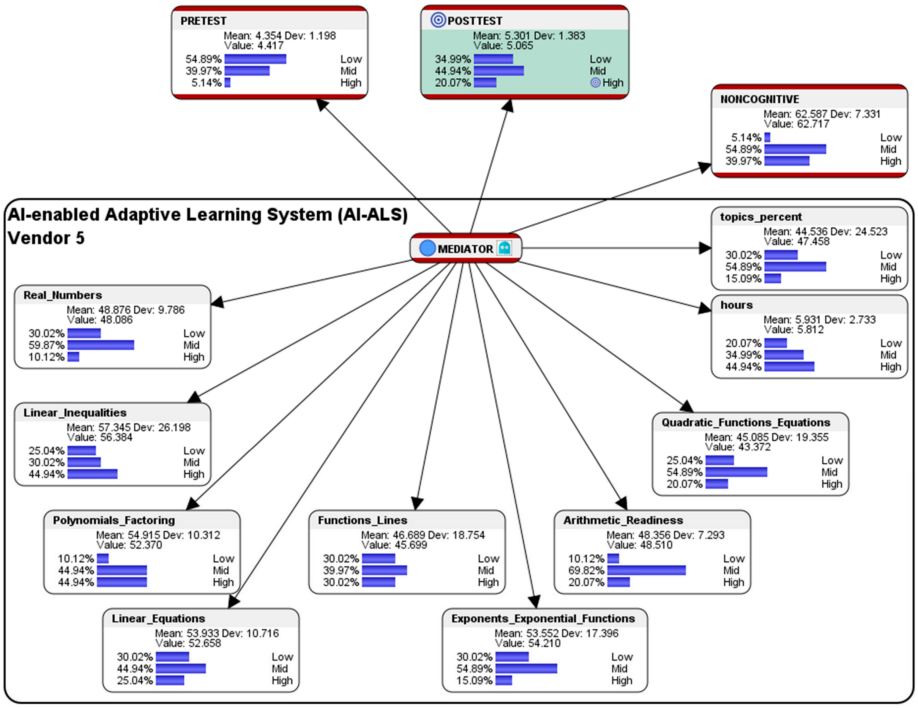

Figure 11. BN model of the students who had used the AI-ALS from Vendor 5 (N = 20 students).

4.7.4. Sensitivity Analysis of the Mathematics Topics that Contribute to the Performance of the Students who had Used the Five Dissimilar AI-ALS from the Five Vendors

Posterior Probability of the Post-test can be performed on the data from each school, while using tornado diagrams (see Figure 12). Sensitivity analysis can be activated in Bayesialab via these steps: *Bayesialab (validation mode) > Analysis > Visual > Sensitivity > Tornado diagrams on Total Effects*.

Each blue tornado chart of the total effects presents the performance (in the learning progress) of the students in each mathematics topic within the AI-ALS, in terms of the posterior probability of achieving high-level scores in the paper-based post-test. This implies that, in the AI-ALS proved by each vendor, the problem-solving practice that the students had in certain mathematics topics might have contributed to the high scores that were achieved by the students in the paper-based post-test. The longer blue bars represent higher sensitivity, in terms of how changes in the score of each mathematics topic (that is, their learning progress within each AI-ALS) could potentially affect the outcome in the paper-based post-test. Further coordination between the education stakeholders and the vendor of each respective AI-ALS should be carried out to understand how the teachers can focus on providing the students remediation of the more sensitive mathematics topics (represented with longer blue bars), as they seem to be important in affecting the performance of their students who could score high marks in the paper-based post-test.

Each red tornado chart of the total effects presents the performance of the students in each mathematics topic within the AI-ALS, in terms of the posterior probability of achieving low-level scores in the paper-based post-test. This implies that, in the AI-ALS proved by the vendor, the problem-solving practice that the students had in the mathematics topics might have contributed to the high scores that were achieved by the students in the paper-based post-test. The longer red bars represent higher sensitivity, in terms of how changes in the score of each mathematics topic (that is, their learning progress within

each AI-ALS) could potentially affect the outcome in the paper-based post-test. Further coordination via discussions between the education stakeholders and each respective vendor of the AI-ALS should be carried out to understand how the teachers can focus on providing the students remediation of the more sensitive mathematics topics (represented with longer red bars), as they seem to be affecting the performance of their students who could only score low marks in the paper-based post-test.

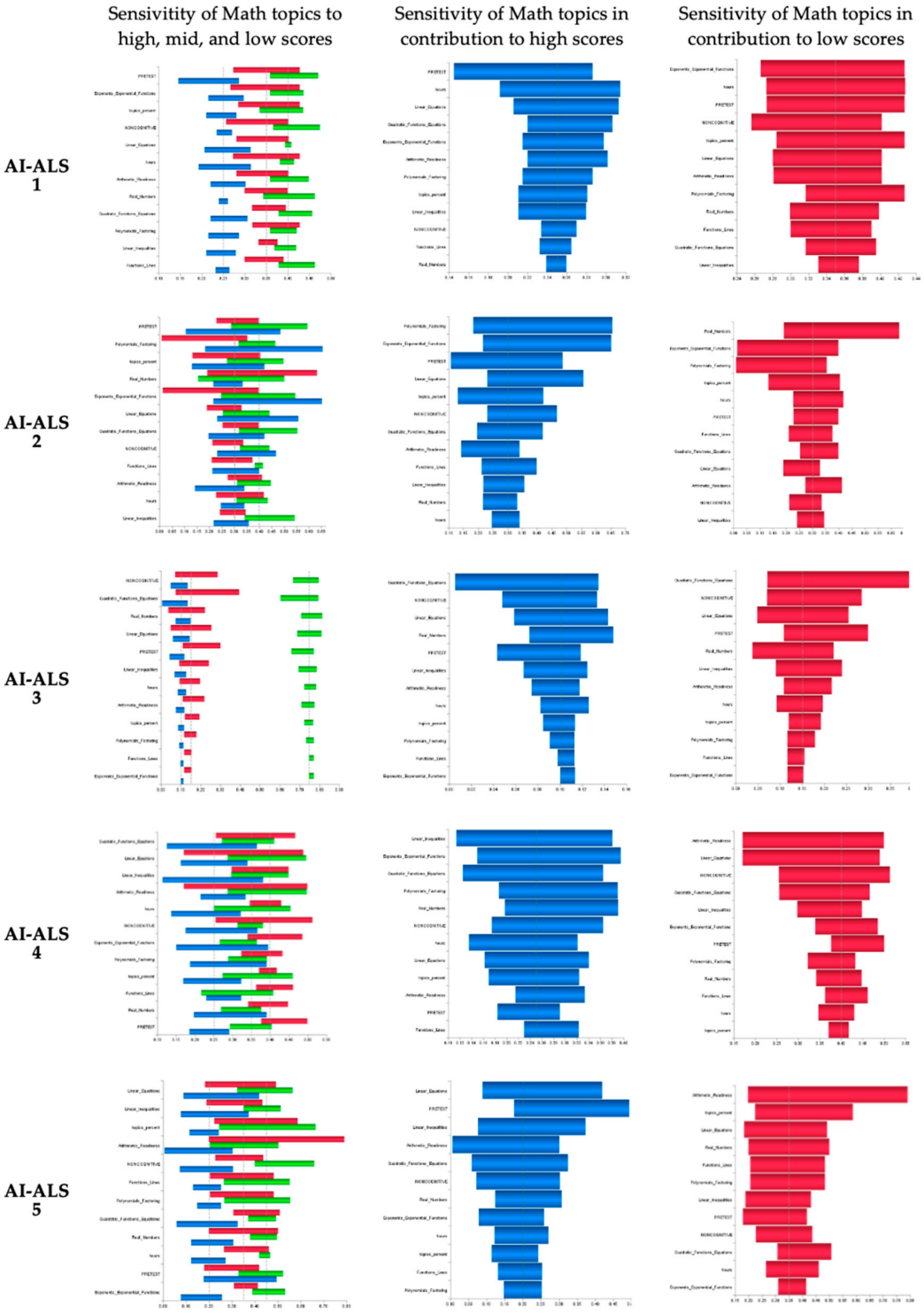

Figure 12. Visualizations of the sensitivity analysis data of the five groups of students in their respective AI-ALS, regarding how their learning progress of the mathematics topics within each AI-ALS could potentially affect their outcomes in the paper-based post-test.

4.7.5. Descriptive Analytics: Oversight Using Curves Analysis of the AI-ALS from the Five Vendors

Another way to visualize the influence of the students' mastery of the various mathematics topics on their paper-based post-test can be accomplished by using this tool in Baysialab via these steps on the menubar: Bayesialab (validation mode) > Analysis > Visual > Target > Target's Posterior > Curves > Total Effects.

As observed in Figure 13, the plots of the data reveal that the relationships between the total effects and the various factors on the target node (that is, the paper-based post-test) could be linear or curvilinear. The curvilinear lines suggest that there might be "peaks" or "valleys" in some of the relationships between the input variables (e.g., the number of hours spent using the AI-ALS, or the quality of the noncognitive factors, or the scores achieved by the students within each AI-ALS, or the percentage of mathematics topics mastered within the AI-ALS) and their respective educational outcomes in the paper-based post-test. With these curves analysis charts, further discussions could be initiated amongst the policy makers, technology vendors, teachers, parents, and students to help improve the learning experiences of the students.

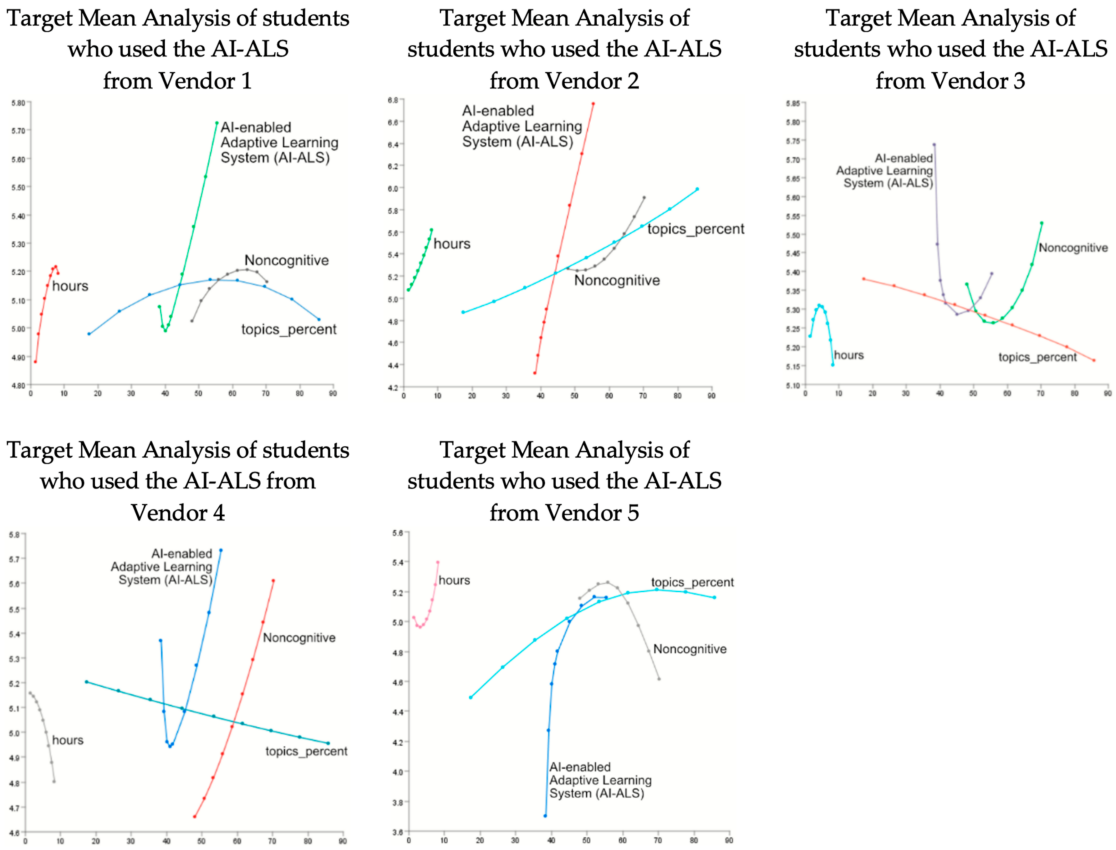

Figure 13. Target Mean Analysis of five different groups of students, each of which had used an AI-ALS from a different vendor.

