Validation Issues in Educational Data Mining – the case of HTML-Tutor and iHelp

Mihaela Cocea[1,2], Stephan Weibelzahl[1]
[1]National College of Ireland, School of Computing, Mayor Street, Dublin 1, Ireland
[2]London Knowledge Lab, Birkbeck College, University of London, 23-29 Emerald Street, London, WC1N 3QS, UK

INTRODUCTION

Validation is one of the key aspects in data mining and even more so in *educational* data mining (EDM) owing to the nature of the data. In this chapter, a brief overview of validation in the context of EDM is given and a case-study is presented. The field of the case study is related to motivational issues in general and disengagement detection in particular. There are several approaches to eliciting motivational knowledge from a learner's activity trace; in this chapter the validation of such an approach is presented and discussed.

The chapter is structured as follows. In the next section an overview of validation in the context of EDM is presented. Section 2 presents the case-study, including previous work on motivation in e-Learning, details of data and methods, and results. Section 3 presents some challenges encountered and lessons learned and, finally, Section 4 concludes the paper.

I. VALIDATION IN THE CONTEXT OF EDM

The term validation in educational data mining is used in two different meanings: (a) the validation of a model for the current context and similar users or (b) validation of a model in a

new context and/or for other users. The former is the typical evaluation of models in data mining, while the latter is more specific to social/educational research, when a model/theory is considered to be valid when it goes beyond the data on which the model has been built upon.

For the first type of validation, which is the most frequent one, different criteria are used, often depending on the methods applied. For example, when modeling student proficiency, criteria like relative closeness to real scores and mean absolute error [5], mean absolute deviation [1], R square and Bayesian Information Criterion [12] are used. For association rules many criteria are considered as representative; for example, twelve such measures were used in [22] among which are Chi-squared, correlation coefficient and predictive association. Prediction models often use accuracy, kappa, true positives and false positives measures [16] [26].

Validation against an external measurement, such as a standardized test, was proposed in [11]. Another possibility is to use different methods and compare their results. For example, [25] compared results of three methods: randomized controlled trials, learning decomposition and knowledge tracing; they argue that the qualitative consistency of results provides evidence for the validity of the results and of the methods.

In contrast, the validation of a model for a new context or for a new population is less frequently used due to the difficulty of building models that could work in different contexts and/or for different users. However, there is research that investigates these aspects; for example, validation of a model for "gaming the system" was successful for new lessons (i.e., different content) and new students (i.e., different users) [4]. Validation in a different context helps to understand to what degree findings can be generalized. It can thus contribute to educational theory building identifying relationships between concepts or patterns in behavior.

The case study presented in this chapter also investigates validation of a predictive approach in a different context, and more specifically, in a different e-Learning system. The development of the predictive approach and the validation are presented in the following.

II. DISENGAGEMENT DETECTION VALIDATION – A CASE STUDY

A. Detection of Motivational Aspects in e-Learning

Several approaches for motivation detection from learner's interactions with e-Learning systems have been proposed ranging from rule-based approaches to latent response models. Some of these approaches are briefly presented below.

First, a rule-based approach based on ARCS Model [14] has been developed to infer motivational states from learners' behavior using a ten-question quiz [9]. A set of 85 inference rules was produced by the participants who had access to replays of learners' interactions with the system and to learners' motivational traits.

Secondly, another approach [18] also based on ARCS Model is used to infer three aspects of motivation: confidence, confusion and effort, from the learner's focus of attention and inputs related to learners' actions.

Thirdly, engagement tracing [6] is an approach based on Item Response Theory that proposes the estimation of the probability of a correct response given a specific response time for modeling disengagement; two methods of generating responses are assumed: "blind guess" when the student is disengaged, and an answer with a certain probability of being correct when the

student is engaged. The model also takes into account individual differences in reading speed and level of knowledge.

Fourthly, a dynamic mixture model combining a hidden Markov model with Item Response Theory was proposed in [13]. The dynamic mixture model takes into account student proficiency, motivation, evidence of motivation, and the student's response to a problem. The motivation variable can have three values: a) motivated, b) unmotivated and exhausting all the hints in order to reach the final one that gives the correct answer, categorized as unmotivated-hint and c) unmotivated and quickly guessing answers to find the correct answer, categorized as unmotivated-guess.

