Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Using Data Mining Techniques to Assess Students' Answer Predictions

Alisa LINCKE[a*], Marc JANSEN[a,b] , Marcelo MILRAD[a,] Elias BERGE[c]
[a]*Computer Science and Media Technology Department, Linnaeus University, Sweden*
[b]*Affiliation B, University of Applied Sciences Ruhr West, Germany*
[c]*Hypocampus AB, Sweden*
*alisa.lincke@lnu.se

Abstract: Estimating students´ knowledge and performance, modeling their learning behaviors, and discovering and analyzing their different characteristics are some of the main tasks in the field of research called educational data mining (EDM). According to Chounta (2017), the predicted probabilities that a student will answer a question correctly can provide some insights into the student´s knowledge. Based on this point of departure, the main objective of this paper is to apply different data mining techniques to predict the probabilities that students will answer questions correctly by using their interaction records with a web-based learning platform called Hypocampus. Five different machine learning algorithms and a rich context model were used on the Hypocampus dataset. The results of our evaluation indicate that the gradient-boosted tree and the XGboost algorithms are best in predicting the correctness of the student's answer.

Keywords: adaptive learning, machine learning, rich context model, answers probability prediction

1. 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, 2015). In order to efficiently provide students with personalized and adaptive digital content and a meaningful learning experience, it is crucial that the learning system gets over time an *understanding* not only of the students' current knowledge level but also his/her progression. One traditional way of assessing the knowledge level is letting students take a placement test (Hodara, 2015). However, to make a placement test adaptive, the system needs to be able to draw conclusions from every answered question. According to Chounta (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 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. If we know the probabilities of students answering questions correctly then we can optimize the studies with regard to engagement and knowledge level. Using probabilities, we can objectively measure a student's knowledge on a particular given subject. This measurement can be used as a valuable feedback to the students. Previous studies have suggested that an adaptive fail rate in a quiz increases student engagement (Papoušek, 2015). 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 were the average student experiences the quiz challenging without being too difficult. Therefore, this paper aims to estimate the probability that students will answer questions correctly by using different data mining techniques on data provided by the Hypocampus[1] system.

[1] https://www.hypocampus.se

Hypocampus is an adaptive web-based learning platform used by medical students. It contains a library with many interactive reading materials (e.g., 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, 2008).

Our research contributes to the student knowledge estimation research area with a particular focus on: (a) *providing a set of features that allows getting good prediction accuracy on students´ answers (Table 2)*; (b) *an approach that works on the subject based level*; (c) *an approach that uses different data types besides the binary representation of students' knowledge state*. The remainder of this paper is structured as follows. Section 2 provides a short overview of the challenges and existing research related to modeling and predicting students' knowledge. Section 3 describes the proposed approach, our dataset, feature extraction, and the models used in this study. In Section 4, we present and discuss the evaluation results of our efforts while in Section 5 we provide our conclusions and outline and present possible lines of future work.

2. Related Work

Modeling and predicting the knowledge of students in online learning systems is a well-identified problem (Piech, 2015; Duong, 2013; Pelánek, 2017; Pardos, 2011; Tato et al., 2017). Unknown students' knowledge background, access to the learning resources such as reading material, quizzes, exams, courses at any time and order brings different learning behaviors (e.g., accessing the learning material in different sequences, some students just doing quizzes, 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 reading resources besides those provided by the learning platform (such as books and notes from the classes). All of these are examples of challenges to model students' knowledge.

One approach for measuring academic achievements and student's knowledge is the Item Response Theory (IRT) models (Reise, 2014; Chen, 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 (Chaundhry, 2018; Galvez, 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 (Chaundhry, 2018; Khajah, 2014).

Students generate a vast amount of interactional data in online learning platforms that allows to use data mining and machine learning techniques to better estimate students' knowledge and performance. A Recurrent Neural Network (RNN) is a type of artificial neural network for processing sequential data such as text, videos, sensors data, and stock markets (Hecht-Nielsen, 1992). RNN takes into account the current information and the previous/historical information to model student's knowledge. Student´s study behavior data collected in web-based learning systems can be represented as sequences of study behaviors. For example, a sequence of performed exercises (Piech, 2015), a sequence of questions answered in a quiz, a sequence of reading and quiz sessions, a sequence of attempts and hints to solve a programming task (Wang, 2017; Duong, 2013). These study behavior sequences can be used as an input to RNN in order to predict whether the student will complete successfully a new exercise, or to predict the next type of learning activity (reading/quiz). Usually, the data consists of binary variables indicating whether the student will complete the task successfully or not and the label indicating the skill name or knowledge component (Wilson, 2016, Piech, 2015). The possible concern with RNN is the vanishing and exploding gradients (Xiong, 2016) that influence the accuracy and performance of the algorithm. Long short-term memory (LSTM) is introduced as a solution to the vanishing problem (Hochreiter, 1997). LSTM is an extension for RNN that allows to remember the inputs over a long period of time and decide whether to store, not store or delete the information based on the importance of that information (Hochreiter, 1997). In a study predicting students' learning gains, Lin (2017) compared BKT, RNN and LSTM models. The LSTM model showed the highest accuracy in predicting students' learning gains.

