Analysis of Bias in Student Modelling and Prediction Using Learning Analytics on edX MOOC Data

Pedro Manuel Moreno-Marcos[1,*], Alicia Ramirez-Arrabe[1,2], Carlos Alario-Hoyos[1], Pedro J. Muñoz-Merino[1], Iria Estévez-Ayres[1] and Carlos Delgado Kloos[1]

[1]Department of Telematic Engineering, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), Spain, ROR: 03ths8210
[2]Deparment of Language and Information Systems, Universidad Nacional de Educación a Distancia (UNED), 28040, Madrid, Spain

Abstract

Massive Open Online Courses (MOOCs) usually have very high dropout rates, and this has produced a vast amount of research about how to determine students' abilities and predict dropout and student success. However, many of those works make some assumptions, e.g., filter data, or decisions when designing the models, that may lead to bias. These biases have already been studied in Intelligent Tutoring Systems but can be adapted to MOOCs. For example, if the difficulty of exercises is computed considering students who do not engage with the course, this difficulty might be overestimated. In this line, this work aims to analyze some possible bias in the estimation of students' abilities, difficulty of exercises and dropout and success prediction. In order to do that, data from a MOOC on Java programming are used. Results show that the students' ability and difficulty of exercises computed with the Item Response Theory (IRT) are more balanced when only considering students who have engaged with graded exercises, and other situations (e.g., students who have engaged with at least one video or the forum) might produce significant bias. Regarding predictive models, results show that using cumulative data is better than using only data from one week, and students' ability generated with the IRT significantly improves the models. In addition, the attrition bias is significant and may produce differences up to 0.15 in RMSE, unlike the mastery attrition bias, which is not frequent in the MOOC.

Keywords

Learning Analytics, Item Response Theory, Bias, Prediction, MOOCs

1. Introduction

There is a great interest in developing predictive models in the literature, mainly models to predict dropout and student success [1]. This research can be done in many different contexts, including undergraduate and postgraduate programs to determine if students drop out the program [2], higher education courses, [3] etc., and particularly in Massive Open Online Courses (MOOCs), where many works have focused due to its high attrition rates [4].

In order to develop the predictive models, many variables have been considered including variables related to engagement with videos [5], exercises [6], forum [7], course activity [8], students patterns [9], etc. In addition, it is possible to use more complex techniques to get information about students' ability. For example, students' skills can be inferred using the Bayesian Knowledge Tracing (BKT) [10] or the Item Response Theory (IRT) [11]. The latter technique allows estimating students' skills and the difficulty of exercises based on students' answers to different items. Moreover, it has three models depending on the number of parameters, and the most complete one also allows obtaining how discriminative an exercise can be and the probability of guessing an exercise. Each exercise has a

LASI Spain 26: Learning Analytics Summer Institute Spain 2026, Zamora, May 07–08, 2026
*Corresponding author.
pemoreno@it.uc3m.es (P. M. Moreno-Marcos); aramirez@lsi.uned.es (A. Ramirez-Arrabe); calario@it.uc3m.es (C. Alario-Hoyos); pedmume@it.uc3m.es (P. J. Muñoz-Merino); ayres@it.uc3m.es (I. Estévez-Ayres); cdk@it.uc3m.es (C. Delgado Kloos)
https://www.it.uc3m.es/pemoreno/ (P. M. Moreno-Marcos); https://www.it.uc3m.es/calario (C. Alario-Hoyos); https://www.it.uc3m.es/pedmume/ (P. J. Muñoz-Merino); https://www.it.uc3m.es/iria/ (I. Estévez-Ayres); https://www.it.uc3m.es/cdk/ (C. Delgado Kloos)
0000-0003-0835-1414 (P. M. Moreno-Marcos); 0000-0002-3082-0814 (C. Alario-Hoyos); 0000-0002-2552-4674 (P. J. Muñoz-Merino); 0000-0002-1047-5398 (I. Estévez-Ayres); 0000-0003-4093-3705 (C. Delgado Kloos)

corresponding Item Characteristic Curve (ICC).