4.7.6. Descriptive Analytics: Pearson Correlation Analysis

Descriptive analytics can also be performed using the Pearson correlation analysis tool in Bayesialab. It can be used for the corroboration of the relationship analyses between the students' learning performances in the AI-ALS and their corresponding performances in the paper-based post-test. The visualizations of the Pearson correlations can be presented, so that it is easier to see the positive correlations highlighted in blue, and the negative correlations faded out in red (see Figure 14). This tool can be activated in Bayesialab via these steps on the menubar: Analysis > Visual > Overall > Arc > Pearson Correlation.

One suggestion for interpretation of the negative Pearson correlations could be that the red lines and nodes might represent the regions where the weaknesses of the students were "surfaced" or educed (drawn out) by the AI-ALS. It might not necessarily be an undesirable situation, provided that the teacher could provide remediation to the students so that the gaps that the AI-ALS could not bridge for the students (e.g., if the AI-ALS could not read the students' workings to pin-point where the mathematical calculation mistakes were for the students) were addressed.

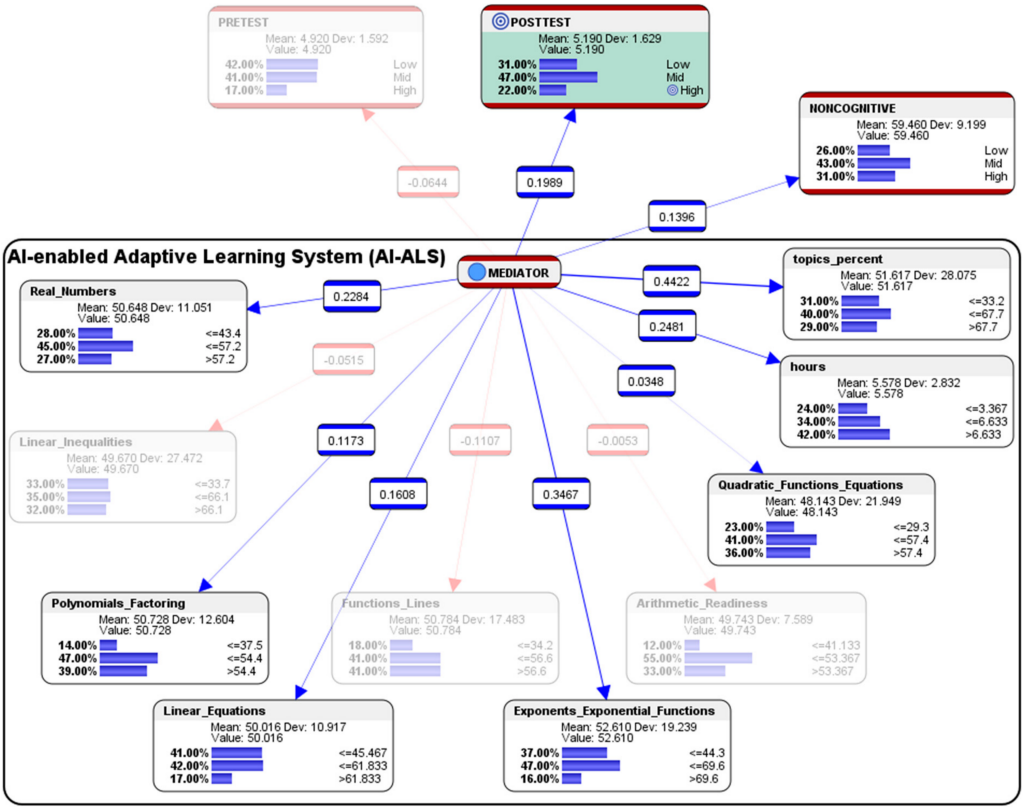

Figure 14. Pearson correlations between the students' learning progress of the mathematics topics within the AI-ALS and their corresponding performances in the paper-based post-test.

4.7.7. Descriptive Analytics: Oversight of the Gains in the Different Groups of Students

No gain in performance (scores in the post-test vis-à-vis the pre-test) was observed for the students who had used AI-ALS from Vendor 2, and negative gain (the scores in the post-test were lower than those in the pre-test) was observed for the students who had used the AI-ALS from Vendor 3, as observed in Table 1 and Figure 15. However, it might not be the fault of the AI-ALS that those students underperformed. Further qualitative interviews with the students might reveal the possible reasons for these preliminary observations.

Table 1. Comparisons between scores within the five AI-ALS and the paper-based post-tests.

AI-ALS Vendor	AI-ALS Low-Level Score (% of Students)	AI-ALS High-Level Score (% of Students)	Post-Pre Test High-Level Score Gain (% of Students)
1	35.00	30.10	4.97
2	50.10	29.89	0.00
3	25.04	44.24	−9.95
4	35.06	30.05	14.92
5	29.63	45.32	14.93

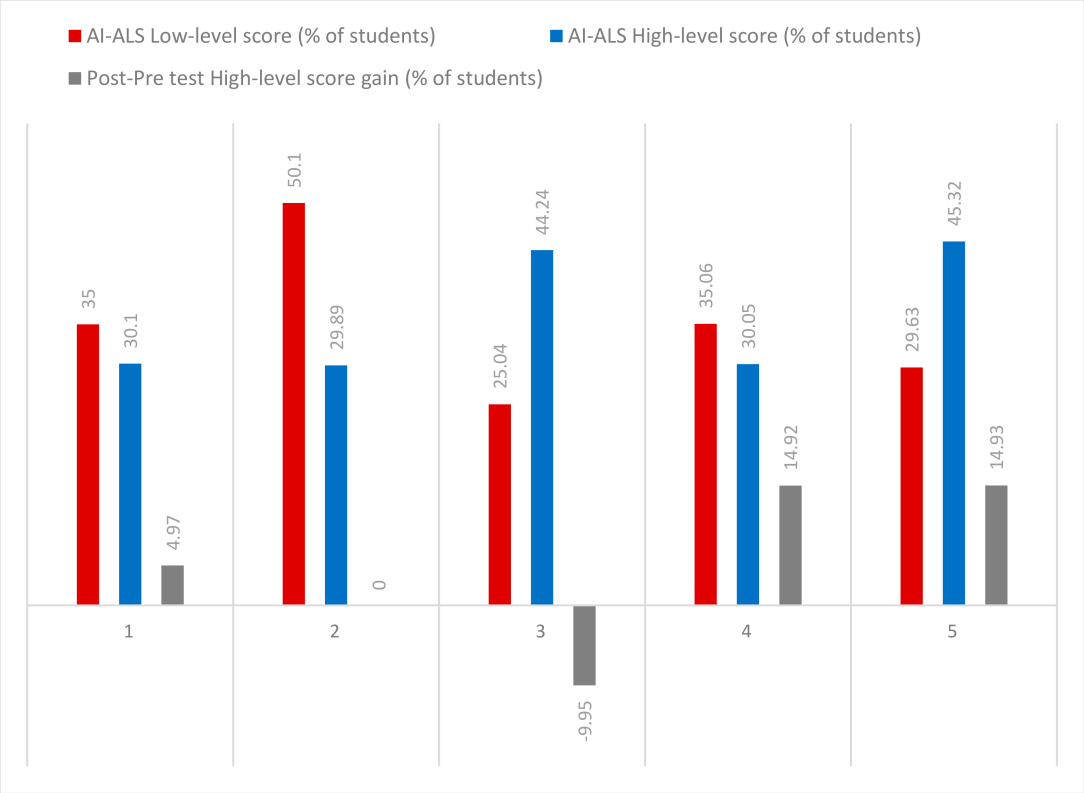

Figure 15. Histograms depicting the performance of each class of students: the low-level scores within each AI-ALS are presented in red; the high-level scores within each AI-ALS are presented in blue; their corresponding high-level score gains in the paper-based post-test are represented in gray.

There seemed to be no clear pattern of correlation between the difficulty of scoring high-level scores or low-level scores within each AI-ALS and the gains in the high-level scores in the paper-based post-test, contrary to what was initially hypothesized by the researcher in Section 2.2. In other words, making it easy (or even difficult) for the students to score at the high-level might not necessarily result in corresponding high-level gains in the paper-based post-test, probably because of the uniqueness of each AI-ALS and each class of students.

However, although direct comparisons between the five AI-ALS might seem challenging, it would still be possible to predict how the performance of each group of students within their respective AI-ALS could be optimized to achieve high scores in the paper-based post-test. To demonstrate that, "what-if?" predictive analytics would be utilized in the subsequent section.

5. "What-If?" Predictive Analytics

In this section, the following predictive analytics reports will be presented unabridged, in order to delineate how human-centric reasoning could be applied to interpret the counterfactual results that were generated by the AI-based BN model. For better readability, the first letter of the names of the BN nodes and entities would be presented in the upper case.

5.1. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 1

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 1, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions needed in the AI-ALS from Vendor 1 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

To predict the conditions that would enable 100% of the students in Class 1, who had used Vendor 1's AI-ALS to score at the High-level in the paper-based Post-test, hard evidence was set on it (by double-clicking on the High-level histogram bar in Bayesialab). The following counterfactually simulated results of score-clusters were observed (see Figure 16):