Different classification models such as decision trees, random forest classifier, logistic regression classifier and vector support machine were used to predict whether a student would pass or fail a final exam (Bucos, 2018; Bunkar et al., 2012). In their study, the support vector machine and logistic regression obtained the best accuracy in predicting failing or passing the exam (Bucos, 2018). The features dataset that they have used contains information about students' gender, average grade, past examination grades, class attendance and others that are not always available in distance learning web based educational systems. Mueen (2016) has analyzed collected data from learning management systems (LMS) (such as forum data, assignments grades, learning material access duration, quizzes) in order to predict the students' performance by using different machine learning algorithms. In his study, the best accuracy score (86%) was achieved by a Naïve Bayes classification algorithm (Mueen, 2016). Tato et al. (2017) used probabilistic approach, a Bayesian neural network of logical reasoning skills to predict the learner's knowledge state with around 85% accuracy.

In summary, extensive research has been carried out on modeling students' knowledge and performance. However, to the best of our knowledge, it is still not clear how existing models handle different features collected by on-line learning platforms (for example, time answering the questions, time since last reading the material, time since last time doing quiz, repetitions) and how they model subject based knowledge and not skill based knowledge, without experts annotations or input, without information about grades, class attendances, and personal information (e.g., gender, age, study year).

3. Our proposed approach

The approach suggested in this paper is to predict the probability whether a question will be answered correctly based on a general machine-learning (ML) pipeline (Pentreath, 2015), see Figure 1 below. There are five main steps: data collection (Dataset), data preprocessing, feature extraction, selection and transformation, model application, and the last step is the model evaluation. This process is iterative and can be repeated many times until the model that performs best will be defined. In terms of technologies and tools, we have used Apache Spark MLlib and Scikit-learn libraries for performing data preprocessing, feature extraction, and transformation steps. In the model application step we use different machine learning models from the Apache Spark MLlib (Meng, 2016), the Keras deep learning library, and the Contextualization Service (Sotsenko, 2016).

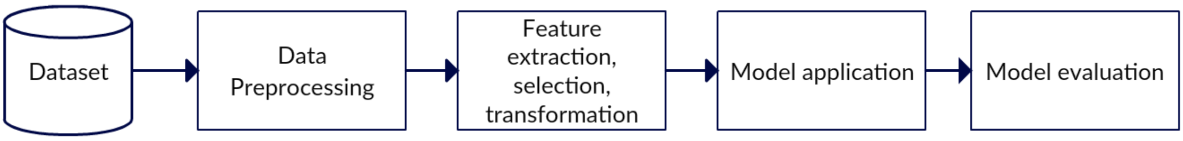

Figure 1. ML pipeline for predicting the probability if a question will be answered correctly.

3.1 Dataset

For this study, we have obtained the data from the Hypocampus web-based learning platform. The platform offers reading material for various subjects (e.g., Dermatology, Surgery, Gynecology, Internal Medicine and others). Every subject has a number of topics called chaptergroups. Every chaptergroup consists of chapters and quizzes. Most of the quizzes contain 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. 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 for 300 medical students used as part of our study that took place over a period of 10 months between 2017-2018.

Table 1

Summary of collected data.

	Number of records
Students	300
Quiz records	121 423
Multiple-choice questions (msq)	18 092
Text questions	103 331
Correct answers (msq)	14 580
Incorrect answers (msq)	3 420

The dataset (Table 1) contains information about quiz records from different subjects gathered over a period between one to three months. Quiz records include information about user identification number, question type (multiple-choice or text), question identifier, 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, and question session. After collecting all these data, the preprocessing and feature extraction steps are performed in order to prepare the dataset to be used by data mining techniques.