Fifthly, a Bayesian Network has been developed [2] from log-data in order to infer variables related to learning and attitudes toward the tutor and the system. The log-data registered variables like problem-solving time, mistakes and help requests.

Last, a latent response model [4] was proposed for identifying the students that game the system. Using a pretest–posttest approach, the gaming behavior was classified in two categories: a) with no impact on learning and b) with decrease in learning gain. The variables used in the model were: student's actions and probabilistic information about the student's prior skills. The same problem of gaming behavior was addressed in [23], an approach that combines classroom observations with logged actions in order to detect gaming behavior manifested by guessing and checking or hint/ help abuse.

B. Proposed Approach to Disengagement Detection

In previous research [8] an approach to disengagement prediction for web-based systems that cover both reading and problem-solving activities was proposed. Log files from HTML-Tutor, a web based interactive learning environment, were analyzed. Initially, complete learning sessions, i.e., all activities between login and logout, were analysed [7]. However, it was found that in this set-up the level of engagement could be predicted only after 45 minutes of activity. After such a long duration, most disengaged students would have logged out, leaving no possibility of disengagement prediction and intervention. To overcome this problem, in the sub-sequent studies the sessions were divided in sequences of 10 minutes.

Several data mining techniques were used, showing that the user's level of engagement can be predicted from logged data, mainly related to reading pages and problem-solving activities. The fact that similar results were obtained when using different techniques and different numbers of attributes demonstrated the consistency of prediction and of the attributes used. The best accuracy, i.e. 88%, was obtained using Classification via Regression on a dataset including attributes related to reading, problem solving, hyperlinks and glossary. The best prediction for disengagement (with a true positive rate of 0.93), was obtained using Bayesian Networks.

C. Disengagement Detection Validation

1. Data Considerations

To validate the approach briefly presented above, data from iHelp, the University of Saskatchewan web-based system, was analyzed. The iHelp system includes two web-based

applications designed to support both learners and instructors throughout the learning process: the iHelp Discussion System and iHelp Learning Content Management System. The latter is designed to deliver online courses to students working at a distance, providing course content (text and multimedia) as well as quizzes and surveys. The students' interactions with the system are preserved in a machine readable format.

The same type of data about the interactions was selected from the logged information to perform the same type of analysis as the one performed on HTML-Tutor data. An HTML course was also chosen to prevent differences in results caused by differences in subject matter. Data from 11 students was used, meaning a total of 108 sessions and 450 sequences (341 of exactly 10 minutes and 109 less than 10 minutes). While at first glance a sample size of 11 students may seem rather small, it should be noted that the total time observed (i.e., more than 60 hours of learning) as well as the number of instances analyzed (i.e., 450 sequences) is far more important for the validity of the results.

Several attributes (displayed in Table 1) related to reading pages and quizzes were used in the analysis. The terms tests and quizzes will be used interchangeably; they refer to the same type of problem-solving activity, except that in HTML they are called tests and in iHelp they are named quizzes. Total time (of a sequence) was included as attribute for the trials that took into account sequences of less than 10 minutes as well as sequences of exactly 10 minutes. Compared to the analysis of HTML-Tutor logs, for iHelp there are fewer attributes related to quizzes: information about the number of questions attempted and about the time spent on them is included, but information about the correctness or incorrectness of answers given by users was not available at the time of the analysis. Two new meta-attributes that were not considered for HTML-Tutor were introduced for this analysis: the number of pages above and below a certain

time threshold, described in the subsequent section; they are meta-attributes because they are not among the raw data, but they are derived from it.