However, the use of predictive models and techniques to estimate abilities might be biased because of different factors. For example, the attrition bias [12] occurs when only the students who usually get the correct answers attempt the exercises, and this produces an underestimation of the difficulty of exercises (they are considered easier). In contrast, in Intelligent Tutoring Systems (ITS), students usually stop doing exercises when they master the skills. This might produce that some exercises are only attempted by students with lower level of the skills, and therefore the difficulty of the exercise may be overestimated. This last phenomenon is usually referred to as mastery attrition bias [12].

For the case of MOOCs, as dropout is very common, it is possible that only a small percentage of enrolled students actually engage with the course. In order to address this issue, researchers often filter the dataset to exclude students who have not interacted with the MOOC. Nevertheless, the way to filter data may vary depending on the study. For example, authors in [13] used different filtering criteria, such as students who had started in the first two weeks, students who had filled a pre-course survey, etc. Authors in [14] only considered learners who had participated in the forum. However, this filter may also affect the results and produce certain bias. A clear example occurs when all students, including those who have never interacted with the MOOC are considered. In this case, most of the learners drop out and that could bias the models. In addition, students' ability might be biased if students who drop out are considered with low ability, as this might also affect the ability of the rest.

While these factors might considerably bias the results, they are not usually addressed. In order to analyze this situation, this paper aims to analyze possible biases that appear when estimating ability and difficulty of exercises with IRT, and when developing predictive models. This way, the paper contributes with insights about how some factors might bias the results. In order to structure the analyses, two main research questions are presented:

- How can different subsets of data influence the estimation of the IRT parameters and students' abilities?
- How can different model configurations and data-related factors, including well-known biases, influence the predictive power in dropout and success prediction?

The remainder of the paper is as follows. Section 2 details the methodology of the paper, including the course context, variables, analytical methods, and metrics. Section 3 presents the main results of this paper. Section 4 summarizes the main conclusions and outlines the limitations and future research directions.

2. Methodology

2.1. Course context

The analyses were carried out using data from a MOOC about Java programming (Introduction to Programming with Java Part 1: Starting to Program in Java). This course was hosted in edX in a synchronous mode. The course was structured in five weeks, and each of them had a graded exam, with 5-8 exercises each. The weight of each test was 15% and the 25% remaining corresponded to two coding projects to be handed by students. However, due to data limitations, this analysis focuses only on graded tests and the final grade used for prediction is the average grade of the five graded tests. In addition to the tests, the course contained several videos per week (72 in total) and formative exercises (135 in total). In order to pass the course, students needed to get 60% of the points at least, and 88,324 users performed at least one action in the course. Regarding the demographic information, USA was the country of residence for the majority of students (25%), followed by India (16%).

2.2. Variables

In order to conduct the analyses, tracking logs from edX were used. Particularly, the following events were considered [15]: (1) *problem_check*, (2) *problem_show*, (3) *play_video*, (4) *pause_video*, (5) *stop_video*,

(6) *edx.forum.thread.created*, (7) *edx.forum.response.created*, and (8) *edx.forum.comment.created*. With these events, independent and dependent variables were computed. The full list of independent variables is shown in Table 1. For these variables, it is important to note that all the grades of the exercises have been normalized; that is, all variables which represent any type of grade have values between 0 (totally incorrect) and 1 (perfectly solved). In addition to these variables, students' ability is computed using IRT and analyzed in Section 3.1, and it is also used in the predictive models as an independent variable in Section 3.2.

Table 1
Summary of the independent variables

Type	Name	Value range	Description
Exercises	n_attempted	0,1,2,3...	Total number of times students have attempted exercises (each attempt is counted separately)
	per_attempted	0.0-1.0 (%)	Percentage of problems attempted by a student over the total number of exercise
	CFA	0.0-1.0 (%)	Percentage of exercises which the student has solved correctly on their first attempt
	avg_attempts	0,1,2,3...	Average number of attempts a student has taken in total to solve the exercises.
	avg_grade	0.0-1.0 (decimal)	Average grade of a student of the exercises they have attempted at least once.
	n_show	0,1,2,3...	Total number of times a student asks for the solution of a problem to be shown.
Videos	n_videos_watched	0,1,...,69	Total number of videos the student has watched
	per_videos	0.0-1.0 (%)	Percentage of content visualized over the total
	n_videos_opened	0,1,2,3...	Number of times a student opened a video
	n_pauses	0,1,2,3...	Number of times a student paused a video
Forum	n_threads	0,1,2,3...	
	n_responses	0,1,2,3...	Total number of threads created by a student
	n_comments	0,1,2,3...	number of responses posted by each student.
Activity	n_days	1,2,...,49	Total number of comments posted by a student

