Research and Practice in Technology Enhanced Learning

RESEARCH Open Access

The performance of some machine learning approaches and a rich context model in student answer prediction

Alisa Lincke[1*], Marc Jansen[1,2], Marcelo Milrad[1] and Elias Berge[3]

*Correspondence:
alisa.lincke@lnu.se
[1] Department of Computer Science and Media Technology, Linnaeus University, PG Vejdes väg, 351 95 Växjö, Sweden
Full list of author information is available at the end of the article

Abstract

Web-based learning systems with adaptive capabilities to personalize content are becoming nowadays a trend in order to offer interactive learning materials to cope with a wide diversity of students attending online education. Learners' interaction and study practice (quizzing, reading, exams) can be analyzed in order to get some insights into the student's learning style, study schedule, knowledge, and performance. Quizzing might be used to help to create individualized/personalized spaced repetition algorithm in order to improve long-term retention of knowledge and provide efficient learning in online learning platforms. Current spaced repetition algorithms have pre-defined repetition rules and parameters that might not be a good fit for students' different learning styles in online platforms. This study uses different machine learning models and a rich context model to analyze quizzing and reading records from e-learning platform called Hypocampus in order to get some insights into the relevant features to predict learning outcome (quiz answers). By knowing the answer correctness, a learning system might be able to recommend personalized repetitive schedule for questions with maximizing long-term memory retention. Study results show that question difficulty level and incorrectly answered previous questions are useful features to predict the correctness of student's answer. The gradient-boosted tree and XGBoost models are best in predicting the correctness of the student's answer before answering a quiz. Additionally, some non-linear relationship was found between the reading learning material behavior in the platform and quiz performance that brings added value to the accuracy for all used models.

Keywords: Adaptive quiz, Quiz performance, Machine learning, Rich context model, Answer probability prediction

Introduction

One of the biggest challenges for educators is to meet the individual needs of students while facing the constraints of time. One way to personalize education is by using adaptable learning systems (Papoušek and Pelánek 2015). In order to efficiently provide students with personalized, adaptive digital content, and scheduled practice, it is crucial that the learning system gets over time an understanding not only of the students' current knowledge level but also of his/her progression and self-regulated learning strategy.

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

One traditional way of assessing the knowledge level is letting students take a placement test (Hodara et al. 2012). However, to make a placement test adaptive, the system needs to be able to draw conclusions from every answered question. According to one study (Chounta et al. 2017), predicted probabilities that students will answer questions correctly can provide some insights into students' knowledge. By using the answers to predict the probability of answering correct on other questions, a learning system might be able to schedule the repetition frequency of the question or recommend questions with a suitable level of difficulty. This would make the placement test more efficient, i.e., needing fewer questions to get an accurate picture of the students' knowledge level.

Predicting the probabilities of students' answering correct may also be valuable in order to maximize students' engagement (Joseph 2005). If we know the probabilities of students answering questions correctly, then we may optimize the studies with regard to engagement, memory retention, and knowledge level. Previous studies have suggested that an adaptive fail rate in a quiz increases student engagement (Papoušek and Pelánek 2015; Ross et al. 2018). By choosing questions with a difficulty level that increases the chances of a student answering correct, around 60% of the questions seems to hit a sweet spot where the average student experiences the quiz challenging without being too difficult. Moreover, one study (Simon-Campbell et al. 2016) showed that the use of adaptive quizzes in nursing education increased learning material mastery, helped to predict final course grades, and positively influenced on the retation rates. Another study (House et al. 2016) explored the student's satisfaction with adaptive quizzing and showed that this learning strategy increased their knowledge in course content and helped better to prepare for final exams.

Modeling the student's knowledge and performance in online education systems is a well-established problem (Pardos and Heffernan 2011; Piech et al. 2015; Pelánek 2017; Duong et al. 2013). For instance, unknown students' knowledge background (academic performance, grades), personal information (age, gender), cognitive skills, and access to the learning resources such as reading material, quizzes, exams, and courses any time and order brings different learning behaviors and strategies (e.g., accessing the learning material in different orders, some students just doing quizzes and exams without reading the material on the web platform, others first read the material and afterwards doing quizzes to check their knowledge about this material). Furthermore, students can use other learning resources besides the ones provided/offered by the learning platform (such as books, university course content).

One approach for measuring academic achievements and student's knowledge is the Item Response Theory (IRT) models (Reise and Revicki DA 2014; Chen et al. 2005). IRT models allow measuring different students' abilities (intelligence, individual learning ability, attitude, academic achievements) by using answers on questions as test-based assessment. It predicts the probability that a student will answer the question correctly as a function with two parameters: student's knowledge level and the question difficulty (Chaudhry et al. 2018; Galvez et al. 2009). This modeling approach showed good practical use in estimating students' performance and making adaptive quizzes (dynamically decide which question to show based on student's answers). However, this approach does not model the evolution of students' knowledge over time (Chaudhry et al. 2018; Khajah et al. 2014).

In a comprehensive review (Dunlosky et al. 2013), researchers compared ten different learning techniques and claimed that retrieval practice and distributed practice were the only two techniques with robust effects on learning that generalized widely. The retrieval practice is a learning strategy that focuses on the active recalling of information from memory which has a positive effect on future recall attempts (Roediger III and Butler 2011). What is seen in studies is that the effect of test-enhanced learning is greatest if repeating the tests with increasing intervals (Roediger III and Karpicke 2006). Repeated re-study is another common study strategy where repetition is used to measure the student's activeness with the study materials (Karpicke and Roediger 2008; Thiede and Dunlosky 1999).

Most research on retrieval practice has been carried out in supporting learning in classroom settings rather than in MOOC environment. Therefore, how to effectively support retrieval practice in online learning environments has not yet been thoroughly examined (Davis et al. 2016). Students generate a vast amount of interactional data in MOOCs that allows to use data mining and machine learning techniques new insights on the student's learning strategies and improve retrieval process (Maldonado-Mahauad et al. 2018; Choffin et al. 2020; Tabibian et al. 2019; Settles and Meeder 2016; Davis et al. 2016). Quizzing and video-based learning material are one of the sources of retrieval practice in online learning platforms (Van der Zee et al. 2018; Fellman et al. 2020). Recent study (Davis et al. 2018) has developed an approach for creating adaptive quizzes in MOOC with automatically and intelligently delivering quiz questions by analyzing the student's learning behavior in previous courses. Their study results did not show a positive effect on the knowledge and learning; thus, the benefits of using the retrieval practice in online learning environment need further research as well as creation of new retrieval practice algorithms for online learning settings.