3.2 Data Preprocessing

As part of the data preprocessing step we apply three filters: (a) selecting records that were collected in the system production mode; (b) selecting only multiple-choice question types, because this type of questions are more reliable for our model evaluation than text type questions (text type question correctness given to students and not to the system); (c) removing records that have missing values related to the question id information.

3.3 Feature Extraction and Transformation

We performed preprocessing on the collected data and extracted and selected specific features. After performing several iterations in the ML pipeline (Figure 1) the following 18 features were identified and described as presented in Table 2. There are five categorical features (F1, F2, F3, F4, F5), three time-related features (in seconds) (F6, F7, F8), and ten numerical features (F9, F10, F11, F12, F13, F14, F15, F16, F17, F18). Some of the features are directly used from the original dataset (e.g., user id, chapter id, question id, question session) and other features are calculated (e.g., number of correct/incorrect answers, attempt number) for each record in the dataset.

Table 2

Feature overview

	Feature Type	**Name**	**Description**
F1	Categorical	User ID	User identification
F2	Categorical	Chapter ID	Chapter identification
F3	Categorical	Question ID	Question identification
F4	Categorical	Question Session ID	Question session identification
F5	Categorical	Time of the day	Morning, lunch, afternoon, evening, night
F6	Time related features	Time since last doing quiz	Represents the time duration since last time the student was doing a quiz
F7	Time related features	Time since last reading a chapter	Represents the time duration since last time the student was reading a chapter
F8	Time related features	Reading time	Total reading time per chapter (in

			seconds)
F9 F10 F11	Numerical	Correct	Number of questions answered correctly per chapter, per attempt, per question session
F11 F12 F13	Numerical	Incorrect	Number of questions answered incorrectly per chapter, per attempt, per question session
F14	Numerical	Attempt number	Number of times a question was answered by the student
F15	Numerical	Reading sessions	Number of times a student read the learning material (chapter)
F16	Numerical	Question facility index	Represents the question difficulty and is facility index of the students correct answers in first attempt in range between 0 and 1 (where 0 is very difficult and 1 is very easy).
F17	Numerical	Question number1	Question number in the question session
F18	Numerical	Question number2	Accumulated question number: total number of questions answered correctly for a specific chapter.

The features (in Table 1) have been transformed to an appropriate format (Vectors) for machine learning models by using the Vector Assembler component from the machine learning library (Mllib) in Spark.

3.4 Models

To predict the probability if a question will be answered correctly, we use six models: linear regression, logistic regression, gradient-boosted tree regression (Apache Spark Mllib), XGBoost (Chen, 2016) (Python library), feed-forward neural network (Keras deep learning library) and a rich context model (RCM) from Contextualization Service (Sotseko, 2016). We select two simple models (linear and logistic regression), and four more advanced models (two decision trees models, deep neural network and RCM) because they are most commonly used in regression problems (predicting probabilities) and taking contextual information into account (RCM). 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. Furthermore, some of these models were used in predicting students' performance (Bucos, 2018; Shahiri, 2015; Zaidah, 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, 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 as described in Table 2. We use this model to check whether there is a linear relationship between the learning activity (quiz) 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. The following features are selected from Table 2: numerical features (F9-F18, except F15), one time related feature (F6), and one categorical feature (F5 transformed from categorical representation to numerical by using hour of the day from 0-24) based on decision that these models work with numerical data types. Features F1-F4 were used as labels in all machine learning models.

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, 2008; Chawla, 2004). Gradient boosted tree and XGBoost algorithm have been shown great success in winning machine learning competitions such as Kaggle[2]. To the best of our knowledge, in the literature review this model was not applied to student's knowledge prediction. Therefore, we decide to test this model and to apply it in our study. The same set of features was selected and used as in linear regression model.

Deep neural networks become more popular in EDM tasks (Coelho, 2017; Guo, 2015). We use a multi-layer sequential feed-forward neural network. Different model parameters were tested, the following parameters were found empirically by minimizing the error: input layer with 11 neurons activation function '*relu'*, one hidden layer with 11 neurons and activation function '*relu'*, output layer with 1 neuron and activation function "*sigmoid*", optimizer '*adam'* and loss '*binary_crossentropy'*. We use the same set of features in linear regression model.

We use also rich context model models contextual information in a multidimensional vector space model (MVSM) to provide recommendations based on the current context of a user. This model can handle different data types therefore we include categorical features in predicting the answer correctness. The 16 features (except the reading time features F7, F8 and F15 that are left for future work) used to model student's quiz records in RCM. RCM requires an *examples* 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. The *examples* dataset is transformed to one-dimensional vector and placed in MVSM (e.g., Quiz Record X shown on Figure 2).