Regarding the dependent variables, two variables have been defined. The first one is *dropout*, which is a binary variable indicating whether or not the student has given up the course. In this analysis, dropout is considered when the student stopped doing the exam and failed the course. Thus, two conditions are considered: (1) they did not do the exams belonging to Week 4 and Week 5 and did not pass the course, or (2) they did not do the exam corresponding to Week 5 and did not pass the course. The second variable is *final_grade*. This variable represents the final grade of each student. It is a continuous variable. Moreover, since grades have been normalized, it can take values from 0 to 1.

2.3. Analytical methods and metrics

For the analyses, several techniques were used. First, IRT was used to analyze possible biases in the estimation of students' ability and difficulty of exercises. For this case, the 35 graded problems are considered and the problems are polytomous (grade is continuous from 0 to 1, with unattempted exercises treated as zero to analyze the bias when using this imputation). With these problems, the Three-Parameter Logistic Model has been used to obtain more information. Particularly, this model offers three parameters: difficulty, discrimination and guessing.

Second, for the predictive analyses, several machine learning algorithms were considered: (1) Regression (Logistic Regression for dropout prediction and Ridge Regression for final grade prediction), (2) Decision Trees, (3) Random Forest, (4) K-Nearest Neighbors, (5) Support Vector Machines, and (6) XGBoost. These algorithms have been trained using the Pipeline and GridSearchCV Scikit-Learn tools in order to explore different combinations of the models parameters and find the best model as possible. At last, the training and evaluation approach has consisted on a train set containing 75% of the data and

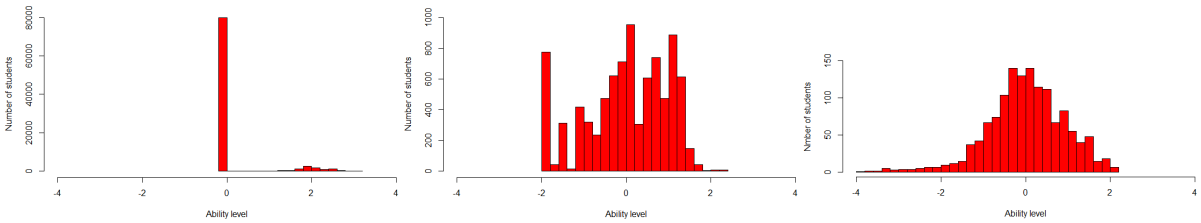

Figure 1: Histogram of ability distribution for the three configurations.

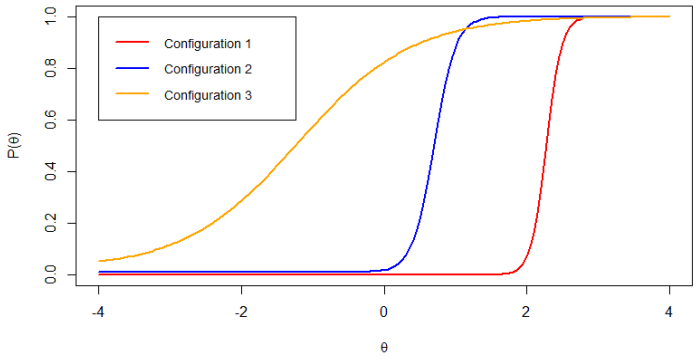

Figure 2: Configurations 1, 2 and 3: Average item ICC

a test set with 25% of the data. As for the metrics, AUC was used for dropout prediction and RMSE was used for final grade prediction, following recommendations in [16].

3. Results

3.1. Analysis of bias in IRT

IRT has been widely used in the literature to estimate students' ability and difficulty of exercises. However, there might be biases when many students do not attempt all the exercises. This can have a crucial impact in MOOCs where most of the student drops out. In order to analyze this, three configurations of students have been considered to filter students differently and analyze the IRT with different subsets of students. The configurations are as follows:

- Configuration 1: All students
- Configuration 2: Students who answered at least one graded exercise.
- Configuration 3: Students who answered all graded exercises.