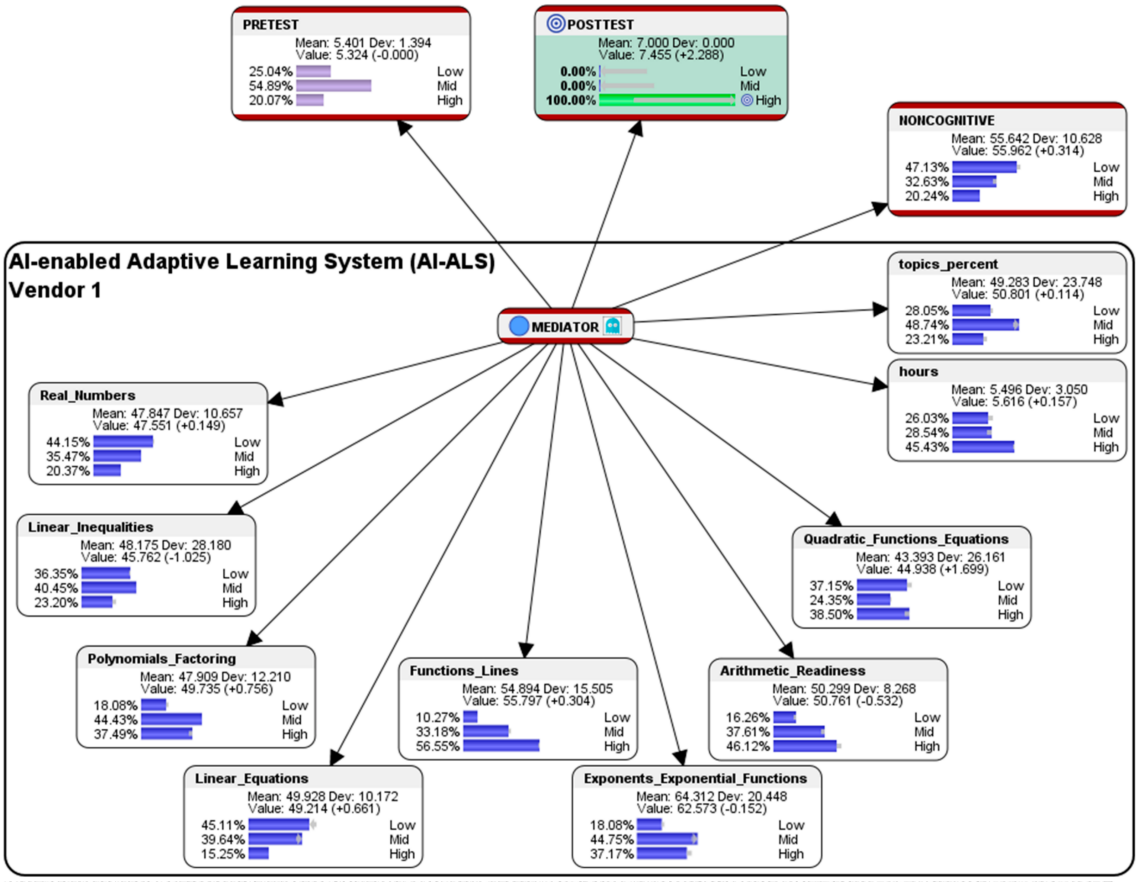

Figure 16. Simulation of counterfactual results for 100% of the students who had used Vendor 1's AI-ALS to score at the high-level in the post-test.

Within the AI-ALS from Vendor 1, in the aggregated Noncognitive factor, ideally 47.13% of the students who had used the AI-ALS from Vendor 1 should be at the so-called Low-level (a lower difference of 2.79% when compared to the original 49.92% of the students who were at the Low-level); 32.63% should be at the Mid-level (a higher difference of 2.61% compared to the original 30.02% of students who were at the Mid-level); and 20.64% should be at the High-level (an almost negligible higher difference of 0.57% as compared to the original 20.07% of students who were at the High-level).

Within the AI-ALS from Vendor 1, in the topic of Real Numbers, ideally 44.15% of the students should score at the Low-level (a slightly lower difference of 0.79% compared to the original 44.94% of the students who scored at the Low-level), 35.47% of the students should score at the Mid-level (a slightly higher difference of 0.48% as compared to the original 34.99% of the students who scored at the Mid-level), and 20.37% of the students should score at the High-level (a slightly higher difference of 0.3% when compared to the original 20.07% of the students who scored at the High-level. The simulated results for the topic of Real Numbers suggest that Vendor 1's AI-ALS was already performing close to optimum in terms of contributing the students scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Linear Inequalities, ideally 36.35% of the students should score at the Low-level (a higher difference of 1.36% as compared to the original 34.99% of the students who scored at the Low-level); 40.45% of the students should score at the Mid-level (an almost negligible higher difference of 0.48% when compared to the original 39.97% of the students who scored at the Mid-level); and, 23.20% of the students should score at the High-level (a slightly lower difference of 1.84% as compared to the original 25.04% of the students who scored at the High-level. The simulated results suggest that, if Vendor 1's AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Polynomials and Factoring, ideally 18.08% of the students should score at the Low-level (a substantially lower difference of 16.91% as compared to the original 49.92% of the students who scored at the Low-level); 44.43% should score at the Mid-level (a higher difference of 9.44% when compared to the original 34.99% of the students who scored at the Mid-level); and, 37.49% should score at the High-level (a substantially higher difference of 22.40% as compared to the original 15.09% of the students who scored at the High-level). The simulated results suggest that, if Vendor 1's AI-ALS could ideally make it easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Linear Equations, ideally 45.11% of the students should score at the Low-level (a lower difference of 4.81% when compared to the original 49.92% of the students who scored at the Low-level); 39.64% should score at the Mid-level (a lower difference of 4.65% when compared to the original 34.99% of the students who scored at the Mid-level); and, 15.25% should score at the High-level (an almost negligible higher difference of 0.16% when compared to the original 15.09% that scored at the High-level). The simulated results suggest that, if Vendor 1's AI-ALS could ideally make it slightly easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Functions and Lines, ideally 10.27% of the students should score at the Low-level (an almost negligible higher difference of 0.15% as compared to the original 10.12% of the students who scored at the Low-level); 33.18% should score at the Mid-level (a lower difference of 1.81% when compared to the original 34.99% of the students who scored at the Mid-level); and, 56.55% should score at the High-level (a higher difference of 1.66% compared to the original 54.89% of the students who scored at the High-level). The simulated results suggest that, if Vendor 1's AI-ALS could ideally make it slightly easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Exponents and Exponential Functions, ideally 18.08% should score at the Low-level (a lower difference of 1.99% as compared to the original 20.07% of the students who scored at the Low-level); 44.75% should score at the Mid-level (a higher difference of 4.78% when compared to the original 39.97% of the students who scored at the Mid-level); and, 37.17% should score at the High-level (a lower difference of 2.8% compared to the original 39.97% of the students who scored at the High-level). The simulated results suggest that, if Vendor 1's AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Arithmetic Readiness, ideally 16.26% should score at the Low-level (a slightly higher difference of 1.17% as compared to the original 15.09% of the students who scored at the Low-level); 37.61% should score at the Mid-level (a higher difference of 2.62% when compared to the original 34.99% of the students who scored at the Mid-level); and, 46.12% should score at the High-level (a lower difference of 3.8% compared to the original 49.92% scored at the High-level). The simulated results suggest that, if Vendor 1's AI-ALS could ideally make it slightly

more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Quadratic Functions and Equations, ideally 37.15% of the students should score at the Low-level (a lower difference of 2.82% as compared to the original 39.97% of the students who scored at the Low-level); 24.35% should score at the Mid-level (an almost negligible lower difference of 0.69% when compared to the original 25.04% who scored at the Mid-level); and, 38.50% should score at the High-level (a higher difference of 3.51% as compared to the original 34.99% of the students who scored at the High-level). The simulated results suggest that if Vendor 1's AI-ALS could ideally make it slightly easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS by Vendor 1, in the average number of hours spent by each student, ideally 26.03% of the students should be at the Low-level (a lower difference of 3.99% as compared to the original 30.02% of the students who were at the Low-level); 28.54% should be at the Mid-level (a higher difference of 3.5% when compared to the original 25.04% of the students who were at the Mid-level); and 45.43% should be at the High-level (an almost negligible higher difference of 0.49% as compared to the original 44.94% who were at the High-level). The simulated results suggest that more time spent using the AI-ALS might contribute to their probability of scoring at the High-level in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the percentage of the total number of topics that were mastered by the students in the AI-ALS by Vendor 1, ideally 28.05% of the students should be at the Low-level (a slightly lower difference of 1.97% as compared to the original 30.02% of the students who were at the Low-level); 48.74% should be at the Mid-level (a higher difference of 3.8% compared to the original 44.94% who were at the Mid-level); and, 23.21% should be at the High-level (a lower difference of 1.83% when compared to the original 25.04% who were at the High-level). The simulated results suggest that Vendor 1's AI-ALS was effective in providing adaptive learning to the students and was contributing well to their probability of scoring high marks in the paper-based Post-test.

5.2. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 2

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 2, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions needed in the AI-ALS from Vendor 2 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

To predict the conditions that would enable 100% of the students in Class 2 who had used Vendor 2's AI-ALS to score at the High-level in the paper-based Post-test, hard evidence was set on it (by double-clicking on the High-level histogram bar in Bayesialab). The following counterfactually simulated results of the score-clusters were observed (see Figure 17):

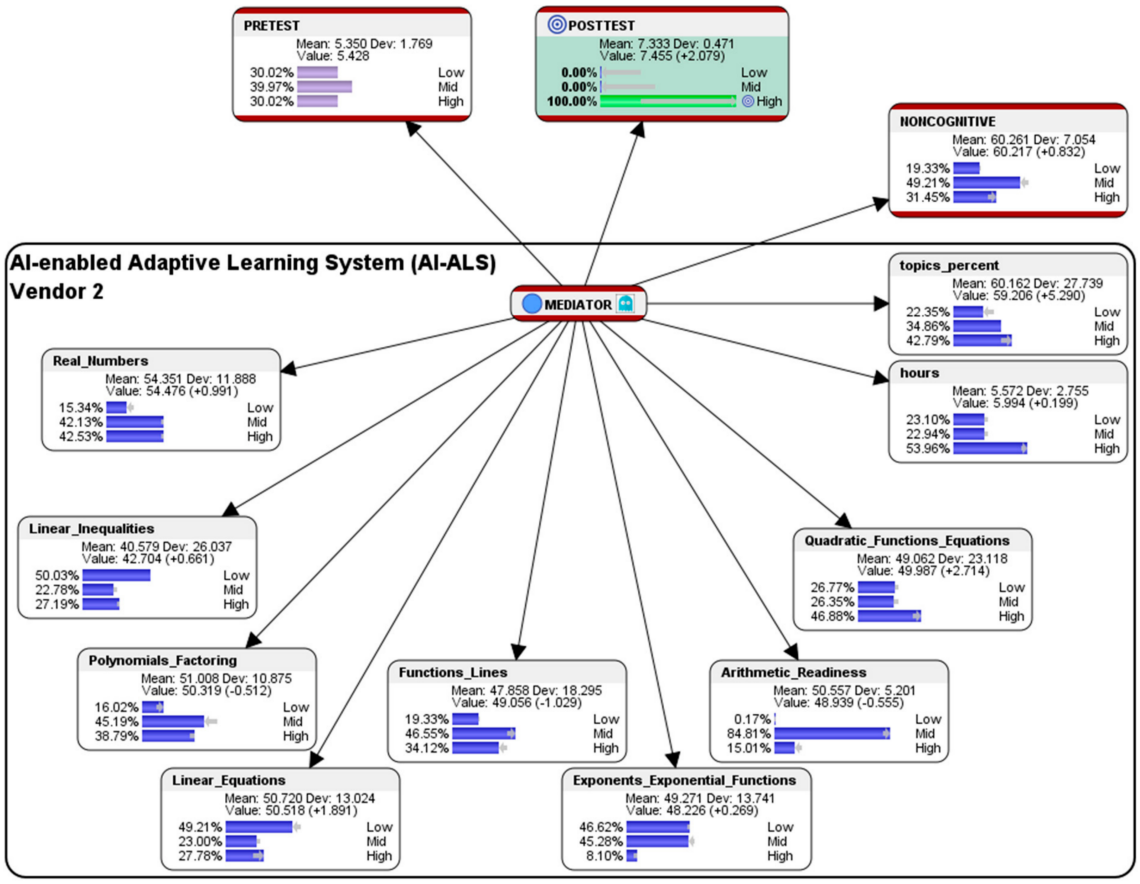

Figure 17. Simulation of counterfactual results for 100% of the students who had used Vendor 2's AI-ALS to score at the high-level in the post-test.