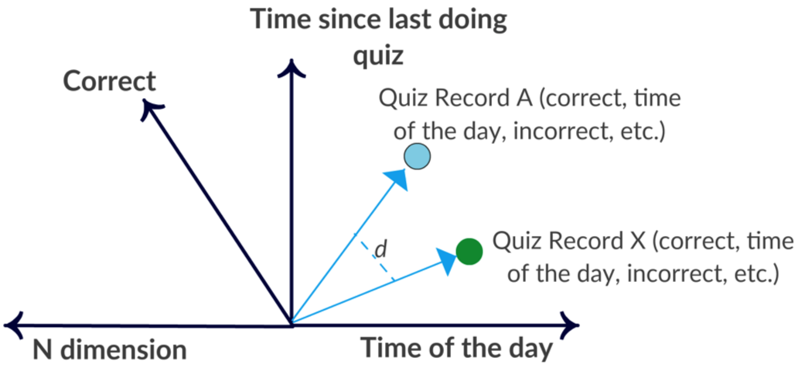
Time since last doing quiz
Correct
Quiz Record A (correct, time of the day, incorrect, etc.)
d
Quiz Record X (correct, time of the day, incorrect, etc.)
N dimension
Time of the day

Figure 2. Representation of quiz records in RCM

The *current context* datasets are transformed into a one-dimensional vector (Quiz Record A in Figure 2) and Euclidean similarity metric (*d*) is used to calculate the distance to find the most similar quiz record in *examples* dataset (Quiz Record X). The most similar quiz record that has minimal distance defines if the student will answer correct or incorrect.

4. Evaluation

The evaluation in this study is conducted using the 10-fold cross validation approach (Kohavi, 1995) and the train-validation split approach provided by the Spark Mllib library for hyper-parameter tuning (Gounaris, 2018). The dataset used consists of 18092 quiz records from medical students that have studied using the Hypocampus web-based learning platform (Table 1). As mentioned earlier in this paper, the purpose of this study is to evaluate which model predicts best the probability that a student will answer the question correct. The evaluation results are shown in Table 3 with the following metrics: false positive rate (FP,%), false negative rate (FN,%), precision (Precision,%), recall (Recall,%), accuracy (Accuracy,%), F1-Score (F1,%) and Pearson correlation coefficient (R) between the predicted value by the algorithm (described in Section 3.4) and depended variable (answer on the question). In our answer prediction task: false positive are incorrect answered questions which have

[2] https://www.kaggle.com

been predicted as correct; false negative 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.

Table 3

Evaluation results

Algorithm	FP,%	FN,%	Precision,%	Recall,%	Accuracy,%	F1,%	R
Linear Regression	62	4	82	95	81	88	0,43
Logistic Regression	72	4	80	96	79	88	0,35
Gradient-boosted tree	30	3	91	97	90	94	0,72
XGBoost	25	4	92	96	91	94	0,73
Neural Network	45	3	87	97	86	92	0,61
RCM	63	17	80	83	71	81	0,19

As shown in Table 3, the five machine learning algorithms performed well in predicting the probability that an answer will be correct, with an accuracy rate ranging between 81% and 91%. These results indicates that dataset of features very well correlates with independent variable (student's answer). The RCM model got the lowest accuracy (71%) and more false negative errors (17%) 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, and others. The best algorithms in predicting the probability that a student will answer correctly are Gradient-boosted tree and XGBoost with around 90% accuracy and highest correlation (0,72-0,73) and lowest false positive and false negative errors (FP=25%, FN=3%). However, the false positive error is the biggest error in all algorithms. We manually analyzed the errors (FP and FN) and decided to balance the data in order to reduce FP errors. As mentioned earlier the original dataset contains 81% correct answers and 19% incorrect answers that make algorithms to predict more often that student will answer correctly. Therefore, we applied one of the well-known techniques for data balancing – synthetic minority oversampling technique (SMOTE) (Chawla, 2002). This technique adds incorrect answer records by finding the k-nearest-neighbors for minority class (incorrect answers) and randomly choosing one and using it to create a similar record. We applied SMOTE only to training dataset and validation dataset we did not change. Same evaluation metrics calculated for balancing the dataset and presented in Table 4 below.