These configurations were selected to capture the extremes of student engagement, from the total population to those showing total commitment to the course. This allows for a clear comparison of how different levels of engagement affect IRT parameters. With these configurations, students' ability and difficulty of exercises for all the 35 exercises have been computed. The histogram of the distribution of ability and the average ICC of the IRT is displayed in Figs. 1 and 2.

The first configuration includes all the 88,324 students and as many of them do not attempt the exercises (around 90%), there is a large bin located between ability levels -0.2 and 0 (with more than 80,000 students). Moreover, there is small bin around ability level 2, for those who took the exercises. In this case, the average difficulty is 2.27, which is too high, as this means that an ability of 2.27 is required to have a 50% chance of answering correctly in the items. Furthermore, the discrimination is extremely

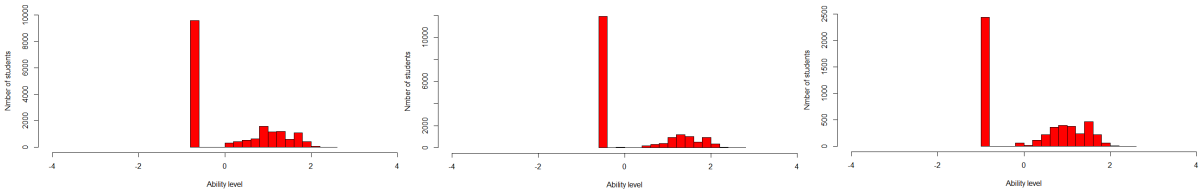

Figure 3: Histogram of ability distribution for configurations 4-6.

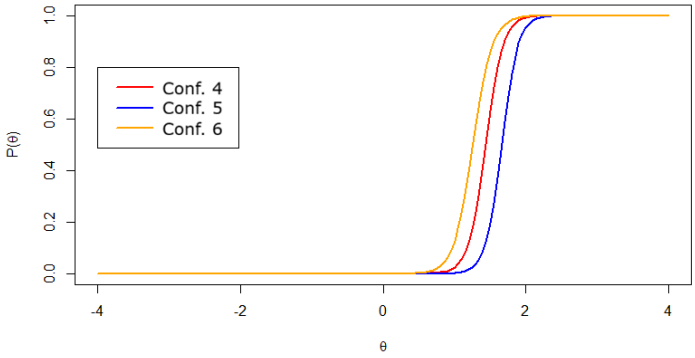

Figure 4: Configurations 4, 5, and 6: Average item ICC

high (9.56). This is observed in the average ICC, as students with an ability smaller than 2 have a very small change of success, while this increases very steeply after this value. This is an unrealistic situation and shows that this configuration produces significant bias in the results.

In the second configuration, only students who have attempted at least one problem are considered. This reduces the sample to 8,708 students. In this case, it can be observed that the ability levels are more distributed than in the first case. In addition, the smallest value is -1.906 and the highest is 2.217, and there is still a significant peak on the left, which belongs to the lowest ability levels. This probably includes students who attempted very few exercises or got very low grades. Regarding the IRT parameters, the difficulty drops to 0.7, which is more reasonable, although the discrimination is still high (6.72).

The third configuration only includes students who answered all the items (n=1,371). In this case, the histogram is a bit left-skewed, with a minimum ability of -3.83. Unlike the previous cases, students located on the left are exclusively students who answered incorrectly. On the other side, the highest value is 2.09. Comparing both extremes, this configuration is not as balanced as the second one, and most of the students are concentrated around 0. Moreover, only 51 students (3.7%) obtain an ability greater than 1.5. In this case, as all students attempt the exercises, it is more difficult to stand out from the rest. As for the IRT parameters. The difficulty is -1.21, which is very different from the other cases. As all the students attempt the exercises now, items are considered easier. The discrimination parameter is also reduced to 1.25. Thus, significant biases might be produced depending on the selection.