Within the AI-ALS from Vendor 2, in the aggregated Noncognitive factor, ideally 19.33% of the students who had used the AI-ALS from Vendor 2 should be at the so-called Low-level (an almost negligible lower difference of 0.74% as compared to the original 20.07% of the students who were at the Low-level); 49.21% should be at the Mid-level (a higher difference of 5.68% when compared to the original 54.89% of students who were at the Mid-level); and, 31.45% should be at the High-level (a higher difference of 6.41% as compared to the original 25.04% of students who were at the High-level). The counterfactual results suggest that, if the mid-level and high-level of noncognitive attributes (e.g., emotional intelligence to manage stress, interest in learning mathematics, motivation, level of engagement, etc.) could be increased, it might contribute to their probability of scoring at the High-level in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Real Numbers, ideally 15.34% of the students should score at the Low-level (a slightly lower difference of 4.73% as compared to the original 20.07% of the students who scored at the Low-level); 42.13% of the students should score at the Mid-level (a slightly higher difference of 2.16% when compared to the original 39.97% of the students who scored at the Mid-level); and, 42.53% of the students should score at the High-level (a slightly higher difference of 2.56% as compared to the original 39.97% of the students who scored at the High-level. The simulated counterfactual results for the topic of Real Numbers suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Linear Inequalities, ideally 50.03% of the students should score at the Low-level (an almost negligible higher difference of 0.11% as compared to the original 49.92% of the students who scored at the Low-level); 22.78% of the students should score at the Mid-level (a slightly lower difference of 2.26% when compared to the original 25.04% of the

students who scored at the Mid-level); and, 27.19% of the students should score at the High-level (a slightly higher difference of 2.15% as compared to the original 25.04% of the students who scored at the High-level. The simulated counterfactual results for the topic suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Polynomials and Factoring, ideally 16.02% of the students should score at the Low-level (a lower difference of 5.90% as compared to the original 10.12% of the students who scored at the Low-level); 45.19% should score at the Mid-level (a substantially lower difference of 9.70% when compared to the original 54.89% of the students who scored at the Mid-level); and, 38.79% should score at the High-level (a slightly higher difference of 3.80% compared to the original 34.99% of the students who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Linear Equations, ideally 49.21% of the students should score at the Low-level (a lower difference of 5.68% when compared to the original 54.89% of the students who scored at the Low-level); 23.00% should score at the Mid-level (a lower difference of 2.04% when compared to the original 25.04% of the students who scored at the Mid-level); and 27.78% should score at the High-level (a higher difference of 7.71% when compared to the original 20.07% who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Functions and Lines, ideally 19.33% of the students should score at the Low-level (an almost negligible lower difference of 0.74% as compared to the original 20.07% of the students who scored at the Low-level); 46.55% should score at the Mid-level (a higher difference of 6.58% when compared to the original 39.97% of the students who scored at the Mid-level); and, 34.12% should score at the High-level (a lower difference of 5.85% compared to the original 39.97% of the students who scored at the High-level). The simulated results suggest that if Vendor 2's AI-ALS could ideally make it more difficult for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Exponents and Exponential Functions, ideally 46.62% should score at the Low-level (a slightly higher difference of 1.68% when compared to the original 44.94% of the students who scored at the Low-level); 45.28% should score at the Mid-level (a lower difference of 4.64% as compared to the original 49.92% of the students who scored at the Mid-level); and, 8.10% should score at the High-level (a lower difference of 2.96% compared to the original 5.14% of the students who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Arithmetic Readiness, ideally 0.17% should score at the Low-level (a difference of 0.00% as compared to the original 0.17% of the students who scored at the Low-level); 84.81% should score at the Mid-level (a higher difference of 5.04% compared to the original 79.77% of the students who scored at the Mid-level); and, 15.01% should score at the High-level (a lower difference of 5.06% compared to the original 20.07% who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Quadratic Functions and Equations, ideally 26.77% of the students should score at the Low-level (a lower difference of 3.25% when compared to the original 30.02% of the students who scored at the Low-level); 26.35% should score at the Mid-level

(a lower difference of 3.67% as compared to the original 30.02% who scored at the Mid-level); and, 46.88% should score at the High-level (a higher difference of 6.91% when compared to the original 39.97% of the students who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS by Vendor 2, in the average number of hours spent by each student, ideally 23.10% of the students should be at the Low-level (a lower difference of 1.94% as compared to the original 25.04% of the students who were at the Low-level); 22.94% should be at the Mid-level (a slightly lower difference of 2.10% when compared to the original 25.04% of the students who were at the Mid-level), and, 53.96% should be at the High-level (a slightly higher difference of 4.04% as compared to the original 49.92% who were at the High-level). The simulated results suggest that if the students could spend more time learning mathematics within Vendor 2's AI-ALS, it could contribute to their probability of scoring at the High-level in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the percentage of the total number of topics that were mastered by the students, ideally 22.35% of the students should be at the Low-level (a slightly higher difference of 7.67% as compared to the original 30.02% of the students who were at the Low-level); 34.86% should be at the Mid-level (an almost negligible lower difference of 0.13% compared to the original 34.99% who were at the Mid-level); and, 42.79% should be at the High-level (a higher difference of 7.8% compared to the original 34.99% who were at the High-level). The simulated results suggest that if the students could master a higher percentage of topics within Vendor 2's AI-ALS, it could contribute to their probability of scoring at the High-level in the paper-based Post-test.

5.3. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 3

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 3, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions that are needed in the AI-ALS from Vendor 3 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

Here is an opportunity that the following analysis can be used as a starting point for discussions to foster strategic coordination between the educational stakeholders and Vendor 3 which provided the AI-ALS. As previously observed in Table 1 and Figure 15, there was a decrease in the number of students who scored at the High-level of the marks in the paper-based post-test. Realistically, since the algorithm with which the AI-ALS from Vendor 3 interacts with the students cannot be changed much, if at all, the mathematics teacher would have to provide remediation for the students. The AI-ALS from Vendor 3 might not be a good choice in the selection for in-service deployment from the perspective of the policy makers and educational stakeholders, as it might be realistically impractical to ask Vendor 3 to change their proprietary algorithm to suit the students of Class 3. However, the simulated counterfactual results (see Figure 18) could still be used as a guide for remediation by the teacher to "level-up" the students in the mathematics topics that they might be weaker in.

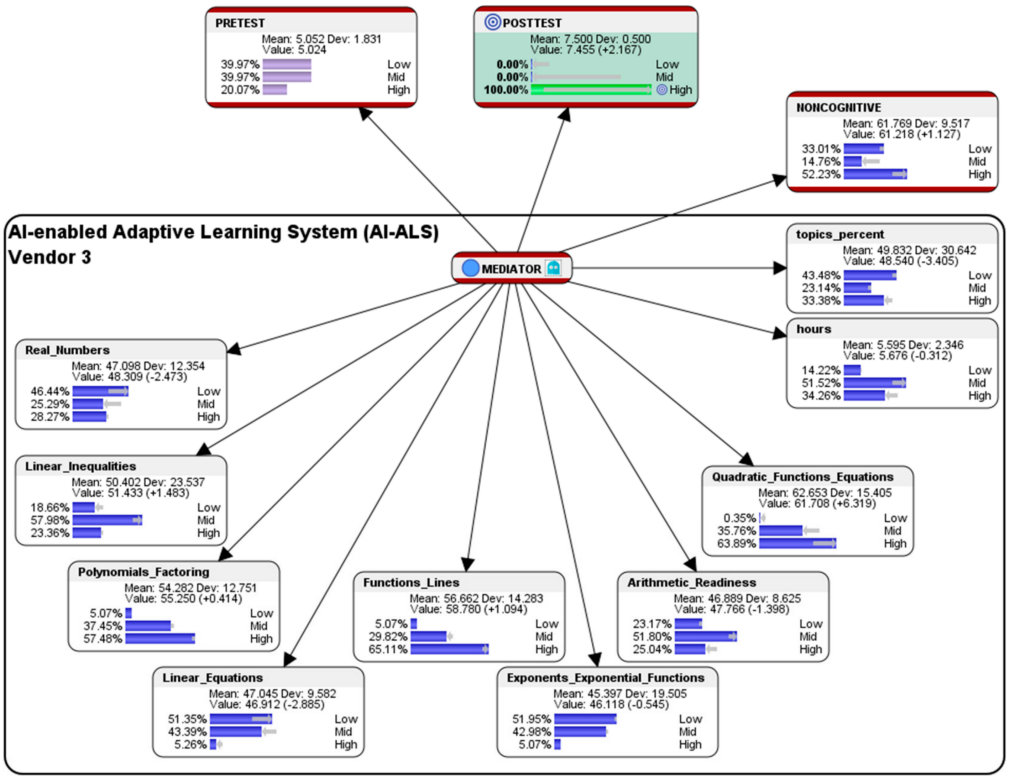

Figure 18. Simulation of counterfactual results for 100% of the students who had used Vendor 3's AI-ALS to score at the high-level in the post-test.

To predict the conditions that would enable 100% of the students in Class 3 who had used Vendor 3's AI-ALS to score at the High-level in the paper-based Post-test, hard evidence was set on it (by double-clicking on the High-level histogram bar in Bayesialab). The following counterfactually simulated results of score-clusters were observed (see Figure 18):

In the aggregated Noncognitive factor, ideally 33.01% of the students who had used the AI-ALS from Vendor 3 should be at the so-called Low-level (a slightly higher difference of 2.99% compared to the original 30.02% of the students who were at the Low-level); 14.76% should be at the Mid-level (a substantially lower difference of 15.26% as compared to the original 30.02% of students who were at the Mid-level); and, 50.23% should be at the High-level (a substantially higher difference of 10.26% as compared to the original 39.97% of students who were at the High-level). The results suggest that noncognitive factors of the students such as motivation, interest, attitude towards mathematics, etc. might need to be improved.

Within the AI-ALS from Vendor 3, in the topic of Real Numbers, ideally 46.44% of the students should score at the Low-level (a substantially higher difference of 16.42% when compared to the original 30.02% of the students who scored at the Low-level), 25.29% of the students should score at the Mid-level (a substantially lower difference of 14.68% as compared to the original 39.97% of the students who scored at the Mid-level), and 28.27% of the students should score at the High-level (a slightly lower difference of 1.75% when compared to the original 30.02% of the students who scored at the High-level. The simulated results suggest that, if Vendor 3's AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Linear Inequalities, ideally 18.66% of the students should score at the Low-level (a lower difference of 6.38% when compared to the original 25.04% of the students who scored at the Low-level); 57.98% of the students should score at the Mid-level (a higher difference of 8.06% as compared to the original 49.92% of the students who scored at the Mid-level); and, 23.36% of the students should score at the High-level (a slightly lower difference of