This study is an extension and follow-up on our previous work (Lincke et al. 2019) that aims to (a) explore the association between online learning activities (such as time spend on reading learning material in the system) and quiz performance in the system and (b) to predict the quiz performance by using different machine learning techniques on data provided by online learning platform called Hypocampus. By knowing in advance if the student will answer the question correctly, it will allow us to create more personalized retrieval practice algorithm rather than the use of standard Leitner system for scheduling the question repetition in a quiz. Specifically, in this study, we increased the dataset (by including both multiple-choice and text type questions), added more user study behavior features (like reading time for learning materials), and added Bayesian-based machine learning model in order to compare the Bayesian model-based approach with previously applied models. We assess the accuracy and response time of training and predicting the quiz performance for the following models: linear regression, logistic regression, gradient-boosted tree, extreme gradient-boosted tree (XGBoost), deep neural network, rich context model (RCM) (Sotsenko 2017), and Bayesian neural network.

Method

In this study, a general machine learning (ML) pipeline (Pentreath 2015) was used for predicting the probability whether a question will be answered correctly as shown on Fig. 1. There are four main steps: data preprocessing of collected data (Dataset); feature extraction and selection; applying seven models: linear regression, logistic regression, gradient-boosted tree regression, extreme gradient-boosted tree (XGBoost) (Chen and Guestrin 2016), feed-forward deep neural network, a rich context model (RCM) (Sotsenko 2017), and Bayesian neural network (BNN); and the last step is the model evaluation.

The model evaluation included the accuracy metrics and performance measures. The following metrics were extracted from confusion matrix (Ting 2010) and used for accuracy evaluation: false-positive rate (FP, %), false-negative rate (FN, %), precision, recall, accuracy, F1-score (F1, %), and Pearson correlation coefficient (*r*) between the predicted value by the model and depended variable (answer on the question). In our answer, prediction task: false positives are incorrect answered questions which have been predicted as correct; false negatives are correctly answered questions which were predicted as incorrectly answered questions; precision is a proportion of correct answers predicted; recall is a proportion of correctly answered questions which are predicted to be correctly answered; accuracy is a proposition of total number of answer predictions that were correct; F1-score is a weighted harmonic average of precision and recall; Pearson correlation coefficient shows how well the true value correlated with the predicted value, where 0 is not correlated and 1 is highly correlated. We also use receiver operator characteristic (ROC) and precision-recall (PR) curves (Davis and Goadrich 2006) to compare the performance of machine leaning models.

For performance evaluation, we measured execution time in milliseconds for training and testing the model. Performance and accuracy evaluation experiments are run on a MacBook Pro with a 2.3 GHz Intel Core i7 processor.

The correlation analysis was applied to check the linear association between online learning activities (described as features in Feature Extraction and Selection section) and quiz performance (answer on questions).

Dataset

For this study, we have obtained the data from the Hypocampus web-based learning platform. Hypocampus is an adaptive web-based learning platform used by medical students in Sweden. It contains a library with many interactive reading materials (e.g.,

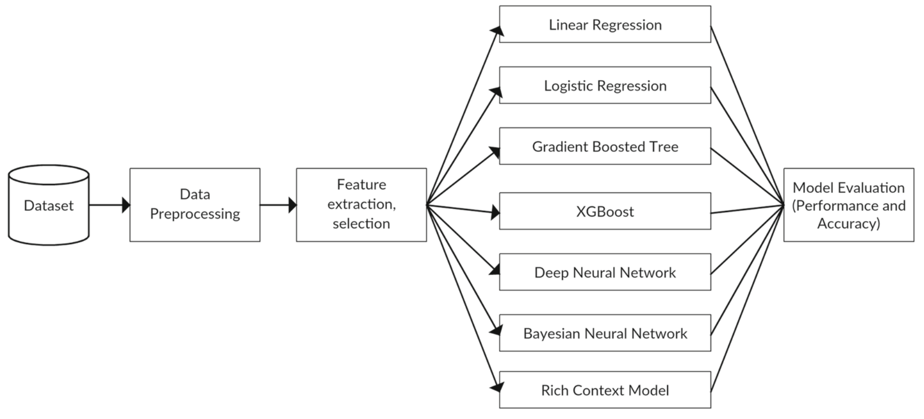

Fig. 1 ML pipeline for predicting the probability if a question will be answered correctly

course literature) that students can use for self-studies in order to learn about a particular subject matter and to revise and review their current knowledge. The platform provides also quizzes for each reading material in order to help students to check and assess their current knowledge for each particular subject matter. In addition, it offers customized learning paths based on quantitative educational studies, visualizations of learning progress for students and teachers, and adaptive individual learning pathways. The learning platform optimizes the learning content according to the principles of retrieval practice (Karpicke and Roediger 2008).

The reading material is structured into various subjects (e.g., Dermatology, Surgery, Gynecology, Internal Medicine, and others). Every subject has a number of topics called chaptergroups. Every chaptergroup represents a learning material and consists of chapters. Every chapter has a quiz containing from 4 to 15 questions; some of them can have up to 28 questions. There are two types of questions: *multiple choice* and *text questions*. The multiple-choice questions have several options to choose and only one is correct. After answering the multiple-choice questions, the system will show the direct feedback to students whether the answer is correct or incorrect. Moreover, depending on the answers of the questions, the system highlights a part of the reading material in green color (that means a student knows this part of the material) or red color (that means a student does not know this part of the material and should read it). The text question contains a problem description and a student should provide a text answer on the problem. After answering a text question, the system does not check the text answer provided by student, but rather shows the answer and explanations to the problem. Once the student has seen the answer, she/he should correct herself by selecting "I knew the answer" (correct) or "I need to read more" (incorrect). Table 1 describes a summary of the collected data of randomly selected 300 medical students over a period of 10 months between 2017 and 2018.