In order to further delve into this issue, three additional configurations are analyzed. Configuration 4 includes students who did at least one non-graded problem (n=17,648). Configuration 5 includes students who watched at least one video (n=17,729). Configuration 6 includes students who contributed to the forum at least once (n=4,964). The equivalent charts are displayed in Figs. 3 and 4.

Results show that all the three cases depict a strong bias on the left side of the scale. This also happened in configuration 1, because it includes students who had not responded to any graded exercise. When analyzing the ICC curves, all of them are very similar, and in all of them, items are considered difficult and the discrimination is very high. Thus, filtering by other variables different from the graded

Table 2
Prediction of dropout and final grades with different algorithms

Algorithm	Dropout Prediction (AUC)	Final grade prediction (RMSE)
Regression	0.934	0.164
Decision Trees	0.981	0.038
Random Forest	0.984	0.034
K-Nearest Neighbors	0.967	0.089
Support Vector Machines	0.987	0.052
XGBoost	0.990	0.028

exercises might produce significant biases. Regarding configurations 2 and 3, they show more balanced results. However, configuration 3 is inappropriate for analyzing dropout as none of the students dropped out. Furthermore, this configuration might overestimate students' knowledge and exercise are considered too easy. Consequently, the second configuration might be more balanced and also appropriate for dropout analyses. Nevertheless, other intermediate configurations (e.g., answering at least X exercises) could also be appropriate, and what is relevant is to understand the possible biases when selecting the subset of data so as to obtain the best possible subset for the analysis.

3.2. Analysis of model configurations, data-related factors and biases in prediction

In this section, several parameters that are relevant in prediction are analyzed to determine how much models might vary depending on the decision of the models.

The first analysis is on the algorithms. In order to do that, several algorithms have been used to predict dropout and final grades using the filtering of configuration 2. The results are presented in Table 2 and show little differences between the algorithms, although Regression and K-Nearest Neighbors show significant worse results, the differences between the rest are small, as suggested in [1]. For the next analyses, XGBoost is used as it achieves the highest predictive power.

The next analysis aims to explore if it is better using cumulative data from the beginning or data from only the last week (non-cumulative mode). To avoid data leakage, each prediction only uses data available up to that point in time, ensuring that models do not have access to future information. Fig. 5 shows the comparison between both modes (blue and orange lines). Results show that in both cases, the cumulative mode produces better results, and e.g., the difference in the last week is 0.128 in AUC and 0.298 in RMSE. Moreover, it is observed that data from certain week may not be the best. For example, data from the last week does not improve the models in the non-cumulative mode, probably because module 5 can be done in two weeks (weeks 6 and 7) and many students might have already finished.

Then, it is analyzed whether the student's ability calculated using the IRT (as presented in the previous subsection) could enhance the models, even with the possible biases this variable may have because of the sample. In this case, Fig. 5 shows a clear improvement of the models when adding this variable. For example, AUC in week 7 with ability is 0.991 and without ability is 0.972 (and 0.129 and 0.028 in RMSE, respectively). In addition, the anticipation of the models might be affected. For example, AUC is around 0.80 from week 3 with ability and week 4 without ability.

Finally, the effect of attrition bias and mastery attrition bias is analyzed. Attrition bias appears when there are less-skilled participants who drop the course [17] and for example, estimations of difficulty of the IRT are affected because only the most-skilled participants interact with the exercises. In this case, it may happen that many students drop out the course and they are still considered when training the models. In order to analyze this, when students are flagged as dropout, they are removed from training. For this analysis, students who do not take all the assessment but pass are not considered as dropout. Thus, 8,708 students are considered in week 1 following configuration 2, but this number is reduced each week. Fig. 6 shows the difference between considering this bias. In this case, only grade prediction is considered as students who dropout are removed. From this analysis, it can be observed that from week 2, the model that removes dropped students perform significantly better. This reflects