9.74% as compared to the original 25.04% of the students who scored at the High-level. The simulated results suggest that, ideally, if Vendor 3's AI-ALS could make it slightly more difficult for students to score at the High-level, but yet, not so difficult that students find it too challenging to score at the Mid-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Polynomials and Factoring, ideally 5.07% of the students should score at the Low-level (an almost negligible lower difference of 0.07% when compared to the original 5.14% of the students who scored at the Low-level); 37.45% should score at the Mid-level (an almost negligible lower difference of 0.04% as compared to the original 39.97% of the students who scored at the Mid-level); and, 57.48% should score at the High-level (a slightly higher difference of 2.59% when compared to the original 54.89% of the students who scored at the High-level). The simulated results suggest that Vendor 3's AI-ALS might already be close to optimally adapting to the students in Class 3 in training them to score at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Linear Equations, ideally 51.35% of the students should score at the Low-level (a substantially higher difference of 16.36% when compared to the original 34.99% of the students who scored at the Low-level); 43.39% should score at the Mid-level (a substantially lower difference of 11.5% when compared to the original 54.89% of the students who scored at the Mid-level); and, 5.26% should score at the High-level (a lower difference of 4.86% when compared to the original 10.12% scored at the High-level). The simulated results suggest that, if Vendor 3's AI-ALS could ideally make it much more difficult for students to score at the High-level and at the Mid-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Functions and Lines, ideally 5.07% of the students should score at the Low-level (an almost negligible lower difference of 0.07% when compared to the original 5.14% of the students who scored at the Low-level); 29.82% should score at the Mid-level (a lower difference of 5.17% as compared to the original 34.99% of the students who scored at the Mid-level); and, 65.11% should score at the High-level (a higher difference of 5.24% compared to the original 59.87% of the students who scored at the High-level). The simulated results suggest that if Vendor 3's AI-ALS could ideally make it slightly easier for students in Class 3 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Exponents and Exponential Functions, ideally 51.95% should score at the Low-level (a lower difference of 2.03% as compared to the original 49.92% of the students who scored at the Low-level); 42.98% should score at the Mid-level (a lower difference of 6.94% when compared to the original 49.92% of the students who scored at the Mid-level); and, 5.07% should score at the High-level (an almost negligible lower difference of 0.07% when compared to the original 5.14% of the students who scored at the High-level). The simulated results suggest that Vendor 3's AI-ALS might already be close to optimally adapting to the students in Class 3 in training them to score at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Arithmetic Readiness, ideally 23.17% should score at the Low-level (a slightly higher difference of 3.1% as compared to the original 20.07% of the students who scored at the Low-level); 51.80% should score at the Mid-level (a higher difference of 6.86% when compared to the original 44.94% of the students who scored at the Mid-level); and, 25.04% should score at the High-level (a lower difference of 9.95% as compared to the original 34.99% scored at the High-level). The simulated results suggest that if Vendor 3's AI-ALS could ideally make it much more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Quadratic Functions and Equations, ideally 0.35% of the students should score at the Low-level (a lower difference of 4.79% as compared to the original 5.14% of the students who scored at the Low-level); 35.76% should score at the Mid-level

(a substantially lower difference of 14.16% when compared to the original 49.92% who scored at the Mid-level); and, 63.89% should score at the High-level (a substantially higher difference of 18.95% as compared to the original 44.94% of the students who scored at the High-level). The simulated results suggest that if Vendor 3's AI-ALS could ideally make it much easier for students in Class 3 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS by Vendor 3, in the average number of hours were spent by each student, 14.22% of the students should be at the Low-level (an almost negligible lower difference of 0.87% compared to the original 15.09% of the students who were at the Low-level), 51.52% should be at the Mid-level (a substantially higher difference of 11.55% as compared to the original 39.97% of the students who were at the Mid-level), and 34.26% should be at the High-level (a substantially lower difference of 10.68% compared to the original 44.94% who were at the High-level). The simulated results suggest that if students spend less time within Vendor 3's AI-ALS, it could contribute to their probability of scoring at the High-level in the paper-based Post-test. Perhaps, one way of interpreting this could be: spending less time within the Vendor 3's AI-ALS could help prevent diminishing marginal returns, as the students would not have to suffer from undue fatigue or stress.

In the percentage of the total number of topics that were mastered by the students in the AI-ALS by Vendor 3, ideally 43.48% of the students should be at the Low-level (a slightly lower difference of 3.51% as compared to the original 39.97% of the students who were at the Low-level); 23.14% should be at the Mid-level (a slightly higher difference of 3.07% when compared to the original 20.07% who were at the Mid-level); and 33.38% should be at the High-level (a lower difference of 6.59% as compared to the original 39.97% who were at the High-level). The simulated results suggest that mastering the topics at a slower pace within Vendor 3's AI-ALS could contribute to their probability of scoring at the High-level in the paper-based Post-test. At first glance, this might seem counterintuitive. However, one way of interpreting this might be: a slower pace of mastering the mathematics topics could be more beneficial, as it could potentially contribute to a deeper level of understanding of the subject matter by the students.

5.4. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 4

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 4, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions needed in the AI-ALS from Vendor 4 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

The following simulated counterfactual results for the conditions that would optimize the performance of students who had used the AI-ALS from Vendor 4 (see Figure 19) and Vendor 5 (see Figure 20) would only be presented in summarized graphical form due to space constraints for publication, since they also had positive gains in the High-level marks in the Post-test (as presented earlier in Table 1 and Figure 15), and they could be considered to be similar to the case in which the students had used the AI-ALS from Vendor 1 (see Section 5.1).

Overall, within the AI-ALS from Vendor 4, the simulated counterfactual results suggest that, in order to train them in score at the High-level in the paper-based Post-test, the finer details of the predictions that recommend whether it should be made easier or more difficult in the various mathematics topics could be perused in Figure 19.

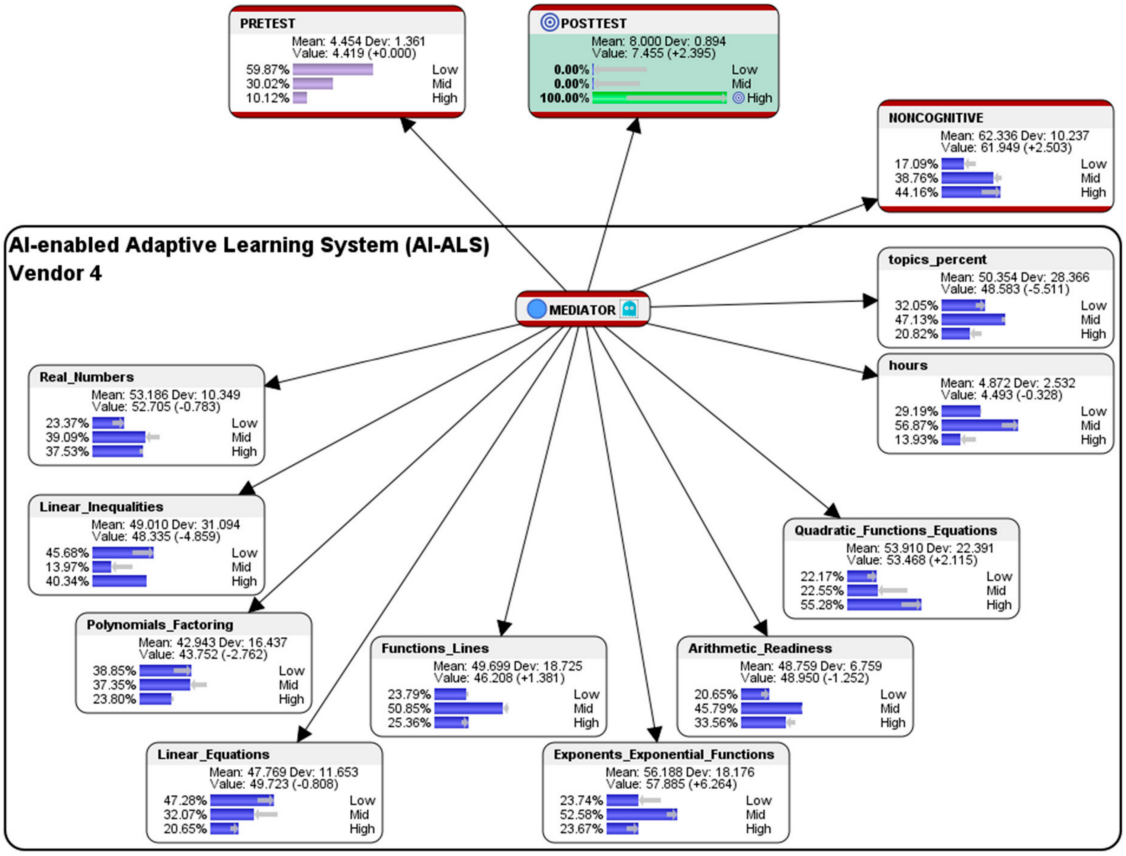

Figure 19. Simulation of counterfactual results for 100% of the students who had used Vendor 4's AI-ALS to score at the high-level in the post-test. Grey arrows recommended whether there should be an increase (pointing to the right), or a decrease (pointing to the left) in each respective mathematics topic's score-clusters for Low-, Mid-, and High-level.

5.5. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 5

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 5, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions that are needed in the AI-ALS from Vendor 5 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the Post-test?

Overall, within the AI-ALS from Vendor 5, the simulated counterfactual results suggest that, in order to train them in score at the High-level in the paper-based Post-test, the finer details of the predictions that recommend whether it should be made easier or more difficult for the various mathematics topics could be perused in Figure 20.

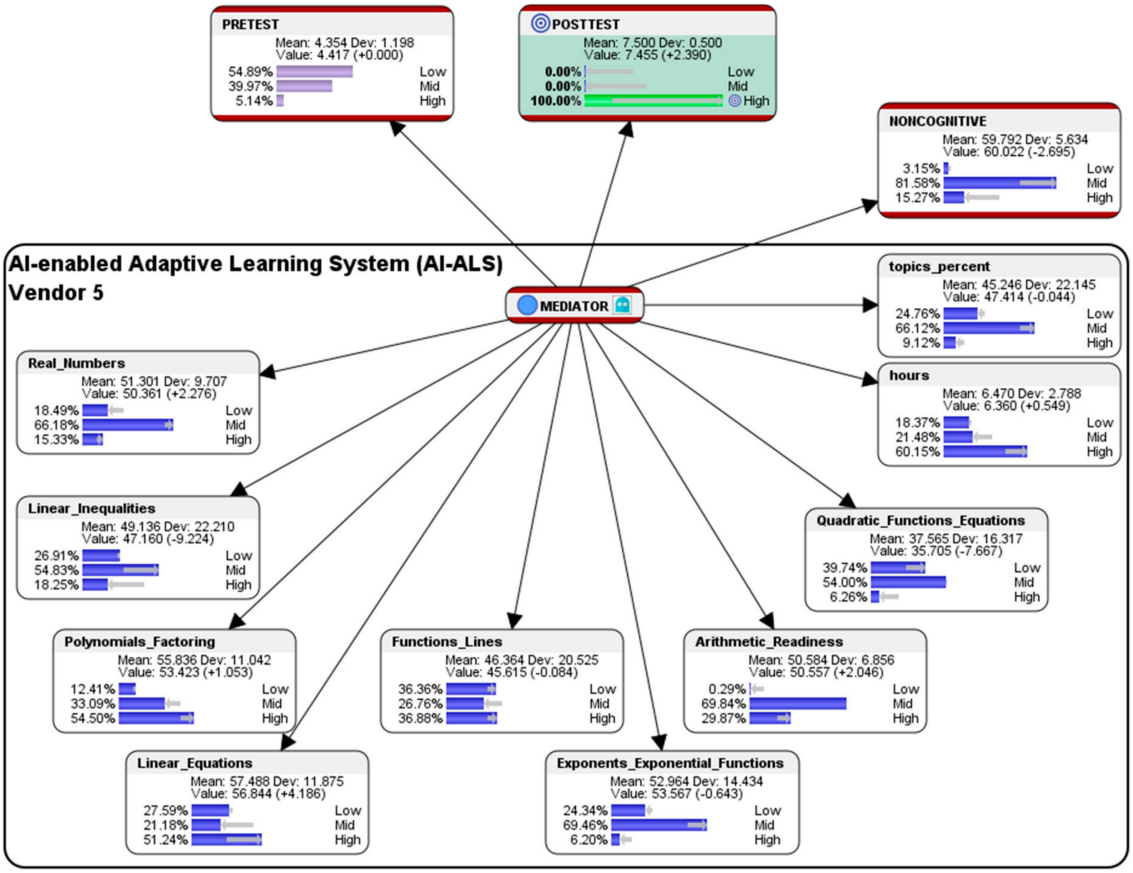

Figure 20. Simulation of counterfactual results for 100% of the students who had used Vendor 5's AI-ALS to score at the high-level in the post-test. Grey arrows recommended whether there should be an increase (pointing to the right), or a decrease (pointing to the left) in each respective mathematics topic's score-clusters for Low-, Mid-, and High-level.