Table 1 shows that selected 300 students tend to read more than doing quizzes in the platform (38% quiz sessions and 62% reading sessions). However, they invest more time on quizzing (1127 h) then reading (467 h) in the system. In average, students spend around 2 min to read a single learning material (chapter) and 41 s to answer a question.

The collected dataset includes information about user identification number, question type (multiple choice or text), question identifier, chapter id to which the question id belongs, time answering a question, time reviewing the feedback from the system after answering the question, student's answer (true — correct and false — incorrect), student's text answer on the text questions, timestamp, course identifier, question session, chapter group number, chapter identification number, focused reading time, and others. After

Table 1 Summary of collected data

Name	Number of records
Number of students	300
Number of questions	121,423
Multiple-choice questions	18,092
Text questions	103,331
Correct answers	94,433
Incorrect answers	26,725
Quiz sessions	6081

collecting all these data, the preprocessing and feature extraction steps are performed in order to prepare the dataset to be used by machine learning techniques and RCM.

Data preprocessing

As part of the data preprocessing step, we performed data cleaning by removing records that have missing values related to the question id information and chapter id information. Original dataset stored into three tables: *answer_material*, *_material*, and *_time*. First, we joined *answer_material* and *review_material* tables into *answer_review_material* table in order to have full information about quizzing in a single table for further feature extraction process. The read_time table contains information about interactions with learning material (chaptergroup, chapter, or quiz). In order to access the quiz in Hypocampus system, a student needs first to select chapter group and scroll over all chapters to the bottom of the page in order to access a quiz. That means that *read_time* table contains records that may correspond to the user navigation to the quiz and not to the reading of the learning material. Moreover, the user session in *read_time* table may include a sequence of different activities (e.g., reading, scrolling, reading, quizzing, reading, quizzing, scrolling). Therefore, the decision was taken to derive the user sessions that relate only to reading activities in order to identify how much time spend on reading a single material (chapter). In this study, a reading session defined and calculated as inactivity with the system for more than 15 min, and the duration of the reading session should be more than 5 s (otherwise, it is scrolling over learning material). The chosen numbers for inactivity and scrolling were selected as starting point and could be changed if necessarily. After reading session calculation, the records in two tables (*read_time* and *answer_review_material*) should be synchronized according to timestamp. This was the last operation in data preprocessing step that returns prepared dataset for feature extraction process.

Feature extraction

The feature extraction step was performed on the preprocessed data presented in Table 1. In this study, a *feature* represents an interaction with the system (quizzing and reading) at specific point in time. A feature is called *direct* if it is directly used from raw data (e.g., user id, chapter id) or *derived* if it is computed from direct or indirect features (e.g., number of correct answers, time spend on reading in the system, etc.). We have selected 4 direct features from the original dataset and used it as labels: *user id, chapter id, question id*, and *question session*. After performing several interactions in the ML pipeline (Fig. 1), the following 15 derived features were identified and calculated for each question record (see in Table 2).

The extracted features capture different learning activities in Hypocamus system and relevant to the predictor variable (student's answer). For example, learning activities such as reading learning materials (F3,F4,F12), quizzing (quiz accuracy (F5-F10,F15), question difficulty level (F13), quiz frequency (F11,F14)), and one categorical variable F1 is used to describe one of the dimensions of current context of the student (time).

Models

We selected seven models: two simple models (linear and logistic regression) were selected, and five more advanced models (two decision trees models, deep neural

Table 2 Feature overview

N	Feature type	Name	Description
F1	Categorical	Time of the day	Morning, lunch, afternoon, evening, night
F2	Numerical	Time since last doing quiz	Represents the time duration (in seconds) since last time the student was doing a quiz
F3	Numerical	Tslr	Represents the time duration (in seconds) since last time the student was reading a chapter
F4	Numerical	Reading time	Total reading time of learning material (chapter)
F5	Numerical	Correct per chapter	Number of questions answered correctly per chapter in current quiz session
F6	Numerical	Correct per attempt	Number of questions answered correctly per attempt
F7	Numerical	Correct per session	Number of questions answered correctly per session
F8	Numerical	Incorrect per chapter	Number of questions answered incorrectly per chapter
F9	Numerical	Incorrect per attempt	Number of questions answered incorrectly per attempt
F10	Numerical	Incorrect per session	Number of questions answered incorrectly per session
F11	Numerical	Attempt number	Number of times a question was answered by the student
F12	Numerical	Reading sessions	Number of times a student read the learning material (chapter)
F13	Numerical	Question facility index	Represents the question difficulty and in range between 0 and 1 (where 0 is very difficult and 1 is very easy)
F14	Numerical	Question counter	Question number in the quiz session
F15	Numerical	Correct total	Total number of questions answered correctly for a specific chapter

network, Bayesian neural network) because they are commonly used in regression problems (predicting probabilities) and taking contextual information into account (RCM). Furthermore, some of these models were successfully used in predicting students' performance (Bucos 2018; Shahiri et al. 2015; Ibrahim and Rusli 2007).

Linear and logistic regressions are one of the simplest machine learning models used to predict one dependent variable based on the set of independent variables (Seber and Lee 2012). Linear regression assumes that there is a linear relationship between dependent and independent variables. In our scenario, the dependent variable is the answer to a question (correct/incorrect) and independent variables are features that described user study behavior data (see Table 2). We use this model to check whether there is a linear relationship between the learning activities (quizzing, reading) and students' answers on the questions. Logistic regression is applied when the dependent variable is binary. In our prediction problem, linear and logistic regressions predict the probability from 0 to 1 if the question will be answered correctly.

More advanced models such as decision trees (gradient-boosted tree and improved version extreme gradient-boosted tree (XGBoost)) help to reduce factors such as bias, variance, and dealing with unbalanced data (Cieslak and Chawla 2008; Chawla et al. 2004). To the best of our knowledge, in the reviewed literature, this model was not applied to student's performance or knowledge prediction. Therefore, we decide to test this model and to apply it in our study.

Deep neural networks become more popular in educational data mining (EDM) tasks (Coelho and Silveira 2017; Guo et al. 2015). We use a feed-forward deep neural network. After empirically testing different model parameters in our previous study (Lincke et al. 2019), the following parameters were found having minimal error: input layer with 13 neurons and activation function *"relu"*, one hidden layer with 13 neurons and activation function *"relu"*, output layer with 1 neuron and activation function *"sigmoid"*, optimizer *"adam"* and loss *"binary_crossentropy"*.

