Predicting the Outcomes of PBL: Comparison of Two Methods

Barbi Svetec, Blaženka Divjak, Damir Horvat, Katarina Pažur Aničić

University of Zagreb

Faculty of Organization and Informatics

Pavlinska 2, 42000 Varaždin, Croatia

{barbi.svetec,blazenka.divjak,damir.horvat1,
katarina.pazur}@foi.unizg.hr

***Abstract**. We used the Random Forest and Decision Trees algorithms to develop predictive learning analytics models for problem- and project-based learning (PBL). Predictive modeling was done on the data collected from two university courses, with a total sample of 309 students. Different phases of PBL were analyzed, with problem-solving found to be more predictable than problem-posing or peer-assessment. Students were divided into three classes based on their performance in PBL, and the RF-based models were found to be the most efficient in predicting the lower performing students. The DT-based models were generally found to be less efficient, but more interpretable.*

Keywords. predictive learning analytics, random forest, decision trees, problem-based learning, project-based learning, assessment

1 Introduction

Today, in the age of education highly supported and enhanced by technologies, higher education institutions (HEIs) have access to more ample and more diverse student data than ever before (Prinsloo, 2020). Exploiting this opportunity, predictive learning analytics (LA) can contribute to the quality of education by using historical data to make predictions and inferences about possible future outcomes (Susnjak et al., 2022).

Problem- and project-based learning (PBL), in which students work to solve real-life problems, have been recognized as innovative approaches that foster critical thinking, creativity, autonomy, collaboration and interdisciplinarity (Brassler & Dettmers, 2017; Dole et al., 2015; Savery, 2006). Nevertheless, educators are still challenged by course structuring, student progress monitoring, and giving guidance. LA provides a promising new perspective in terms of course design and monitoring. (Wang et al., 2023)

Some studies have been conducted at the intersection of PBL and LA. For example, a study investigating the key factors predicting online PBL performance (Wang et al., 2023) identified self-regulation, posted messages, message words, and peer-learning engagement as predictors. However, the overall body of research combining PBL and LA to support the development of data-driven student-centered courses, and especially linking LA with particular steps of PBL, is still relatively scarce (Wang et al., 2023).

The aim of this study was to investigate the predictive power of parts of the assessment program on PBL performance, and to analyze the efficiency of predictive models based on two prediction algorithms (Random Forest and Decision Trees).

2 Theoretical Background

2.1 Predictive Learning Analytics

An important area in LA refers to forecasting academic outcomes based on patterns in past and present data. This *predictive* LA strongly relies on machine learning and deep learning algorithms, learning from historic datasets and using various kinds of student data to make predictions. (Sghir et al., 2023; Susnjak et al., 2022) Predictive LA can provide important support to achieving LOs, informing learning design, identifying at-risk students, as well as increasing students' satisfaction (Sghir et al., 2023). Predictive modeling, as an important practice in LA, is done with the use of several algorithms: in particular, the most often used Artificial Neural Networks are followed by the Random Forest (RF) and Gradient Boosting algorithms, and these algorithms were also found to have the highest prediction accuracy compared to other algorithms (Sghir et al., 2023). Particularly, an educational research study (Kabathova & Drlik, 2021) found that the RF classifier was associated with the best accuracy and precision compared to Logistic Regression, Support Vector Machine, Decision Tree (DT), Neural Networks and Naïve Bayes. The DT algorithm was among the algorithms that came out second best.

Some research has been conducted to identify the most important predictors. Importantly, it has been found that cognitive data (entry test and quiz scores) present the best predictors, as opposed to basic LMS data (Tempelaar et al., 2015). Research has also shown that formative assessment is an important predictor of students' performance (Bulut et al., 2023; Divjak et al., 2024).