6. Evaluation of the Predictive Performance of the Bayesian Network Model

The predictive performance of a model could be evaluated by using measurement tools, such as the gains curve (see Figure 21), the lift curve (see Figure 22), and via cross-validation by bootstrapping to 100,000 samples (see Figure 23).

6.1. Gains Curve

In the Gains curve (see Figure 21), there were around 21% of participants with the target value >6 for the Post-test (yellow line intercepting with the % total axis). The blue diagonal line represented the gains curve of a pure random policy, which refers to prediction without this predictive model. The red lines represented the gains curve while using this predictive model, which was observed to be above the blue diagonal line. The Gini index of 12.73% and relative Gini index of 16.32% suggested that the gains of using this predictive model vis-à-vis not using it was acceptable.

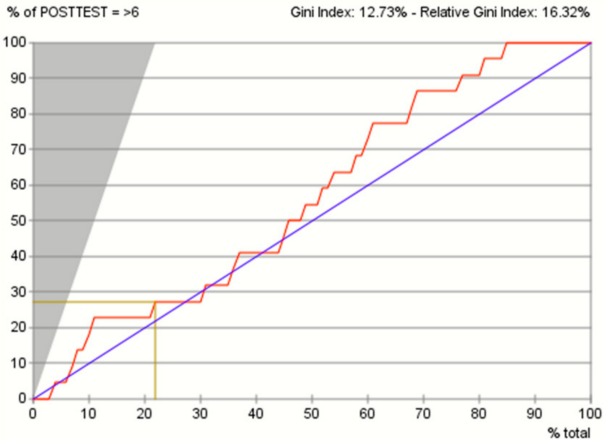

Figure 21. Gains curve.

6.2. Lift Curve

The lift curve (see Figure 22) was built upon the results from the gains curve (see Figure 21). The value of the best lift around 21% was interpreted as the ratio between 100% and 2.07% (optimal policy divided by random policy). The lift decreased when more than 2.07% of the participants were considered and was close to 1 when all the participants were considered. The lift index of 1.1257 and relative lift index of 45.09% suggested that the performance of this predictive model was acceptable.

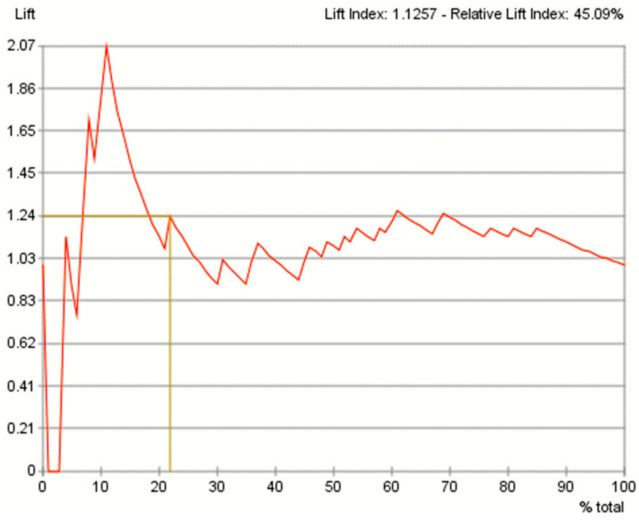

Figure 22. Lift curve.

6.3. Target Evaluation Cross-Validation by Statistical Bootstrapping of the Data 100,000 Times in Every Node

The purpose of this section is not to laud the effectiveness of the BN model that was used in the exemplars thus far, but to illustrate how the evaluation of the model can be done in Bayesialab. Therefore, the results will be honestly reported, regardless of whether it is good or bad. Bootstrapping to 100,000 times in every node would ensure that it is statistically sufficiently large enough for generating a parametric data distribution. As observed in the results that were generated by Bayesialab after performing bootstrapping 100,000 times on the data distribution of each node in the BN by using the Naïve Bayes algorithm, the Overall Precision was 65.8963%; the Mean Precision was 63.7718%; the Overall Reliability was 65.7522%; the Mean Reliability was 64.3817%; the Mean Gini Index was 55.1787%; the Mean Relative Gini Index was 69.9366%; the Mean Lift Index was 2.0297; the Mean Relative Lift Index was 79.6543%; the Mean ROC Index was 84.9939%; the Mean Calibration Index

was 56.0688%; the Mean Binary Log-Loss was 0.3619; the Correlation Coefficient R was 0.5096; the Coefficient of Determination R2 was 0.2597; the RMSE was 1.3883; and, the NRSME was 19.8329%. These results suggested that the predictive performance of the BN model could be considered to be acceptable.

Figure 23 presents a confusion matrix after bootstrapping the data 100,000 times in every node of the BN model. The confusion matrix provided additional information regarding the computational model's predictive performance. The leftmost column in the matrix contained the predicted values, while the actual values in the data were presented in the top row. The following three confusion matrix views would be available by clicking on the corresponding tabs. The Occurrences Matrix (see Figure 23) would indicate the number of cases for each combination of the predicted versus actual values. The diagonal shows the number of true positives. The Reliability Matrix (see Figure 24) would indicate the probability of the reliability of the prediction of a state in each cell. Reliability measures the overall consistency of a prediction. A prediction could be considered to be highly reliable if the computational model could produce similar results under consistent conditions. The Precision Matrix (see Figure 25) would indicate the probability of the precision of the prediction of a state in each cell. Precision is the measure of the overall accuracy which the computational model can correctly predict.

Confusion Matrix

Occurrences | Reliability | Precision

Value	Low (3131011)	Mid (4746574)	High (2122415)
Low (3106141)	1915086	859161	331894
Mid (4895543)	861950	3458809	574784
High (1998316)	353975	428604	1215737

Figure 23. Occurrences confusion matrix after performing target evaluation cross-validation by bootstrapping the data in each node 100,000 times.

Occurrences | Reliability | Precision

Value	Low (3131011)	Mid (4746574)	High (2122415)
Low (3106141)	61.65%	27.66%	10.69%
Mid (4895543)	17.61%	70.65%	11.74%
High (1998316)	17.71%	21.45%	60.84%

Figure 24. Reliability confusion matrix after performing target evaluation cross-validation by bootstrapping the data in each node 100,000 times.

Occurrences | Reliability | Precision

Value	Low (3131011)	Mid (4746574)	High (2122415)
Low (3106141)	61.17%	18.1%	15.64%
Mid (4895543)	27.53%	72.87%	27.08%
High (1998316)	11.31%	9.03%	57.28%

Figure 25. Precision confusion matrix after performing target evaluation cross-validation by bootstrapping the data in each node 100,000 times.

6.4. Limitations of the Study

The exploratory nature of predictive analytics in this study using BN modeling renders the simulated counterfactual results suggestive, rather than conclusive. Further, it is only applicable to

this BN model, which was generated from the current datasets. Therefore, caution must be exercised when interpreting the potential relationships between the variables (nodes) in the BN model.

The current study only utilized 100 students' data. However, the Bayesian approach that is delineated in the current paper could still be used as an alternative approach by educational stakeholders in small-scale pilot studies to independently explore the pedagogical motifs of any AI-ALS, in order to coordinate the analyses of datasets procured from the servers different AI-AL, and to strategically educe (draw out) the problem-solving abilities of the students.

The Bayesian network model that was used in the current study was based on the Naïve Bayes algorithm, as it is suitable for exploratory studies that do not assume relations between nodes to be causal in nature. As in any study that involves simulations, the results are dependent on the dataset that generated the computational model. Moreover, educational stakeholders and researchers should consider alternative models that could better depict the relations between the variables in the dataset.

Thus far, the tools for descriptive as well as predictive analytics, and the tools in Bayesialab that could be used for the evaluation of the predictive performance of the BN model have been clearly depicted. The limitations of the study have also been described. In the next section, the discussion and concluding remarks will be presented.

7. Discussion and Concluding Remarks

Strategic coordination between schools to analyze the AI-ALS that they are using could yield useful information for educational stakeholders. The current paper has put forth a Bayesian approach for educational stakeholders to independently explore the underlying pedagogical motifs of five different AI-ALS. Even in realistic school situations where the number of students in classes might be low, and even if there is no control group, the Bayesian implementation of Response Surface Methodology [66–69] could still be used to keep individual parameters constant, whilst others could be changed to simulate different hypothetical scenarios. Specific examples have been provided to demonstrate how this AI-based Bayesian approach could be used to analyze the underlying pedagogical motifs of five AI-ALS that were used in five different schools. Potentially, these hypothetical scenarios with fully controllable parameters could be used to better inform educational stakeholders about the suitability of each AI-ALS for broader adoption after the pilot study.

Beyond the conventional observation of gains in the cognitive pre-test vis-à-vis post-test, this proposed Bayesian approach also generated hypothetical scenarios that might be of interest for noncognitive researchers to consider in future studies. The implication for education is that the AI-ALS should not be solely relied upon to improve the students' learning of mathematics; rather, the gaps in the learning of mathematical concepts that the AI-ALS could not bridge for the students should be addressed by their mathematics teachers. For example, if the student had scored low marks in the AI-ALS, but could surprisingly score high marks in the paper-based post-test, it might be due to the opportunities that were provided by the AI-ALS to the student to experience vicarious trial and error (VTE). Hence, active inference [36] could be successfully accomplished to solve similar problems in the paper-based post-test. Conversely, if the student could score high-level marks in the AI-ALS for a particular mathematics topic, but could not do so for the paper-based post-test, the teacher should intervene to find out why the student was unable to accomplish active inference from the concepts that were taught by the AI-ALS to the paper-based post-test.

The call by the Foresight Institute [30] and other researchers to study the machine behavior of artificial superintelligence has provided an inspirational impetus to embark on the research outlined in the current paper. To help the reader envision how explainable AI technology could be harnessed to better understand the computational results produced by artificial superintelligence, a human-centric analytical approach based on BN has been proffered. After discursive reasonings of the analytical results by the educational stakeholders, future-ready actionable advice could be engendered to assist teachers in bridging the gaps in the learning process for their students. With this approach, policy makers could also be better informed regarding the use of AI in education. Usage of explainable

AI technology in this BN approach empowers us to gain insights from the past (via descriptive analytics of "what has happened"), enabling us to look beyond the horizon of the present, and peer into alternative variations of the future (via "what-if" predictive analytics of simulated hypothetical scenarios). While facing off a relentless T-800 in the movie *Terminator*, Sarah Connor defiantly seethed, "The future is not set." Knowing about the potential behavior of AI systems under various different conditions using this future-ready approach could also allow us to defy the odds, and turn them in our favor, regardless of which AI systems the schools choose to deploy.

Supplementary Materials: As mentioned in Section 4.1 of the current paper, the zip file containing the datasets can be downloaded from https://doi.org/10.6084/m9.figshare.8206976.

Funding: This research received funding provided by the Education Research Funding Programme (ERFP) via the Office of Education Research in the National Institute of Education, Nanyang Technological University, Singapore. [grant number: ERFP OOE]

Acknowledgments: The author sincerely thanks the editors, the staff of the journal, and the anonymous reviewers for helping to improve the manuscript. This paper and the research behind it would not be possible without the visionary leadership of David Hung, and the exceptional support of Eva Moo, Lek-Hong Teo, and Shih-Fen Wah from the Office of Education Research in the National Institute of Education of Nanyang Technological University Singapore. The author is grateful to the Education Technology Division of the Ministry of Education of Singapore for providing guidance about artificial intelligence in education and adaptive learning technology. The author is especially thankful to Sin-Mei Cheah of Singapore Management University for proofreading the drafts of the manuscript.