Bayesian neural network combines probabilistic modeling and neural network in order to benefit from the bounds and guarantees of probabilistic modeling (Mullachery et al. 2018). In BNN, weights are a probability distribution instead of a single value, which describe the uncertainty in weights and in predictions. The output of the BNN is an entire distribution of answers. BNN can be used for classification and regression problems. In our study, we use BNN for predicting the probability that an answer will be correct (regression problem). Therefore, perceptron is logistic regression with 1 hidden layer and with 15 neurons. Weights initialized with normal distribution, tanh activation function used for hidden layer and sigmoid for output layer with Bernoulli likelihood function. We approximate the posterior using variational inference method called ADVI provided by PyMC3 library (Kucukelbir et al. 2017).

Rich context model models contextual information in a multidimensional vector space model (MVSM) in order to provide recommendations based on the current context of a user. It is important to understand the student's current context (e.g., time of the day: morning, lunch, afternoon, evening, night; location; number of difficult questions answered correctly) in order to provide personalized learning tasks/quizzes. This model can handle different data types (numerical, categorical features) in predicting the answer correctness. RCM requires an example set of vectors that represent the basis (or training set used in machine learning approach) and a current context set of vectors to obtain recommended result (or testing set used in machine learning approach). In this study, the quiz records are divided into examples and current context datasets with 7:3 ratios. Results from our previous study (Lincke et al. 2019) showed that RCM require to have balanced example dataset. Therefore, the example dataset contains 30% of data with similar distribution of incorrect and correct answers (where incorrect answer represents class 0 and correct is class 1). The example dataset is transformed to one-dimensional vector representing the quiz record (QR) and placed in MVSM (e.g., from 1 to N as shown on Fig. 2).

The current context datasets are transformed into a one-dimensional vector (as current quiz record (CQR) shown in Fig. 2), and Euclidean distance (d) is used to find the most similar quiz record in the example dataset (QR1, QR2, …QRN). The most similar quiz record that has minimal distance defines if the student will answer correct or incorrect. In more detail, we calculate distances to each example in MVSM. The vector distance is a two-dimensional array where the first dimension is distance and the second dimension is the answer (correct 1 or incorrect 0). In this study, distances are a two-dimensional array sorted by distance in ascending order, because we used only one distance measure (Euclidean) for all dimensions. The first element in the array will have minimal distance to the current quiz record, and its answer value defines whether the student will answer correct or incorrectly.

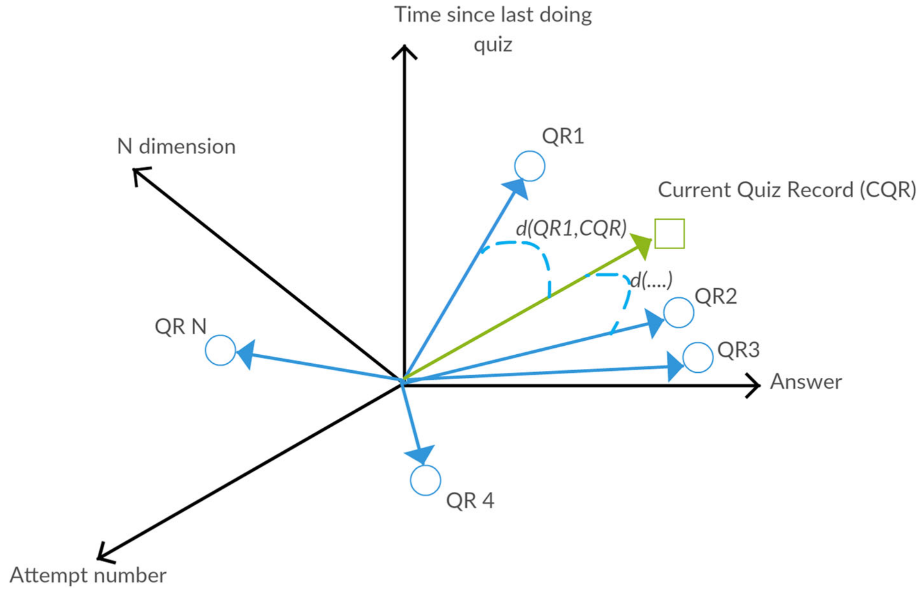

Fig. 2 Representation of quiz records in RCM

In terms of technologies and tools, we have used Matlab version R2020a (`?matlab`) for correlation analysis and Apache Spark MLlib and Scikit-learn libraries for machine learning pipeline (performing data preprocessing, feature extraction, and transformation steps). For building the models, we used different frameworks and libraries: Apache Spark MLlib (Meng et al. 2016) for linear, logistic, and gradient-boosted trees; XGBoost python library for extreme gradient-boosted tree; the Keras deep learning library for deep neural network; the PyMC3 library was used for building Bayesian neural network; and our Contextualization Service (Sotsenko et al. 2016b) for building RCM.

Results

Correlation analysis

The results of the correlational analysis between features and response variable (question answer) are presented in Table 3. There is a low significant correlation between question difficulties and quiz performance ($r = 0.26$, p-value $= 0.000$). There is a negative low correlation (significant) between incorrectly answered questions per chapter ($r = -0.21$, $p = 0.000$), per session ($r = -0.23$, p-value $= 0.000$), and quiz performance. No significant linear correlation between reading time spend on learning materials and quiz performance was found. Overall, most of the features are weakly correlating with response variable (student's answer).

Prediction of quiz performance

First, we run the experiment for dataset of 121,423 quiz records (multiple-choice and text question types) without reading features (F3, F4, and F12). The evaluation is conducted using the 10-fold cross-validation approach (Kohavi and et al 1995), the train-validation split approach provided by the Spark Mllib library for hyper-parameter tuning (Gounaris and Torres 2018), and one split into training and validation datasets for the Bayesian model. The results of the first experiment are shown in Table 4.