2.2 Problem- and Project-Based Learning

Problem- and project-based learning are student-centered, constructivist approaches to teaching and learning (Dole et al., 2015). Both are based on activities done by students collaboratively, independently, in order to achieve a mutual goal (Brassler & Dettmers, 2017; Savery, 2006). Problem-solving is usually done in teams, and this peer-learning is followed by peer-assessment. While in problem-based learning students investigate to solve ill-structured problems, not necessarily leading to a concrete artifact (Savery, 2006), in project-based learning, students' problem-solving is oriented towards the production of a final artifact (Dole et al., 2015). Moreover, while guidance is important in both approaches, in project-based learning, teachers provide more direct instructions to students on how to produce the final artifact, but still fostering self-regulated learning (Savery, 2006). In the problem-solving context, we distinguish between problem-posing and problem-solving, with the former found to be more demanding for students, as it is something they are less exposed to (Divjak, 2015). Problem-posing generally refers to students generating new problems, which can be based on given situations or existing problems (Cai & Rott, 2024).

For the aim of this paper, the two approaches will be referred to in conjunction (using the abbreviation "PBL"), as the aspects relevant for this research are common to both.

3 Methodology

The focus of this study has been on two research questions:

1. In these PBL case studies, how do different prediction methods compare in terms efficiency and explainability?
2. What elements of an assessment program are the best predictors of PBL performance?

3.1 Study Setting and Data Collection

The study was conducted at the University of Zagreb, Faculty of Organization and Informatics, a HEI offering study programs in ICT, and included data from two courses, collected in the academic year 2022/2023. The courses were chosen so as to represent two different levels of study (undergraduate and graduate) and two different subject areas (ICT and Mathematics), include a considerable PBL experience for students, as well as to have a number of students large enough to enable meaningful machine learning data analysis.

At the undergraduate level, the study included the Informatics Services Management (ISM) course (second year). The course has a student workload of 4 ECTS credits. The assessment program includes formative assessment (assignments done in laboratory exercises) and summative assessment (two periodical exams, a PBL assignment in the middle of the semester and at the end). In the PBL assignment, students work in groups on projects given by the course teacher, based on client requirements, focused on developing a prototype of an IT service. In the first phase, students have the task to determine the characteristics of end-users and make a service proposal in the form of a wireframe (a type of problem-posing). The first version of the student solution is presented to the teacher in the middle of the semester, in the form of a business meeting, in which teachers have the role of potential clients. In the second part of the semester students work on the solution (service prototype as a type of problem-solving) according to the feedback received in the first business meeting. The PBL assignment contributes 50% to the final grade. The sample included in this study consisted of 191 students.

At the graduate level, the study included the Discrete Mathematics with Graph Theory (DMGT) course (first year). The course has a student workload of 6 ECTS credits. The course's assessment program comprises formative assessment (weekly quizzes) and summative assessment (two periodical exams, a PBL assignment). The extensive PBL assignment includes five weeks of independent work done by students in teams, and has three stages. In the first stage, students find and describe a real-world situation in which graph theory and discrete mathematics can be applied to solve a problem, and submit a project description with the characteristics of solutions (problem-posing). Feedback and grades are provided by course teachers, and these project proposals are distributed to other teams to solve. In the second stage, students work on projects, solving the problems posed by other teams (problem-solving). Finally, each team assesses the solutions to their project proposals (peer-assessment), together with the course teachers. The final grades consist of aggregated students' peer-assessment grades and teachers' grades (which have a higher weight). Additionally, students also receive a grade for the quality of work on peer-assessment, based on its consistency with the teacher's grade. The PBL assignment contributes 40% to the final grade. The sample included in this study consisted of 118 students.

For both courses, we collected assessment data from the Moodle LMS. The data were anonymized and students' identities protected.

3.2 Data Analysis

For each of the courses, the assessment data were randomly distributed, so that 75% of the data were used for training, and 25% were used for testing.

The data were analyzed using two supervised machine learning algorithms used for classification and regression purposes: Random Forest (RF) and Decision Trees (DTs). The analyses were performed in R, using the *tidymodels* package with *rpart* engine.