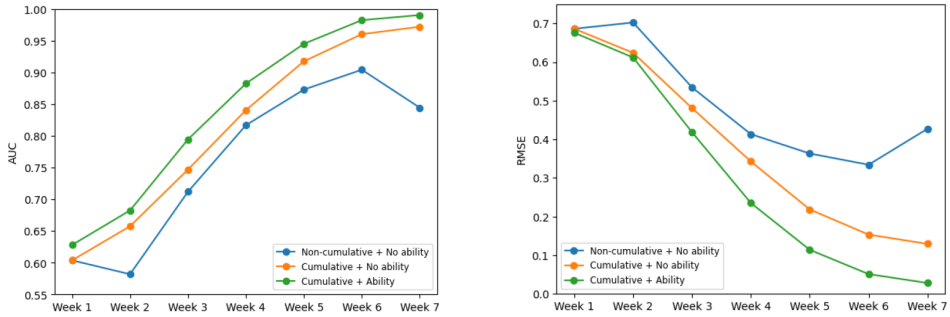

Figure 5: Comparison between prediction with and without ability and cumulative and non-cumulative mode

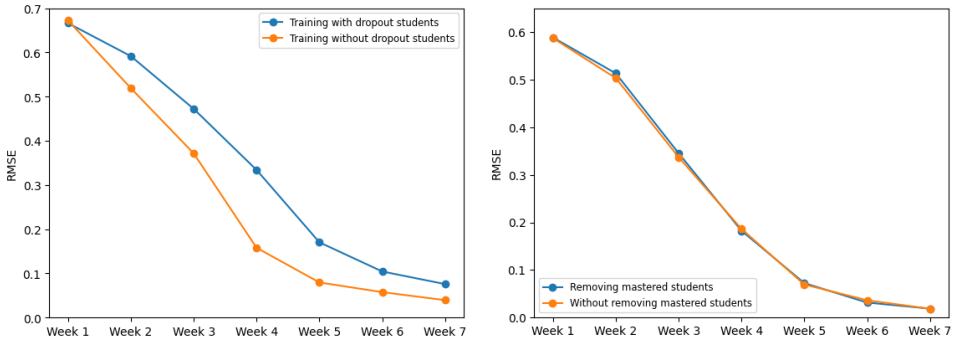

Figure 6: Final grades prediction: Attrition Bias (left) and Mastery Attrition Bias (right)

the fact that considering samples of students who do not interact can produce errors in the models and by removing them, the model can adapt better. Particularly, results improve from 0.075 to 0.039 in week 7 and the difference is above 0.15 in week 4.

Similarly, mastery attrition bias might be produced when students leave the course because they have mastered the contents [18]. In order to analyze this, students are removed as they were leaving because passed the course. A total of 402 students left the course in that condition. In this case, results in Fig. 6 show that mastery attrition bias has an insignificant effect in the MOOC. This is because it is not frequent that students who get 60% of the points (e.g., they get the maximum score in the first three tests and decide not to do the last two) do not engage with the last part of the MOOC, unlike the drop out situation that appears with attrition bias. Nevertheless, this effect might also depend on the situation and this bias might be more frequent in other contexts.

4. Conclusions

This paper has analyzed the effect of several aspects that may influence the estimations of students' abilities and the difficulty of questions using the IRT, and the predictive models. As for the IRT analysis, results showed that subsets of students which did not focus on graded activities (e.g., considering all students or those who interacted at least with formative activities, videos or forum) overestimated the difficulty and discrimination of exercises, and produced an imbalanced ability of students. In contrast, configurations that considered students who attempted one graded exercise at least or all the exercises were more balanced, although the latter case may underestimate the difficulty of exercises.

Regarding the predictive models, results showed that cumulating data from previous weeks is a better approach and the addition of the students' ability as predictor enhances the models even when this variable might suffer certain bias. Moreover, when analyzing well-known bias, it was shown that the effect of attrition bias is significant as many students drop out the MOOC, and removing students might produce significant biases (e.g., the difference of RMSE in week 4 was above 0.15). In contrast, the

mastery attrition bias was insignificant, probably because the condition for this bias was not frequent in this MOOC.

Despite having obtained these findings, there are some limitations that are worth mentioning. First, the analysis is based on only one MOOC; therefore, the generalizability of the findings is limited, and the conclusions should be framed cautiously. In addition, the grades were obtained based only on the quizzes and not the programming assignments, as their grades were not available. Furthermore, the set of configurations were defined so as to be representative, but other possibilities could have been chosen.