Table 3 Correlation coefficients of students' online learning activities and quiz performance

N	Feature name	Quiz performance (answer), *r* (*p*-value)
F1	Time of the day	0.0087 (0.002)
F2	Time since last doing quiz	− 0.0155 (0.000)
F3	Tslr	0.0089 (0.002)
F4	Reading time	− 0.0015 (0.594)
F5	Correct per chapter	0.1194 (0.000)
F6	Correct per attempt	0.118 (0.000)
F7	Correct per session	0.1161 (0.000)
F8	Incorrect per chapter	− 0.2087 (0.000)
F9	Incorrect per attempt	− 0.1256 (0.000)
F10	Incorrect per session	− 0.2264 (0.000)
F11	Attempt number	0.0639 (0.000)
F12	Reading sessions	0.0216 (0.000)
F13	Question facility index	0.2597 (0.000)
F14	Question counter	0.0931 (0.000)
F15	Correct total	0.1264 (0.000)

As shown in Table 4, all of the models performed well in predicting the probability that an answer will be correct, with and accuracy rate ranging between 76 and 88%. These results indicate that the dataset of features has some non-linear association with response variable (student's answer). The RCM model got the lowest accuracy (76%) and more false-negative errors (15%) than machine learning approaches. This could be explained by the lack of using contextual information in the model such as student's location, age, gender, a device type, and others. Bayesian neural network (with logistic regression perceptron) outperformed only on 1% in accuracy in comparison to traditional logistic regression model. The best algorithms in predicting the probability that a student will answer correctly are gradient-boosted tree and XGBoost with around 88% accuracy and highest correlation (0.62–0.63) and low false-positive and false-negative errors (FP = 39–40%, FN = 4–5%).

Receiving operation characteristic (ROC) is an established metric for comparing the performances of classifiers (Bradley 1997; Fawcett 2006). ROC curves summarize the trade-off between true-positive rate and false-positive rate using different probability thresholds. In our study, ROC curve is for two classes (correct/incorrect answer) and based on plotting true-positive (TP) rate on the y-axis and the false-positive (FP) rate on the x-axis. Figure 3 shows ROC curves for all machine learning models. The current implementation of RCM does not provide the probabilities; therefore, it was excluded from the ROC analysis.

Table 4 Evaluation results without reading features

Model	FP, %	FN, %	Precision, %	Recall, %	Accuracy, %	F1, %	*r*
Linear regression	80	3	81	97	80	89	0.29
Logistic regression	81	3	81	97	79	88	0.25
Gradient-boosted tree	39	5	90	95	88	92	0.62
XGBoost	40	4	90	96	88	93	0.63
Deep neural network	54	2	86	98	86	92	0.55
Bayesian NN	87	2	80	98	80	88	0.23
RCM	53	15	85	85	76	85	0.31

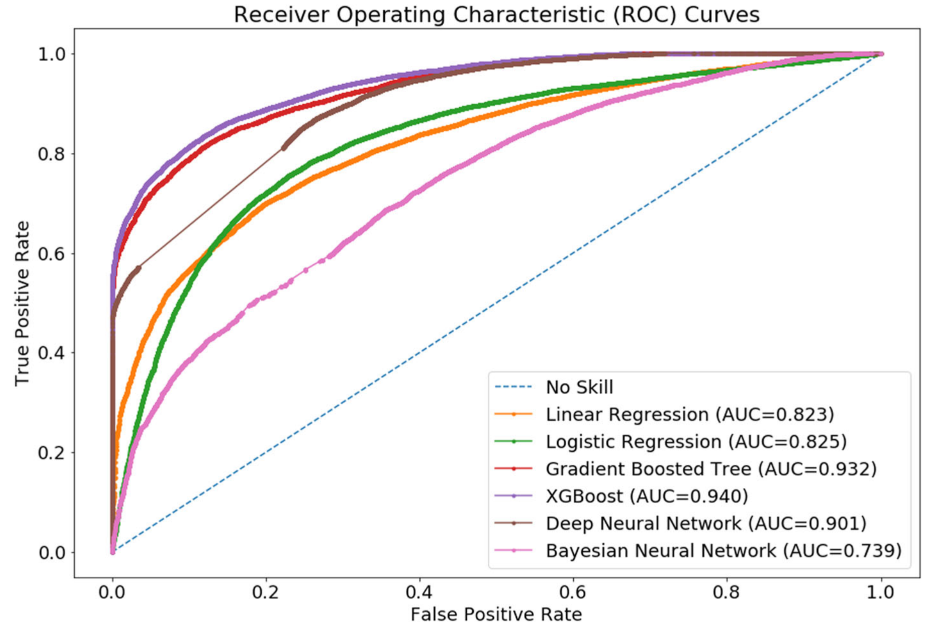

Fig. 3 ROC curves for machine learning models

ROC curves indicate that all models have good prediction performance. The dominant models are gradient-boosted tree and XGBoost models with highest AUC (0.903–0.94) and perform better in early-retrieval area. The early-retrieval area is used for evaluating a small part of the dataset with high-ranked items (Saito and Rehmsmeier 2015; Truchon and Bayly 2007) and also useful to check when the dataset is unbalanced (Weng and Poon 2008). XGBoost and gradient-boosted tree perform better at lower FP rate and reach 100% TP rate much earlier than other models. This indicates that these two models perform well on both classes (correct and incorrect answers). Bayesian neural network received the lowest AUC (0.739) and ROC curve is lower than for linear regression and logistic regression, even though the accuracy of Bayesian neural network (80%) is higher of equal to linear and logistic regression (see Table 3). The ROC curve of deep neural network has missing TP and FP values for probability threshold 0.04–0.25. However, it performs better than linear regression, logistic regression, and Bayesian neural network in the early-retrieval area.