2. Annotation of the Level of Engagement

Annotations of the level of engagement for each sequence (of 10 minutes or less) were made by an expert with tutoring experience, in a similar manner as for the HTML-Tutor data; each sequence was annotated with the label *engaged* or *disengaged*. The expert annotated sequences based on *all logged attributes*, not just the ones used in the analyses. On top of these annotations, two additional rules related to the two new attributes (regarding number of pages that are above or below a threshold, depending on time spent reading) were used. These rules were applied after having obtained the expert annotations and as a result of a common pattern observed for both HTML-Tutor and iHelp. Consequently, the two new meta-attributes were added to investigate their contribution to prediction and their potential usage for a less time consuming process for annotation.

Initially, we intended to use the average time spent on each page across all users, as suggested by [19], but analyzing the data, we have seen that some pages are accessed by a very small number of users, sometimes only one; this problem was also encountered in other research (e.g. [10]). Consequently, we decided to use the average reading speed known to be in between 200 and 250 words per minute [20], [21]. Out of the 652 pages accessed by the students, 5 pages needed between 300 and 400 seconds to be read at average speed, 41 pages needed between 200 and 300 seconds, 145 needed between 100 and 300 seconds, and 291 needed less than 100 seconds. Some pages included images and videos; however, only two students attempted to

watch videos, one giving up after 3.47 seconds and the other one watching a video (or being on the page with the link to a video) for 162 seconds (almost three minutes). Taking into account this information, less than five seconds or more than 420 seconds (seven minutes) spent on a page were agreed to indicate disengagement.

For the HTML-Tutor logs, the level of engagement was established by human experts that looked at the log files and established the level of engagement for each sequence (of 10 minutes or less), in a similar way to the analysis described by [9]. The same procedure was applied for iHelp, plus the two rules aforementioned.

Accordingly, the level of engagement was determined for each sequence of 10 minutes or less. If in a sequence the learner spent more than seven minutes on a page or test, he/she was considered disengaged during that sequence. In relation to pages accessed less than five seconds, a user was considered disengaged if 2/3 of the total number of pages were below that time.

With HTML-Tutor, the rating consistency was verified by measuring inter-coding reliability. A sample of 100 sequences (from a total of 1015) was given to a second rater and results indicated high inter-coder reliability: percentage agreement of 92%, Cohen's kappa measurement of agreement of .826 ($p<.01$) and Krippendorff's alpha of .845 [15]. With iHelp only one rater classified the level of engagement for all sequences.

3. Analysis and Results

Using the attributes described in Section C.1, an analysis was conducted to investigate disengagement prediction with iHelp data and to compare the results with the ones from HTML-Tutor. Waikato Environment for Knowledge Analysis (WEKA) [24] was used to perform the

analysis. The same methods (presented below) as the ones used in our previous research were applied and four datasets were used: (i) Dataset 1 including all attributes and all sequences, (ii) Dataset 2 was obtained from Dataset 1 by eliminating the two additional attributes (NoPgP, NoPgM), (iii) Dataset 3 included all attributes, but only sequences of exactly 10 minutes and (iv) Dataset 4 was obtained from Dataset 3 by eliminating the two additional attributes (NoPgP, NoPgM). Dataset 2 and 4 were used to compare the results with the ones from HTML-Tutor. Table 2 presents the datasets with the corresponding attributes and sequences.

The eight methods [17] [24] used for the analysis are: (a) Bayesian Networks with K2 algorithm and maximum 3 parent nodes (BN); (b) Logistic regression (LR); (c) Simple logistic classification (SL); (d) Instance based classification with IBk algorithm (IBk); (e) Attribute Selected Classification using J48 classifier and Best First search (ASC); (f) Bagging using REP (reduced-error pruning) tree classifier (B); (g) Classification via Regression (CvR) and (h) Decision Trees (DT) with J48 classifier based on Quilan's C4.5 algorithm. The experiments were done using 10-fold stratified cross-validation iterated 10 times.

Results are displayed in Table 3, including accuracy and its standard deviation across all trials, true positive (TP) rate for disengaged class, precision (TP/ (TP + false positive)) for disengaged class, mean absolute error and kappa statistic. In our case, TP rate is more important than precision because TP rate indicates the correct percentage from *actual* instances of a class and precision indicates the correct percentage from *predicted* instances in that class. In other words, we want to identify as many disengaged students as possible. If an engaged students is misdiagnosed as being disengaged and receives special treatment for re-motivation, this will cause less harm than the opposite situation.