As future work, it would be relevant to expand the analyses to other courses with different methodologies so as to validate the results and overcome the current scope limitations. Moreover, it would be relevant to test the predictive models with different configurations, add more variables and include more advanced algorithms. Finally, it would be relevant to apply these models in real scenarios to validate their impact and how early prediction and estimation of abilities and difficulty of items could help instructors to support their students.

Acknowledgments

This work was supported by FEDER / Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación through the grant PID2023-146692OB-C31 (GENIE Learn project) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/UE. Moreover, it received support from the UNESCO Chair of "Scalable Digital Education for All" at UC3M and by the grant RED2022-134284-T funded by MICIU/AEI/10.13039/501100011033.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3 Flash in order to carry out grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Gardner, C. Brooks, Student success prediction in MOOCs, User Modeling and User-Adapted Interaction 28 (2018) 127–203.

[2] C. Olivares-Rodríguez, P. M. Moreno-Marcos, E. Scheiling García, P. J. Muñoz-Merino, C. Delgado Kloos, An actionable learning path-based model to predict and describe academic dropout, Ingeniería e Investigación 44 (2024) 14.

[3] E. Alyahyan, D. Düştegör, Predicting academic success in higher education: literature review and best practices, International journal of educational technology in higher education 17 (2020) 3.

[4] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, C. Delgado Kloos, Prediction in MOOCs: A review and future research directions, IEEE transactions on Learning Technologies 12 (2018) 384–401.

[5] A. A. Mubarak, H. Cao, S. A. Ahmed, Predictive learning analytics using deep learning model in MOOCs' courses videos, Education and Information Technologies 26 (2021) 371–392.

[6] Z. Ren, H. Rangwala, A. Johri, Predicting performance on MOOC assessments using multi-regression models, arXiv preprint arXiv:1605.02269 (2016).

[7] R. Cobos, V. M. Palla, edX-MAS: Model analyzer system, in: Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, 2017, pp. 1–7.

[8] A. Alamri, M. Alshehri, A. Cristea, F. D. Pereira, E. Oliveira, L. Shi, C. Stewart, Predicting MOOCs dropout using only two easily obtainable features from the first week's activities, in: International Conference on Intelligent Tutoring Systems, Springer, 2019, pp. 163–173.

[9] P. M. Moreno-Marcos, M. Sanz-Gómez, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, C. Delgado Kloos, Analysis of Students' Patterns and Prediction Using edX Data and Learning Analytics, in: 2025 IEEE Frontiers in Education Conference (FIE), IEEE, 2025, pp. 1–9.
[10] I. Šarić-Grgić, A. Grubišić, A. Gašpar, Twenty-five years of Bayesian knowledge tracing: a systematic review, User modeling and user-adapted interaction 34 (2024) 1127–1173.
[11] Y.-J. Lee, D. J. Palazzo, R. Warnakulasooriya, D. E. Pritchard, Measuring student learning with item response theory, Physical Review Special Topics—Physics Education Research 4 (2008) 010102.
[12] M. J. Cechak, Exploring Biases in Educational Data, Ph.D. thesis, Masaryk University, 2019.
[13] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, H. Gehlbach, Forecasting student achievement in MOOCs with natural language processing, in: Proceedings of the sixth international conference on learning analytics & knowledge, 2016, pp. 383–387.
[14] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, C. Delgado Kloos, Analysing the predictive power for anticipating assignment grades in a massive open online course, Behaviour & Information Technology 37 (2018) 1021–1036.
[15] Open edX Community, Open edX Tracking Logs Reference, 2024. URL: https://docs.openedx.org/en/latest/developers/references/internal_data_formats/tracking_logs/index.html, accessed: 2026-04-08.
[16] R. Pelánek, Metrics for Evaluation of Student Models, Journal of Educational Data Mining 7 (2015) 1–19.
[17] J. Papoušek, V. Stanislav, R. Pelánek, Evaluation of an adaptive practice system for learning geography facts, in: Proceedings of the sixth international conference on learning analytics & knowledge, 2016, pp. 134–142.
[18] R. Pelánek, J. Rihák, J. Papoušek, Impact of data collection on interpretation and evaluation of student models, in: Proceedings of the sixth international conference on learning analytics & knowledge, 2016, pp. 40–47.