Another useful performance measure is the precision-recall (PR) curve analysis (Boyd et al. 2013; Buckland and Gey 1994). Several studies showed that precision-recall curves are more applicable for imbalanced datasets and ROC curve is more applicable for datasets with equal distribution of classes (Davis and Goadrich 2006; Saito and Rehmsmeier 2015). A dataset is called imbalanced when the ratio between classes has big difference (e.g., 100:1) (Saito and Rehmsmeier 2015; Wu and Chang 2003). Our dataset contains 22% of incorrect answers and 78% of correct answers, which can be considered as an unbalanced dataset with one majority class (correct answers). Therefore, we should also look into precision-recall curve analysis. It summarizes the trade-off between the true-positive rate (TP) and the positive predicted value. In our study, the PR curve is for two classes (correct/incorrect answer) and based on plotting precision value on the y-axis and recall value on the x-axis. Figure 4 shows precision-recall curves for all machine

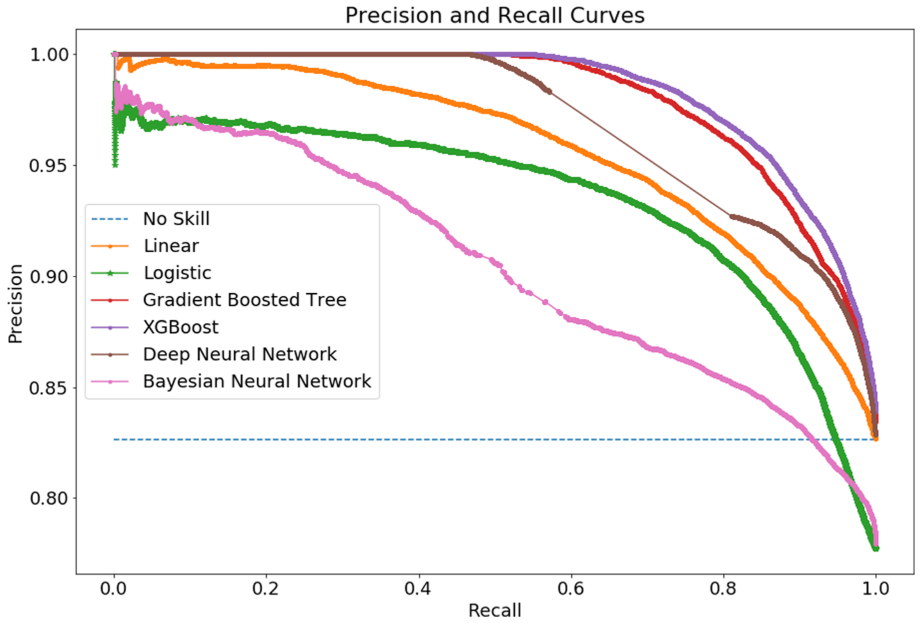

Fig. 4 Precision and recall curves

learning models. The current implementation of RCM does not provide the probabilities; therefore, it was excluded from the PR curve analysis. Similarly to ROC, the dominant models are gradient-boosted tree and XGBoost.

For measuring the response time, the experiment was repeated 100 times and for each model the average response time of training and testing was measured and presented in Fig. 5.

The most computational expensive models are deep neural network (due to running it on CPU and not on GPU) and Bayesian neural network because of variational inference step (Mullachery et al. 2018). RCM received the lowest time for training the model (3 ms) because the training process is to transform original data to vector representation and

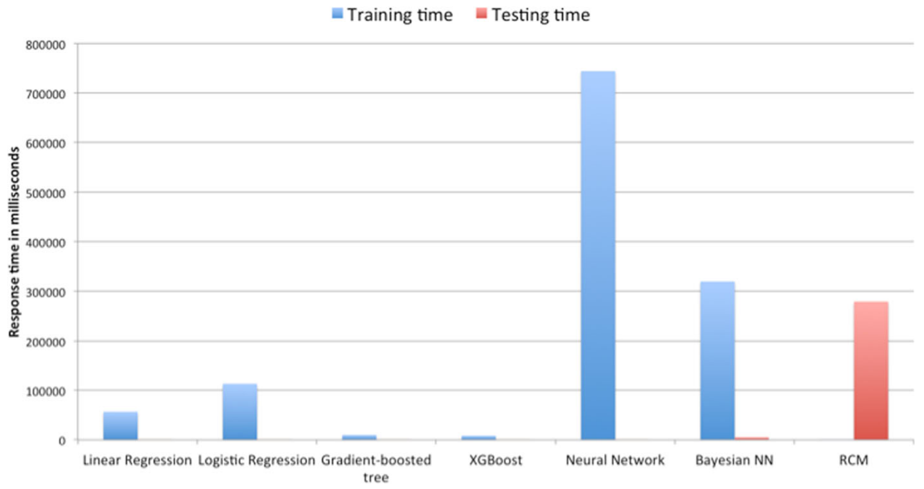

Fig. 5 Response time in milliseconds for training and testing the model

save it in to an array of examples. However, RCM is the most computational expensive in testing mode. This can be explained by implementation solution: currently, the vectors are stored as arrays; therefore, when calculating the distance between vectors, we use loops that are not efficient when the dataset is big because of loops. As a result, gradient-boosted tree and XGBoost models performed best in terms of accuracy and response time (having the lowest response time in training and testing the model).

In the second experiment, we added the reading features (F3, F4, and F12) and repeat it on the 119,230 quiz records (multiple-choice and text question types). The total amount of quiz records changed because some of the students did not read the learning material corresponding to the question in quiz. The results of the second experiment are shown in Table 5.

As shown in Table 5, the accuracy increased in all models in a range between 2 and 6%. That indicates there is some non-linear correlation between reading sessions and quiz performance (r <0.1). The Pearson correlation coefficients decreased almost in all models except RCM. The response times for all models did not change after adding reading features and it is as in the first experiment (see Fig. 5).

Discussion

The purpose of this study is to explore the association between online learning activities (independent variables) and quiz performance (response variable), and experimental assessment of the accuracy and performance of the different approaches to predict the quiz performance (such as probability that a student will answer the question correctly). The correlation analysis results (see Table 3) showed weak linear relationship between online learning activities and quiz performance. Only few features such as question difficulty level, incorrect answers per session, and chapter received low (significant) correlation. No significant correlation was found between reading features and quiz performance. Hence, the machine learning results (presented in Table 5) show that adding reading features has positive effect in increasing the prediction accuracy from 2 to 6%. This might be explained by having non-linear association between reading features and student's answers. In terms of accuracy, the RCM performed similarly to machine learning models and received the lowest accuracy value (82%). However, machine learning models have lack of transparency in their decision processes due to a black-box implementation or a high complexity that is difficult to explain the reasoning of received results (Strobel 2019). The core of our RCM is the multidimensional vector space with computing either distance measures or similarity metrics. Thus, RCM is transparent in making decisions and easy to explain why this decision was taken (e.g., the vector has absolute