Conflicts of Interest: The author of this manuscript declares that there is no conflict of interest.

Ethical Statement: The identities of the students and the AI-ALS vendors have been anonymized. Further, synthetic data has been used in the examples in this paper, so that any good (or bad) performances by the students in the AI-ALS, or in the pre- or post-test will not be unfairly attributed to the level of quality of the AI-ALS, or to the students.

Appendix A

Table A1. Codebook of the columns in the dataset, each of which will become a node in the BN model.

Node/Column Name	Description
student_id	student id
hours	number of hours spent by student using the (AI-ALS) AI-enabled Adaptive Learning System
topics_350	number of topics out of a total of 350 completed by the student in the AI-ALS
topics_percent	percentage of topics completed by the student in the AI-ALS
Arithmetic Readiness (AR)	
AR_FMEF_P	AR_Factors_Multiples_Equivalent_Fractions_Passed
AR_FMEF_RL	AR_Factors_Multiples_Equivalent_Fractions_Ready_For_Learning
AR_ASF_P	AR_Addition_Subtraction_with_Fractions_Passed
AR_ASF_RL	AR_Addition_Subtraction_with_Fractions_Ready_for_Learning
AR_MD_P	AR_Multiplication_Division_with_Decimals_Passed
AR_MD_RL	AR_Multiplication_Division_with_Decimals_Ready_for_Learning
AR_MN_P	AR_Mixed_Numbers_Passed
AR_MN_RL	AR_Mixed_Numbers_Ready_for_Learning
AR_RONL_P	AR_Rounding_Number Line_Passed
AR_RONL_RL	AR_Rounding_Number Line_Ready_for_Learning
AR_ASD_P	AR_Addition_Subtraction_with_Decimals_Passed
AR_ASD_RL	AR_Addition_Subtraction_with_Decimals_Ready_for_Learning
AR_MDD_P	AR_Multiplication_Division_with_Decimals_Passed
AR_MDD_RL	AR_Multiplication_Division_with_Decimals_Ready_for_Learning
AR_CBFD_P	AR_Converting_Between_Fractions_Decimals_Passed
AR_CBFD_RL	AR_Converting_Between_Fractions_Decimals_Ready_for_Learning
AR_RUR_P	AR_Ratios_Unit_Rates_Passed
AR_RUR_RL	AR_Ratios_Unit_Rates_Ready_for_Learning
AR_PDF_P	AR_Percents_Decimals_Fractions_Passed
AR_PDF_RL	AR_Percents_Decimals_Fractions_Ready_for_Learning
AR_IPA_P	AR_Intro_Percent_Applications_Passed
AR_IPA_RL	AR_Intro_Percent_Applications_Ready_for_Learning
AR_UM_P	AR_Units_Measurement_Passed
AR_UM_RL	AR_Units_Measurement_Ready_for_Learning

Table A1. *Cont.*

Node/Column Name	Description
Real Numbers (RN)	
RN_PLOT_P	RN_Plotting_Ordering_Passed
RN_PLOT_RL	RN_Plotting_Ordering_Ready_for_Learning
RN_OSN_P	RN_Operations_Signed_Numbers_Passed
RN_OSN_RL	RN_Operations_Signed_Numbers_Ready_for_Learning
RN_EOO_P	RN_Exponents_Order_Operations_Passed
RN_EOO_RL	RN_Exponents_Order_Operations_Ready_for_Learning
RN_EE_P	RN_Evaluation_Expressions_Operations_Passed
RN_EE_RL	RN_Evaluation_Expressions_Ready_for_Learning
RN_VDSRN_P	RN_Venn_Diagrams_Sets_Real_Num_Passed
RN_VDSRN_RL	RN_Venn_Diagrams_Sets_Real_Num_Ready_for_Learning
RN_PROP_O_P	RN_Properties_Operations_Passed
RN_PROP_O_RL	RN_Properties_Operations_Ready_for_Learning
RN_OSLE_P	RN_One_Step_Linear_Equations_Passed
RN_OSLE_RL	RN_One_Step_Linear_Equations_Ready_for_Learning
Linear Equations (LE)	
LE_MSLE_P	LE_Multi_Step_Linear_Equations_Passed
LE_MSLE_RL	LE_Multi_Step_Linear_Equations_Ready_for_Learning
LE_WEE_P	LE_Writing_Expressions_Equations_Passed
LE_WEE_RL	LE_Writing_Expressions_Equations_Ready_for_Learning
LE_ALE_P	LE_Applications_Linear_Equations_Passed
LE_ALE_RL	LE_Applications_Linear_Equations_Ready_for_Learning
LE_SVDA_P	LE_Solving_Variable_Dimensional_Analysis_Passed
LE_SVDA_RL	LE_Solving_Variable_Dimensional_Analysis_Ready_for_Learning
LE_PROP_P	LE_Proportions_Passed
LE_PROP_RL	LE_Proportions_Ready_for_Learning
LE_MP_P	LE_More_Percents_Passed
LE_MP_RL	LE_More_Percents_Ready_for_Learning
LE_PFL_P	LE_Personal_Financial_Literacy_Passed
LE_PFL_RL	LE_Personal_Financial_Literacy_Ready_for_Learning
Linear Inequalities (LI)	
LI_WGI_P	LI_Writing_Graphing_Inequalities_Passed
LI_WGI_RL	LI_Writing_Graphing_Inequalities_Ready_for_Learning
Functions and Lines (FL)	
FL_TGL_P	FL_Tables_Graphs_Lines_Passed
FL_TGL_RL	FL_Tables_Graphs_Lines_Ready_for_Learning
FL_IF_P	FL_Introduction_Functions_Passed
FL_IF_RL	FL_Introduction_Functions_Ready_for_Learning
FL_AS_P	FL_Arithmetic_Sequences_Passed
FL_AS_RL	FL_Arithmetic_Sequences_Ready_for_Learning
Exponents and Exponential Functions (EEF)	
EEF_PPQR_P	EEF_Product_Power_Quotient_Rules_Passed
EEF_PPQR_RL	EEF_Product_Power_Quotient_Rules_Ready_for_Learning
EEF_IR_P	EEF_Intro_Radicals_Passed
EEF_IR_RL	EEF_Intro_Radicals_Ready_for_Learning
Polynomials and Factoring (PE)	
PE_PM_P	PE_Polynomial_Multiplication_Passed
PE_PM_RL	PE_Polynomial_Multiplication_Ready_for_Learning
PF_FGCF_P	PE_Factoring_Greatest_Common_Factor_Passed
PF_FGCF_RL	PE_Factoring_Greatest_Common_Factor_Ready_for_Learning
PF_FQT_P	PE_Factoring_Quadratic_Trinomials_Passed
PF_FQT_RL	PE_Factoring_Quadratic_Trinomials_Ready_for_Learning
PF_FSP_P	PE_Factoring_Special_Products_Passed
PF_FSP_RL	PE_Factoring_Special_Products_Ready_for_Learning
Quadratic Functions and Equations (QFE)	
QFE_SQEF_P	QFE_Solving_Quadratic_Equations_Factoring_Passed
QFE_SQEF_RL	QFE_Solving_Quadratic_Equations_Factoring_Ready_for_Learning
QFE_SRP_P	QFE_Square_Root_Property_Passed
QFE_SRP_RL	QFE_Square_Root_Property_Ready_for_Learning
Pre-test (PRETEST)	synthetic data for Pre-test Questions 1-10
Post-test (POSTTEST)	synthetic data for Post-test Questions 1-10
Noncognitive (NONCOG)	synthetic data for Noncognitive Survey Questions 1-10

References

1. Association of Computing Machinery. A.M. Turing Award Laureate Dr. McCarthy's Lecture "The Present State of Research on Artificial Intelligence". Available online: https://amturing.acm.org/award_winners/mccarthy_1118322.cfm (accessed on 10 July 2019).
2. Yampolskiy, R.V. *Artificial Superintelligence: A Futuristic Approach*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2015; ISBN 1-4822-3443-2.
3. Wogu, I.A.P.; Misra, S.; Assibong, P.A.; Olu-Owolabi, E.F.; Maskeliūnas, R.; Damasevicius, R. Artificial Intelligence, Smart Classrooms and Online Education in the 21st Century: Implications for Human Development. *J. Cases Inf. Technol.* **2019**, *21*, 66–79. [CrossRef]
4. Egoeze, F.; Misra, S.; Maskeliūnas, R.; Damaševičius, R. Impact of ICT on Universities Administrative Services and Management of Students' Records: ICT in University Administration. *Int. J. Hum. Cap. Inf. Technol. Prof.* **2018**, *9*, 1–15. [CrossRef]
5. Wogu, I.A.P.; Misra, S.; Assibong, P.A.; Ogiri, S.O.; Damasevicius, R.; Maskeliunas, R. Super-Intelligent Machine Operations in Twenty-First-Century Manufacturing Industries: A Boost or Doom to Political and Human Development? In *Towards Extensible and Adaptable Methods in Computing*; Chakraverty, S., Goel, A., Misra, S., Eds.; Springer: Singapore, 2018; pp. 209–224, ISBN 9789811323478.
6. Wilson, C.; Scott, B. Adaptive systems in education: A review and conceptual unification. *Int. J. Inf. Learn. Technol.* **2017**, *34*, 2–19. [CrossRef]
7. Nkambou, R.; Mizoguchi, R.; Bourdeau, J. Introduction: What Are Intelligent Tutoring Systems, and Why This Book? In *Advances in Intelligent Tutoring Systems*; Nkambou, R., Mizoguchi, R., Bourdeau, J., Eds.; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2010; Volume 308, ISBN 978-3-642-14362-5.
8. Phobun, P.; Vicheanpanya, J. Adaptive intelligent tutoring systems for e-learning systems. *Procedia-Soc. Behav. Sci.* **2010**, *2*, 4064–4069. [CrossRef]
9. Garrido, A. AI and Mathematical Education. *Educ. Sci.* **2012**, *2*, 22–32. [CrossRef]
10. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
11. Cen, H.; Koedinger, K.R.; Junker, B. Is Over Practice Necessary?-improving learning efficiency with the cognitive tutor through Educational Data Mining. *Front. Artif. Intell. Appl.* **2007**, *158*, 511.
12. VanLehn, K. The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* **2006**, *16*, 227–265.
13. Hawkins, W.J.; Heffernan, N.T.; Baker, R.S.J.D. Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. In *Intelligent Tutoring Systems*; Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8474, pp. 150–155, ISBN 978-3-319-07220-3.
14. Magoulas, G.D.; Papanikolaou, Y.; Grigoriadou, M. Adaptive web-based learning: Accommodating individual differences through system's adaptation. *Br. J. Educ. Technol.* **2003**, *34*, 511–527. [CrossRef]
15. Szafir, D.; Mutlu, B. ARTFul: Adaptive review technology for flipped learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '13, Paris, France, 27 April–2 May 2013; ACM Press: New York, NY, USA, 2013; p. 1001.
16. Rosenberg, L. Artificial Swarm Intelligence, a Human-in-the-Loop Approach to A.I. In Proceedings of the the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ, USA, 12–17 February 2016; pp. 4381–4382.
17. How, M.-L.; Hung, W.L.D. Educational Stakeholders' Independent Evaluation of an Artificial Intelligence-Enabled Adaptive Learning System Using Bayesian Network Predictive Simulations. *Educ. Sci.* **2019**, *9*, 110. [CrossRef]
18. Harley, J.M.; Lajoie, S.P.; Frasson, C.; Hall, N.C. Developing Emotion-Aware, Advanced Learning Technologies: A Taxonomy of Approaches and Features. *Int. J. Artif. Intell. Educ.* **2017**, *27*, 268–297. [CrossRef]
19. Forushani, N.Z.; Besharat, M.A. Relation between emotional intelligence and perceived stress among female students. *Procedia-Soc. Behav. Sci.* **2011**, *30*, 1109–1112. [CrossRef]
20. McGeown, S.P.; St Clair-Thompson, H.; Clough, P. The study of non-cognitive attributes in education: Proposing the mental toughness framework. *Educ. Rev.* **2016**, *68*, 96–113. [CrossRef]
21. Panerai, A.E. Cognitive and noncognitive stress. *Pharmacol. Res.* **1992**, *26*, 273–276. [CrossRef]