Table 4

Evaluation results (balanced training data)

Algorithm	FP	FN	Precision	Recall	Accuracy	F1	R
Linear Regression	19	17	93	82	81	87	0,57
Logistic Regression	15	14	94	86	85	90	0,65
Gradient-boosted tree	14	10	95	90	88	92	0,71
XGBoost	13	10	95	90	89	93	0,73
Neural Network	24	11	92	89	85	90	0,61
RCM	5	23	94	77	85	84	0,72

Balancing the training data improves the accuracy 14% for the RCM and 6% for the logistic regression. Pearson correlation coefficient increased for all models, more than 20% of FP errors were reduced. Almost no change in accuracy for gradient-boosted tree, XGBoost, and neural network were obtained. This can be explained due to the fact that decision trees work well on unbalanced data and neural networks do not require balancing the data. Furthermore, the FP error is reduced and balanced (FN is increased) in gradient boosted tree and XGBoost. XGBoost algorithm still provides the best results even after balancing the training dataset. RCM received the smallest FP error (5%), biggest FN error (23%) and overall improvements in accuracy, F1-score, and Pearson correlation. Based on the obtained results we can assume that RCM is not invariant to unbalanced data and requires to have balanced *example* data that is placed in MVSM as basis. Overall, this study shows that data generated in web-based learning platforms can be used to predict student's answer in order to estimate his/her knowledge on a subject.

5. Conclusion and Future Work

We applied different machine learning techniques to the problem of estimating students' knowledge by knowing the probability if the student will answer the question correctly. Six algorithms were applied including linear and logistic regression, gradient-boosted tree, XGBoost, deep neural network, and rich context model on a dataset consisting of medical students' answers on quizzes 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 90-91% and lowest false negative and false positive errors (4% and 25%). Additionally, the XGBoost algorithm performs well on unbalanced dataset with two classes that is shown in our case. The obtained results increase the prediction accuracy about 5% in comparison with other studies discussed in the related work section. 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 log data.

As part of our future research efforts, we plan to (a) add more features: reading time and time since last time read the chapter (F7, F8, F15), (b) add text type questions, and (c) apply one of the variants of deep knowledge tracing models (e.g. Bayesian neural network) and compare the obtained results with XGBoost algorithm; (d) create a measurement of students' knowledge based on the prediction probabilities results obtained from the XGBoost algorithm.

References

Baker, R. S., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *International conference on intelligent tutoring systems*. 406-415.

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

Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012). Data mining: Prediction for performance improvement of graduate students using classification. In *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*. 1-5.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

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*.

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, 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*. 785-794.

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

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*. 3-16.

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*. 241-256.

Coelho, O. B., & Silveira, I. (2017, October). 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, No. 1, p. 143).

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". *In EDM*, 316-317.

Galvez, J., Guzman, E., Conejo, R., & Millan, E. (2009). Student Knowledge Diagnosis Using Item Response Theory and Constraint-Based Modeling. In *AIED*. 291-298.

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 student's performance in educational data mining. In *2015 International Symposium on Educational Technology (ISET)*. 125-128.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception*. 65-93.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

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

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 *CEUR Workshop Proceedings*.7-15.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*. Vol. 14 (2). 1137-1145.

Lin, C., & Chi, M. (2017) Comparisons of BKT, RNN and LSTM for Predicting Student Learning Gains. In *AIED.*

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Xin, D. (2016). Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, *17*(1), 1235-1241.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, *8*(11), 36.

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*. 243-254.

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

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

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.

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*, 505-513.

Reise, S. P., & Revicki, D. A. (Eds.). (2014). Handbook of item response theory modeling: Applications to typical performance assessment.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. 329.

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

Sotsenko, A., Jansen, M., Milrad, M., & Rana, J. (2016). 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 (FiCloudW)*. 228-233.

Wang, L., Sy, A., Liu, L., & Piech, C. (2017). Learning to Represent Student Knowledge on Programming Exercises Using Deep Learning. In *Proceedings of the 10th International Conference on Educational Data Mining*. 324-329.

Wilson, K. H., Xiong, X., Khajah, M., Lindsey, R. V., Zhao, S., Karklin, Y., Van Inwegen, E.G., Han, B., Ekanadham, C., Beck, J.E., Heffernan, N. (2016). Estimating student proficiency: Deep learning is not the panacea. In *In Neural Information Processing Systems, Workshop on Machine Learning for Education*.

Xiong, X., Zhao, S., Van Inwegen, E. G., & Beck, J. E. (2016). Going Deeper with Deep Knowledge Tracing. In 9th *International Conference on Educational Data Mining*.

Zaidah I., Daliela R. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression. In *21st Annual SAS Malaysia Forum*.