Table 5 Evaluation results with reading features

Model	FP, %	FN, %	Precision, %	Recall, %	Accuracy, %	F1, %	*r*
Linear regression	90	1	84	99	83	91	0.20
Logistic regression	88	2	84	98	83	91	0.21
Gradient-boosted tree	49	3	90	97	89	93	0.56
XGBoost	48	2	91	97	89.5	94	0.59
Deep neural network	70	1	87	99	87	89.5	0.45
Bayesian NN	96	0	83	100	83	83	0.11
RCM	51	10	89	89	82	89	0.37

minimal values for all distances or similarity metrics). There are several solutions to increase the RCM response time: (a) re-implement our approach as a distributed submodule which allows execution of multiple parallel operations with using analytics engine for Big Data such as Apache Spark; (b) based on our proposed approach in our previous study (Sotsenko et al. 2016a), group similar entities into clusters, then compute similarity to the center of the cluster, and then inside of the cluster in order to reduce the computation load.

Conclusion and future work

The aims of the present research were to explore the association between online learning activities and quiz performance and to predict the quiz performance with using machine learning and RCM. Seven algorithms were applied including linear and logistic regressions, gradient-boosted tree, XGBoost, deep neural network, Bayesian neural network, and rich context model on a dataset consisting of medical students' answers on quizzes (with both multiple-choice and text questions) carried out at the Hypocampus web-based learning platform. The results show that the gradient-boosted tree and the XGBoost algorithms outperform others by obtaining the overall prediction accuracy 88–89% and with lowest response time for training and testing the model. Additionally, adding the reading features had positive effect on the overall accuracy for all models. That indicates there is some relationship between quiz performance and reading learning material time spend in e-learning platform. Hence, no significant correlation was found between reading and quiz performance. Our results indicate that it is possible to predict the probability that a student will answer the question correct before doing a quiz by analyzing student's study behavior on an e-learning platform. Further research is to design personalized spaced repetition algorithm for quizzing.

Abbreviations

IRT: Item response theory; RNN: Recurrent neural network; LSTM: Long-short term memory model; BKT: Bayesian knowledge tracing model; XGBoost: Extream gradient-boosted tree; RCM: Rich context model; ML: Machine learning; BNN: Bayesian neural network; EDM: Educational data mining; PR: Precision-recall curve; ROC: Receiving operation characteristic curve; MVSM: Multidimensional vector space model

Acknowledgements

Not applicable.

Authors' contributions

AL carried out the study and drafted the manuscript. EB provided the dataset for this study. MJ and MM contributed to the review of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data cannot be shared; please contact Elias Berg for getting the access to the dataset.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Computer Science and Media Technology, Linnaeus University, PG Vejdes väg, 351 95 Växjö, Sweden. [2] Department of Computer Science, University of Applied Sciences Ruhr West, Düsternbrooker Weg 20 24105 Bottrop, Germany. [3] Hypocampus AB, Düsternbrooker Weg 20 24105 Bottrop, Sweden.

Received: 29 January 2020 Accepted: 6 April 2021
Published online: 26 April 2021

References

Boyd, K., Eng, K.H., Page, C.D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals, In *Joint European conference on machine learning and knowledge discovery in databases*. https://doi.org/10.1007/978-3-642-40994-3_29 (pp. 451–466): Springer.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, *45*(1), 12–19.

Bucos, M. (2018). Predicting student success using data generated in traditional educational environments. *TEM Journal*, *7*(3), 617.

Chaudhry, R., Singh, H., Dogga, P., Saini, S.K. (2018). Modeling hint-taking behavior and knowledge state of students with multi-task learning. *International Educational Data Mining Society*. https://doi.org/10.29007/dj6b.

Chawla, N.V., Japkowicz, N., Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, *6*(1), 1–6.

Chen, C.M., Lee, H.M., Chen, Y.H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, *44*(3), 237–255.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system, In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). San Francisco California: Association for Computing Machinery New York NY United States.

Choffin, B., Popineau, F., Bourda, Y. (2020). Modelling student learning and forgetting for optimally scheduling skill review. *ERCIM News*, *2020*(120), 12–13.

Chounta, I.A., Albacete, P., Jordan, P., Katz, S., McLaren, B.M. (2017). The "Grey Area": A computational approach to model the Zone of Proximal Development, In *European Conference on Technology Enhanced Learning*. https://doi.org/10.1007/978-3-319-66610-5_1 (pp. 3–16): Springer.

Cieslak, D.A., & Chawla, N.V. (2008). Learning decision trees for unbalanced data, In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. https://doi.org/10.1007/978-3-540-87479-9_34 (pp. 241–256): Springer.

Coelho, O.B., & Silveira, I. (2017). Deep learning applied to learning analytics and educational data mining: A systematic literature review, In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol. 28*. https://doi.org/10.5753/cbie.sbie.2017.143 (p. 143).

Davis, D., Chen, G., Van der Zee, T., Hau_, C., Houben, G.J. (2016). Retrieval practice and study planning in moocs: Exploring classroombased self-regulated learning strategies at scale, In *European conference on technology enhanced learning* (pp. 57–71): Springer.

Davis, J, & Goadrich, M. (2006). The relationship between precision-recall and ROC curves, In *Proceedings of the 23rd international conference on Machine learning*. https://doi.org/10.1145/1143844.1143874 (pp. 233–240).

Davis, D., Kizilcec, R.F., Hau_, C., Houben, G.J. (2018). The half-life of mooc knowledge: a randomized trial evaluating knowledge retention and retrieval practice in moocs, In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 1–10).

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.

Duong, H., Zhu, L., Wang, Y., Heffernan, N.T. (2013). *A prediction model that uses the sequence of attempts and hints to better predict knowledge: "Better to attempt the problem first, rather than ask for a hint"*, (pp. 316–317): EDM.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fellman, D., Lincke, A., Jonsson, B. (2020). Do individual differences in cognition and personality predict retrieval practice activities on moocs? *Frontiers in psychology*, *11*, 2076.