The results presented in Table 3 show very good levels of prediction for all methods, with accuracy varying between approximately 81% and 98%. There are similar results for the disengaged class, the true positive rate and the precision indicator for disengaged class varying between 75% and 98%. The mean absolute error varies between 0.02 and 0.25; the kappa statistic varies between 0.64 and 0.97, indicating that the results are much better than chance. In line with the results for HTML-Tutor, the fact that very similar results were obtained from different methods and trials demonstrates the consistency of the prediction and of the attributes used for prediction. The results for Dataset 1 and 3 are better than the ones from Dataset 2 and 4, suggesting that the two new meta-attributes bring significant information gain.

The highest accuracy was obtained using Instance based classification with IBk algorithm on Dataset 3: 98.59%; the confusion matrix for this method is presented in Table 4. For the disengaged TP rate, the same method performs best on the same dataset: 0.98.

Investigating further the information gain brought by the two meta-attributes, attribute ranking using information gain ranking filter as attribute evaluator was performed and the following ranking was found: NoPgP, AvgTimeP, NoPages, NoPgM, NoQuestions and AvgTimeQ. Hence, the meta-attributes seem to be more important than the attributes related to quizzes. The information gain contributed by NoPgP is also reflected in the decision tree graph displayed in Figure 1, where NoPgP has the highest information gain, being the root of the tree.

4. Cross-system Results Comparison

Comparing the results of iHelp to the ones of HTML-Tutor, an improvement for Datasets 1 and 3 and a small decrease for Datasets 2 and 4 are noticed. For ease of comprehension some of the

results from HTML-Tutor log-file analysis were included. These are only for the dataset with the attributes related to reading and tests and they are presented in Table 5.

The decrease for Dataset 2 and 4 may be due to the two missing attributes related to quizzes: number of correct and number of incorrect answers that were available for HTML-Tutor. The increase for Datasets 1 and 3 could be accounted by the contribution of the two new attributes.

The two missing attributes related to correctness or incorrectness of quiz responses may improve even more the prediction level. Looking at their role in prediction with HTML-Tutor, using three attribute evaluation methods with ranking as search method for attribute selection, these two attributes were found to be the last ones. Thus, according to chi-square and information gain ranking the most valuable attribute is average time spent on pages, followed by the number of pages, number of tests, average time spent on tests, number of correctly answered tests and number of incorrectly answered tests. OneR ranking differs only in the position of the last two attributes: number of incorrectly answered tests comes before number of correctly answered tests. The attribute ranking using information gain filter for iHelp attributes, shows similar positions for attributes related to reading and tests, meaning that attributes related to reading come before the ones related to tests. This suggests that the two missing attributes with iHelp are not essential, but if available they could improve the prediction level. Table 6 summarizes the similarities and differences between the findings from iHelp and HTML-Tutor.

Even with the mentioned differences, the fact that a good level of prediction was obtained from similar attributes on datasets from different systems using the same methods indicate that engagement prediction is possible using information related to reading pages and problem-solving activities, information logged by most e-Learning system. Therefore, our proposed

approach for engagement prediction is potentially system independent and could be generalized for any web-based system that includes both types of activities.

III. CHALLENGES AND LESSONS LEARNED

In defining our approach to disengagement detection, one of the major challenges encountered was the definition of disengagement in terms of the actions of learners when interacting with web-based learning environments. The type of web-based systems investigated, i.e. systems that provide both reading and problem-solving activities, presents an even bigger challenge. Most frequently, research on motivation focused exclusively on problem-solving activities, often characterized by a clearly defined structure which, to a certain degree, facilitates the assessment and modeling of motivational characteristics. To overcome this problem we used human experts that assessed the level of engagement of learners based on their actions and annotated the data; these annotations were subsequently used in building the prediction models. As observed in other research [3], without labeled data it is difficult to validate models.