22. Pau, A.K.H. Emotional Intelligence and Perceived Stress in Dental Undergraduates. *J. Dent. Educ.* **2003**, *67*, 6.
23. Schoon, I. *The Impact of Non-Cognitive Skills on Outcomes for Young People*; Education Endowment Foundation: London, UK, 2013.
24. Manheim, D. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *BDCC* **2019**, *3*, 21. [CrossRef]
25. Perry, B.; Uuk, R. AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk. *BDCC* **2019**, *3*, 26. [CrossRef]
26. Turchin, A.; Denkenberger, D.; Green, B. Global Solutions vs. Local Solutions for the AI Safety Problem. *BDCC* **2019**, *3*, 16. [CrossRef]
27. Umbrello, S. Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *BDCC* **2019**, *3*, 5. [CrossRef]
28. Watson, E.N. The Supermoral Singularity—AI as a Fountain of Values. *BDCC* **2019**, *3*, 23. [CrossRef]
29. Ziesche, S.; Yampolskiy, R. Towards AI Welfare Science and Policies. *BDCC* **2018**, *3*, 2. [CrossRef]
30. Duettmann, A.; Afanasjeva, O.; Armstrong, S.; Braley, R.; Cussins, J.; Ding, J.; Eckersley, P.; Guan, M.; Vance, A.; Yampolskiy, R. *Artificial General Intelligence: Coordination & Great Powers*; Foresight Institute: Palo Alto, CA, USA, 2018.
31. Cheewaprakobkit, P. Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, China, 13–15 March 2013; p. 5.
32. Shahiri, A.M.; Husain, W.; Rashid, N.A. A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Comput. Sci.* **2015**, *72*, 414–422. [CrossRef]
33. Brusilovsky, P.; Karagiannidis, C.; Sampson, D. Layered evaluation of adaptive learning systems. *Int. J. Contin. Eng. Educ. Lifelong Learn.* **2004**, *14*, 402. [CrossRef]
34. Zeng, D. From Computational Thinking to AI Thinking. *IEEE Intell. Syst.* **2013**, *28*, 2–4.
35. How, M.-L.; Hung, W.L.D. Educing AI-Thinking in Science, Technology, Engineering, Arts, and Mathematics (STEAM) Education. *Educ. Sci.* **2019**, *9*, 184. [CrossRef]
36. Pezzulo, G.; Cartoni, E.; Rigoli, F.; Pio-Lopez, L.; Friston, K. Active inference, epistemic value, and vicarious trial and error. *Learn. Mem.* **2016**, *23*, 322–338. [CrossRef]
37. Al-Mutawah, M.A.; Fateel, M.J. Students' Achievement in Math and Science: How Grit and Attitudes Influence? *Int. Educ. Stud.* **2018**, *11*, 97. [CrossRef]
38. Chamberlin, S.A.; Moore, A.D.; Parks, K. Using confirmatory factor analysis to validate the Chamberlin affective instrument for mathematical problem solving with academically advanced students. *Br. J. Educ. Psychol.* **2017**, *87*, 422–437. [CrossRef]
39. Egalite, A.J.; Mills, J.N.; Greene, J.P. The softer side of learning: Measuring students' non-cognitive skills. *Improv. Sch.* **2016**, *19*, 27–40. [CrossRef]
40. Lipnevich, A.A.; MacCann, C.; Roberts, R.D. Assessing Non-Cognitive Constructs in Education: A Review of Traditional and Innovative Approaches. In *The Oxford Handbook of Child Psychological Assessment*; Saklofske, D.H., Reynolds, C.R., Schwean, V., Eds.; Oxford University Press: Oxford, UK, 2013.
41. Mantzicopoulos, P.; Patrick, H.; Strati, A.; Watson, J.S. Predicting Kindergarteners' Achievement and Motivation From Observational Measures of Teaching Effectiveness. *J. Exp. Educ.* **2018**, *86*, 214–232. [CrossRef]
42. Hox, J.; van de Schoot, R.; Matthijsse, S. How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* **2012**, *6*, 87–93.
43. Bayes, T. Letter from the Late Reverend Mr. Thomas Bayes, F.R.S. to John Canton, M.A. and F. R. S. In *The Royal Society, Philosophical Transactions (1683–1775)*; The Royal Society Publishing: London, UK, 1763; Volume 53, pp. 269–271.
44. van de Schoot, R.; Kaplan, D.; Denissen, J.; Asendorpf, J.B.; Neyer, F.J.; van Aken, M.A.G. A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Dev.* **2014**, *85*, 842–860. [CrossRef]
45. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2010; ISBN 978-0-521-89560-6.
46. Pearl, J. Causes of Effects and Effects of Causes. *Sociol. Methods Res.* **2015**, *44*, 149–164. [CrossRef]
47. Pearl, J. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* **1986**, *29*, 241–288. [CrossRef]

48. Lee, S.-Y.; Song, X.-Y. Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivar. Behav. Res.* **2004**, *39*, 653–686. [CrossRef]
49. Button, K.S.; Ioannidis, J.P.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.; Munafao, M.R. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **2013**, *14*, 365–376. [CrossRef]
50. Kaplan, D.; Depaoli, S. Bayesian structural equation modeling. In *Handbook of Structural Equation Modeling*; Hoyle, R., Ed.; Guilford Press: New York, NY, USA, 2012; pp. 650–673.
51. Walker, L.J.; Gustafson, P.; Frimer, J.A. The application of Bayesian analysis to issues in developmental research. *Int. J. Behav. Dev.* **2007**, *31*, 366–373. [CrossRef]
52. Zhang, Z.; Hamagami, F.; Wang, L.; Grimm, K.J.; Nesselroade, J.R. Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* **2007**, *31*, 374–383. [CrossRef]
53. Kaplan, D. Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assess. Educ.* **2016**, *4*, 7. [CrossRef]
54. Levy, R. Advances in Bayesian Modeling in Educational Research. *Educ. Psychol.* **2016**, *51*, 368–380. [CrossRef]
55. Mathys, C. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **2011**, *5*, 39. [CrossRef]
56. Muthén, B.; Asparouhov, T. Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychol. Methods* **2012**, *17*, 313–335. [CrossRef]
57. How, M.-L.; Hung, W.L.D. Harnessing Entropy via Predictive Analytics to Optimize Outcomes in the Pedagogical System: An Artificial Intelligence-Based Bayesian Networks Approach. *Educ. Sci.* **2019**, *9*, 158. [CrossRef]
58. Bekele, R.; McPherson, M. A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition. *Br. J. Educ. Technol.* **2011**, *42*, 395–416. [CrossRef]
59. Millán, E.; Agosta, J.M.; de la Cruz, J.L.P. Bayesian student modeling and the problem of parameter specification. *Br. J. Educ. Technol.* **2002**, *32*, 171–181. [CrossRef]
60. Shannon, C.E. The lattice theory of information. *IRE Prof. Group Inf. Theory* **1953**, *1*, 105–107. [CrossRef]
61. Correa, M.; Bielza, C.; Pamies-Teixeira, J. Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. *Expert Syst. Appl.* **2009**, *36*, 7270–7279. [CrossRef]
62. Cowell, R.G.; Dawid, A.P.; Lauritzen, S.L.; Spieglehalter, D.J. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*; Springer: New York, NY, USA, 1999; ISBN 978-0-387-98767-5.
63. Jensen, F.V. *An Introduction to Bayesian Networks*; Springer: New York, NY, USA, 1999; ISBN 0-387-91502-8.
64. Korb, K.B.; Nicholson, A.E. *Bayesian Artificial Intelligence*; Chapman & Hall/CRC: London, UK, 2010; ISBN 978-1-4398-1591-5.
65. Tsamardinos, I.; Aliferis, C.F.; Statnikov, A. Time and sample efficient discovery of Markov blankets and direct causal relations. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '03, Washington, DC, USA, 24–27 August 2003; p. 673.
66. Guoyi, C.; Hu, S.; Yang, Y.; Chen, T. Response surface methodology with prediction uncertainty: A multi-objective optimisation approach. *Chem. Eng. Res. Des.* **2012**, *90*, 1235–1244.
67. Fox, R.J.; Elgart, D.; Christopher Davis, S. Bayesian credible intervals for response surface optima. *J. Stat. Plan. Inference* **2009**, *139*, 2498–2501. [CrossRef]
68. Miró-Quesada, G.; Del Castillo, E.; Peterson, J.J. A Bayesian approach for multiple response surface optimization in the presence of noise variables. *J. Appl. Stat.* **2004**, *31*, 251–270. [CrossRef]
69. Myers, R.H.; Montgomery, D.C.; Anderson-Cook, C.M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed.; Wiley and Sons, Inc.: Somerset, NJ, USA, 2009; ISBN 978-0-470-17446-3.
70. Collins, J.A.; Greer, J.E.; Huang, S.H. *Adaptive Assessment Using Granularity Hierarchies and Bayesian Nets*; Springer: Berlin/Heidelberg, Germany, 1996; Volume 1086, pp. 569–577.
71. Conati, C.; Gertner, A.; VanLehn, K.; Druzdzel, M. On-line student modelling for coached problem solving using Bayesian networks. In Proceedings of the Sixth International Conference on User Model—UM'97, Sardinia, Italy, 2–5 June 1997; Springer: Berlin/Heidelberg, Germany, 1997; pp. 231–242.
72. Jameson, A. Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling User-Adapt. Interact.* **1996**, *5*, 193–251. [CrossRef]

73. VanLehn, K.; Niu, Z.; Siler, S.; Gertner, A.S. *Student Modeling from Conventional Test Data: A Bayesian Approach without Priors*; Springer: Berlin/Heidelberg, Germany, 1998; Volume 1452, pp. 434–443.
74. Conrady, S.; Jouffe, L. *Bayesian Networks & BayesiaLab: A Practical Introduction for Researchers*; Bayesia: Franklin, TN, USA, 2015; ISBN 0-9965333-0-3.
75. Bayesia, S.A.S. Bayesialab. Available online: https://www.bayesialab.com/ (accessed on 18 March 2019).
76. Bayesia, S.A.S. BayesiaLab: Missing Values Processing. Available online: http://www.bayesia.com/bayesialab-missing-values-processing (accessed on 2 June 2019).
77. Lauritzen, S.L.; Spiegelhalter, D.J. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc.* **1988**, *50*, 157–224. [CrossRef]
78. Kschischang, F.; Frey, B.; Loeliger, H. Factor graphs and the sum product algorithm. *IEEE Trans. Inf. Theory* **2001**. [CrossRef]