Galvez, J., Guzman, E., Conejo, R., Millan, E. (2009). Student knowledge diagnosis using item response theory and constraint-based modeling, In *Artificial Intelligence in Education (AIED-2009)Ů Building learning systems that care: from knowledge representation to affective modelling (Vol. 200)* (pp. 291–298): IOS Press.

Gounaris, A., & Torres, J. (2018). A methodology for spark parameter tuning. *Big data research*, *11*, 22–32.

Guo, B., Zhang, R., Xu, G., Shi, C., Yang, L. (2015). Predicting students performance in educational data mining, In *2015 International Symposium on Educational Technology (ISET)*. https://doi.org/10.1109/iset.2015.33 (pp. 125–128). Wuhan: Institute of Electrical and Electronics Engineers Inc, IEEE Computer Society.

Hodara, M., Jaggars, S., Karp MJM (2012). Improving developmental education assessment and placement: Lessons from community colleges across the country. (CCRC Working Paper No. 51). New York: Community College Research Center.

House, S.K., Sweet, S.L., Vickers, C. (2016). Students' perceptions and satisfaction with adaptive quizzing. *AURCO Journal*, *22*(Spring), 104–110.

Ibrahim, Z, & Rusli, D. (2007). Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression, In *21st Annual SAS Malaysia Forum, 5th September*. Kuala Lumpur, Malaysia.

Joseph, E. (2005). Engagement tracing: using response times to model student disengagement. *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology*, *125*, 88.

Karpicke, J.D., & Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Khajah, M.M., Huang, Y., González-Brenes, J.P., Mozer, M.C., Brusilovsky, P. (2014). Integrating knowledge tracing and item response theory: A tale of two frameworks, In *Proceedings of Workshop on Personalization Approaches in Learning Environments (PALE 2014) at the 22th International Conference on User Modeling, Adaptation, and Personalization* (pp. 7–15). Pittsburgh: University of Pittsburgh.

Kohavi, R., & et al (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1137–1143). San Francisco: Morgan Kaufmann.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, *18*(1), 430–474.

Lincke, A, Jansen, M, Milrad, M, Berge, E. (2019). Using data mining techniques to assess students' answer predictions, In *The 27th International Conference on Computers in Education (Vol. 1)* (pp. 42–50). Kenting: Asia-Pacific Society for Computers in Education.

Maldonado-Mahauad, J., Perez-Sanagustin, M., Kizilcec, R.F., Morales, N., Munoz- Gama, J. (2018). Mining theory-based patterns from big data: Identifying selfregulated learning strategies in massive open online courses. *Computers in Human Behavior*, *80*, 179–196.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al (2016). Mllib: machine learning in apache spark. *The Journal of Machine Learning Research*, *17*(1), 1235–1241.

Mullachery, V., Khera, A., Husain, A. (2018). Bayesian neural networks. arXiv preprint arXiv:180107710.

Papoušek, J., & Pelánek, R. (2015). Impact of adaptive educational system behaviour on student motivation, In *International Conference on Artificial Intelligence in Education* (pp. 348–357). Madrid: Springer.

Pardos, Z.A., & Heffernan, N.T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model, In *International conference on user modeling, adaptation, and personalization* (pp. 243–254). Girona: Springer.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3-5), 313–350.

Pentreath, N. (2015). *Machine learning with spark*. Birmingham: Packt Publishing Ltd.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J. (2015). Deep knowledge tracing, In *Advances in neural information processing systems* (pp. 505–513). Montreal: MIT Press.

Reise, S.P., & Revicki DA (2014). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge: Taylor & Francis, New York & London.

Roediger III, H.L., & Butler, A.C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, *15*(1), 20–27.

Roediger III, H.L., & Karpicke, J.D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249–255.

Ross, B., Chase, A.M., Robbie, D., Oates, G., Absalom, Y. (2018). Adaptive quizzes to increase motivation, engagement and learning outcomes in a first year accounting unit. *International Journal of Educational Technology in Higher Education*, *15*(1), 30.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0118432.

Seber, G.A., & Lee, A.J. (2012). *Linear regression analysis, vol. 329*. New York: Wiley.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning, In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1848–1858).

Shahiri, A.M., Husain, W., et al (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422.

Simon-Campbell, L., Phelan, J., et al (2016). Effectiveness of an adaptive quizzing system as an institutional-wide strategy to improve student learning and retention. *Nurse educator*, *41*(5), 246–251.

Sotsenko, A. (2017). *A rich context model: Design and implementation. PhD thesis, Faculty of Technology, Linnaeus University*. Växjö.

Sotsenko, A., Jansen, M., Milrad, M., Rana, J. (2016a). Using a rich context model for real-time big data analytics in twitter, In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops* (pp. 228–233). Vienna: IEEE Computer Society.

Sotsenko, A., Zbick, J., Jansen, M., Milrad, M. (2016b). Flexible and contextualized cloud applications for mobile learning scenarios. *Mobile, ubiquitous, and pervasive learning*, 167–192. Springer.

Strobel, M. (2019). Aspects of transparency in machine learning, In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 2449–2451). Richland: International Foundation for Autonomous Agents and Multiagent Systems.

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., Gomez-Rodriguez, M. (2019). Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, *116*(10), 3988–3993.

Thiede, K.W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of experimental psychology: Learning, Memory, and Cognition*, *25*(4), 1024.

Ting, K.M. (2010). Confusion Matrix. *Encyclopedia of machine learning*, *1*, 260–260. Springer, Boston.

Truchon, J.F., & Bayly, C.I. (2007). Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, *47*(2), 488–508.

Van der Zee, T., Davis, D., Saab, N., Giesbers, B., Ginn, J., Van Der Sluis, F., Paas, F., Admiraal, W. (2018). Evaluating retrieval practice in a mooc: How writing and reading summaries of videos affects student learning, In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 216–225).

Weng, C.G., & Poon, J. (2008). A new evaluation measure for imbalanced datasets, In *Proceedings of the 7th Australasian Data Mining Conference-Volume, vol. 87*. https://doi.org/10.1109/ijcnn.2011.6033267 (pp. 27–32).

Wu, G., & Chang, E.Y. (2003). Class-boundary alignment for imbalanced dataset learning, In *ICML 2003 workshop on learning from imbalanced data sets, vol. II* (pp. 49–56). Washington.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.