Another challenge was the subject domain; most previous research was conducted in fields like mathematics or programming, which are more systematic and, therefore, more "controllable" than non-technical domains. In our approach, the domain was HTML, which is at the junction between technical and non-technical domains. Still, what seemed a disadvantage may prove to be beneficial, in the sense that the characteristics of this domain may allow an easier generalization across other domains, including non-technical ones; however, this requires further investigation. One important lesson learned from the case study presented is that a lack of domain structure does not necessarily mean that user activity is impossible to model;

nevertheless, the modeling process involves more exploration and is, perhaps, closer to typical data mining, which aims to discover information hidden in the data.

Another challenge was the validation process and its aim: to validate the approach and the attributes involved in the detection of disengagement, rather than the models initially built. The disadvantages involved in this course of action are two-fold: (a) the model(s) need to be built for every new system and (b) annotations are needed to do that. However, the big advantage is that knowledge about the relevant attributes is available and this offers the possibility of building disengagement detectors for web-based systems that include both reading and problem-solving activities. The other way to generalize would be to use models built for other systems and change them or provide them with adaptive mechanisms for the new environment; however, current research indicates that this is still a difficult task, while our proposed approach, although involving some effort, is feasible.

In relation to the above mentioned challenge, the lesson learned is that two stages are needed when aiming to develop an approach that could be extended beyond the data it was initially build on. The first step is an exploratory one, involving research about the relevant attributes and methods, while the second one involves the practical, implementation issues. For example, when developing an approach the use of several methods serves the purpose of inspecting the consistency of results, while in practice it is best to work with one method.

IV. CONCLUSIONS

In this chapter issues related to validation in educational data mining were presented and discussed in the context of a case-study about disengagement detection. The proposed approach

for disengagement detection is simple and needs information about actions related to reading and problem-solving activities, which are logged by most e-Learning systems. Because of these characteristics, we believe that this approach can be generalized to other systems, as illustrated in the validation study presented in this chapter. The similarity of results across different data mining methods is also an indicator of the consistency of our approach and of the attributes used.

REFERENCES

1. Anozie N. and Junker, B. W., Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In Beck, J., Aimeur, E. and T. Barnes (Eds). *EDM: Papers from the AAAI Workshop*. Menlo Park, CA: AAAI Press., 1-6. Technical Report WS-06-05, 2006.
2. Arroyo, I. and Woolf, B.P., Inferring learning and attitudes from a Bayesian Network of log file data. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 33–34, 2005.
3. Baker, R.S.J.d. and Carvalho, A.M.J.A.d., Labeling Student Behavior Faster and More Precisely with Text Replays. In *Proceedings of First International Conference on Educational Data Mining*, 38-47, 2008.
4. Baker, R.S.J.d., Corbett, A.T., Roll, I. and Koedinger, K.R., Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18(3), 287-314, 2008.
5. Beck, J. E. and Sison, J., Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education* 16, 129-143, 2006.

6. Beck, J., Engagement tracing: Using response times to model student disengagement. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 88–95, 2005.
7. Cocea, M. and Weibelzahl, S., Can Log Files Analysis Estimate Learners' Level of Motivation? In *Proceedings of ABIS Workshop, ABIS 2006 - 14th Workshop on Adaptivity and User Modeling in Interactive Systems*, 32–35, 2006.
8. Cocea, M. and Weibelzahl, S., Eliciting Motivation Knowledge from Log Files towards Motivation Diagnosis for Adaptive Systems. In *Proceedings of 11th International Conference on User Modeling*, 197-206, 2007.
9. De Vicente, A. and Pain, H., Informing the Detection of the Students' Motivational State: an empirical Study. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, 933–943, 2002.
10. Farzan, R. and Brusilovsky, P., Social navigation support in E-Learning: What are real footprints. In *Proceedings of IJCAI'05 Workshop on Intelligent Techniques for Web Personalization*, 49–56, 2005.
11. Feng, M., Beck J., Hefferman, N. and Koedinger, K., Can an Intelligent System Predict Math Proficiency as Well as a Standardized Test? In *Proceedings of First International Conference on Educational Data Mining*, 107-116, 2008.
12. Feng, M., Heffernan, N.T. and Koedinger, K., Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. In *Proceedings of the Fifteenth International World Wide Web Conference*, 307-316, 2006.

13. Johns, J. and Woolf, B., A Dynamic Mixture Model to Detect Student Motivation and Proficiency. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, 163-168, 2006.
14. Keller, J.M., Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, 10(3), 2–10, 2007.
15. Lombard, M., Snyder-Duch, J. and Campanella Bracken, C., Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research, 2003. http://www.temple.edu/mmc/reliability (accessed November 6, 2006)
16. Mavrikis, M., Data-driven modelling of students' interactions in an ILE. In *Proceedings of First International Conference on Educational Data Mining*, 87-96, 2008.
17. Mitchell, T.M., *Machine Learning*. McGraw-Hill, New York, 1997.
18. Qu, L., Wang, N. and Johnson, W.L., Detecting the Learner's Motivational States in an Interactive Learning Environment. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 547–554, 2005.
19. Rafter, R. and Smyth, B., Passive Profiling from Server Logs in an Online Recruitment Environment. In *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization*, 35-41, 2001.
20. ReadingSoft.com found at HYPERLINK, http://www.readingsoft.com
21. TurboRead Speed Reading found at HYPERLINK, http://www.turboread.com
22. Ventura, S., Romero, C., and Hervas, C., Analysing Rule Evaluation Measures with Educational Datasets: A Framework to Help the Teacher. In *Proceedings of First International Conference on Educational Data Mining*, 177-186, 2008.

23. Walonoski, J. and Heffernan, N.T., Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In *Proceedings of the Eight International Conference in Intelligent Tutoring Systems*, 382–391, 2006.
24. Witten, I.H. and Frank, E., *Data mining. Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kauffman Publishers, Elsevier, Amsterdam, 2005.
25. Zhang, X., Mostow, J. and Beck, J.E., A Case Study Empirical Comparison of Three Methods to Evaluate Tutorial Behaviors. In *Proceedings of the 9th International Conference on Intelligent Tutoring System*, LNCS vol. 5091, 122-131, 2008.
26. Zhang, X., Mostow, J. Duke, N., Trotochaud, C., Valeri, J. and Corbett, A., Mining Free-form Spoken Responses to Tutor Prompts. In *Proceedings of First International Conference on Educational Data Mining*, 234-241, 2008.

Figure Captions

Figure 1. Decision Tree graph for Dataset 3.

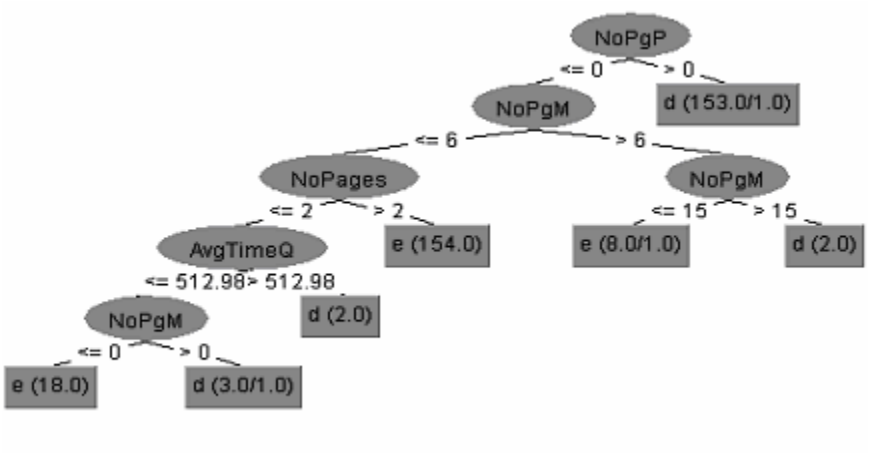

Table 1. The attributes used for analysis

Codes	Attributes
NoPages	Number of pages read
AvgTimeP	Average time spent reading
NoQuestions	Number of questions from quizzes/ surveys
AvgTimeQ	Average time spent on quizzes/surveys
Total time	Total time of a sequence
NoPpP	Number of pages above the threshold established for maximum time required to read a page
NoPM	Number of pages below the threshold established for minimum time to read a page

Table 2. Datasets used in the experiment

Dataset	Sequences	Attributes
Dataset 1	All sequences	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Total time, NoPpP, NoPM
Dataset 2	All sequences	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Total time
Dataset 3	Only 10 minutes sequences	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Total time, NoPpP, NoPM
Dataset 4	Only 10 minutes sequences	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Total time

Table 3. Experiment results summary

Dataset	Measure	BN	LR	SL	IBk	ASC	B	CvR	DT
Dataset 1	**Accuracy**	89.31	95.22	95.13	95.29	95.44	95.22	95.44	95.31
	Std. Dev	4.93	2.78	2.82	2.98	2.97	3.12	3.00	3.03
	TP rate	0.90	0.95	0.95	0.94	0.94	0.94	0.95	0.95
	Precision	0.90	0.95	0.95	0.96	0.97	0.97	0.96	0.96
	Error	0.13	0.07	0.10	0.05	0.08	0.08	0.08	0.07
	Kappa	0.79	0.90	0.90	0.91	0.91	0.90	0.91	0.91
Dataset 2	**Accuracy**	81.73	83.82	83.58	84.00	84.38	85.11	85.33	84.38
	Std. Dev	5.66	5.03	5.12	4.85	5.08	5.17	5.13	5.07
	TP rate	0.78	0.82	0.81	0.79	0.77	0.79	0.80	0.78
	Precision	0.86	0.86	0.86	0.89	0.91	0.91	0.91	0.91
	Error	0.22	0.24	0.26	0.20	0.25	0.23	0.23	0.25
	Kappa	0.64	0.68	0.67	0.68	0.69	0.70	0.71	0.69
Dataset 3	**Accuracy**	94.65	98.06	97.91	98.59	97.65	97.65	97.76	97.47
	Std. Dev	4.47	2.18	2.69	2.11	2.64	2.64	2.65	2.58
	TP rate	0.95	0.97	0.96	0.98	0.96	0.96	0.96	0.96
	Precision	0.94	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Error	0.07	0.02	0.04	0.02	0.05	0.04	0.03	0.03
	Kappa	0.89	0.96	0.96	0.97	0.95	0.95	095	0.95
Dataset4	**Accuracy**	84.29	85.82	85.47	84.91	84.97	85.38	85.26	85.24
	Std. Dev.	5.77	5.90	5.88	5.95	5.61	5.80	5.96	5.91
	TP rate	0.78	0.77	0.76	0.77	0.75	0.76	0.75	0.75
	Precision	0.88	0.92	0.92	0.89	0.92	0.92	0.92	0.92
	Error	0.18	0.22	0.23	0.20	0.25	0.23	0.24	0.24
	Kappa	0.68	0.71	0.70	0.69	0.69	0.70	0.70	0.70

Table 4. The confusion matrix for instance based classification with IBk algorithm

		Predicted	
		Engaged	**Disengaged**
Actual	**Engaged**	180	1
	Disengaged	4	155

Table 5. Experiment results summary for HTML Tutor

	BN	LR	SL	IBk	ASC	B	CvR	DT
Accuracy	87.07	86.52	87.33	85.62	87.24	87.41	87.64	86.58
TP rate	0.93	0.93	0.93	0.92	0.93	0.93	0.92	0.93
Precision	0.91	0.90	0.90	0.91	0.92	0.92	0.92	0.91
Error	0.10	0.12	0.12	0.10	0.10	0.12	0.12	0.11

Table 6. Similarities and dissimilarities between iHelp and HTML-Tutor

Characteristic	**iHelp**	**HTML-Tutor**
Prediction based on reading and tests attributes	81% to 85% with no information on correctness /incorrectness of quizzes and no additional attributes 85% to 98% with the two additional attributes	86-87%
Attribute ranking	Number of pages above a threshold Average time spent reading Number of pages read/ accessed Number of pages below a threshold Number of questions from quizzes Average time spent on quizzes	Average time spent on pages Number of pages Number of tests Average time spent on tests Number of correctly answered tests Number of incorrectly answered tests