DTs are a simple method, but successful in predicting and explaining relationships (Rokach & Maimon, 2014). A DT is a mathematical graph, starting with a single node, with edges branching into next-level nodes based on the probability of a certain outcome. Each node refers to a possible decision, while edges represent their consequences. It consists of repeated tests on the input variable, with the results of each test determining the next test, until the result of the function is certain. DTs can be prone to overfitting if the number of nodes is high relative to the quantity of data. A higher number of nodes decreases the training error, but increases the generalization error. (Rokach & Maimon, 2014)

RF is an ensemble method, more complex, based on multiple DTs, combining their results for final classification. While building DTs, the RF randomly selects various subsets of attributes. RF presents "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Breiman, 2001). Due to the Law of Large Numbers, with RFs, there is no problem of overfitting. Considering all this, RFs have been found effective in making predictions. (Breiman, 2001; Kursa & Rudnicki, 2010) Another advantage of RFs is that they offer the possibility of variable importance measures (VIMs), used for identification of relevant features or selection of variables, including impurity and permutation importance (Nembrini et al., 2018). Regardless of the advantages, the random nature of RFs means they are "not always intuitive and comprehensible", as different trees can lead to inconsistency in interpretations. In this sense, the advantage of single DTs is their interpretability. (Rokach & Maimon, 2014)

We started the analysis with cross-validation to choose the optimal combination of hyperparameters. For RF, we tested 500 randomly chosen combinations of hyperparameters (*mtry, min_n, trees*), and for DTs, we tested 5000 randomly chosen combinations of hyperparameters (*tree_depth, min_n, cost_complexity*) with Latin hypercube sampling. Then, we performed RF and made single DTs on the training datasets, and tested the efficiency of each algorithm on the testing datasets.

For each of the two courses, we built an RF-based and a DT-based model for each phase of the PBL assignment. We tested the efficiency of each of the models based on several metrics, and most importantly using the area under the Receiver Operating Characteristics curve (ROC_AUC), presenting a plot of the true positive vs. the false positive rate. Although the ROC curve is normally used for assessing the performance of binary classification models, its use also extends to multi-class classification (Hand & Till, 2001; Mandrekar, 2010).

In each model, students were divided into three classes, based on their assessment results (0 - 33%, 33 - 67%, 67 - 100% of points obtained in the respective part of the PBL assignment).

We analyzed the confusion matrices for each of the models to assess the models' performance.

Furthermore, we analyzed the importance of predictors, based on three VIMs: the Gini index, permutations and the Boruta extension (Kursa & Rudnicki, 2010).

4 Results

4.1 Efficiency of the Models

The efficiency of the models was tested, most importantly, based on the ROC_AUC value. The values for both courses and both algorithms are presented in tables 1 and 2.

If we look at the training datasets, the RF results indicate a higher discriminatory power than the DT results. Namely, for both courses, the ROC_AUC values in the RF-based models are generally (around or) above 0.9, which is considered outstanding (Hosmer & Lemeshow, 2000). Looking at the DT-based models, the performance is somewhat lower, with ROC_AUC values ranging from (around) 0.7, which is considered acceptable (Hosmer & Lemeshow, 2000), to (above) 0.9. Generally, the models' performance on the training datasets indicates that patterns have been learnt well, but with the RF-based models outperforming the DT-based models.

If we consider the testing datasets, the models turned out to be less efficient, with ROC_AUC values often below 0.7 (less than acceptable). However, looking at the models for problem-solving in both DMGT and ISM, the ROC_AUC values around or above 0.8 indicate excellent efficiency (Hosmer & Lemeshow, 2000) of both RF-based and DT-based models, with the RF-based again performing slightly better. In the case of peer-assessment, this applies only to the DT-based DMGT course model.

The confusion matrices were analyzed for all the models, and here we present those related to problem-solving in DMGT (Fig. 1) and ISM (Fig. 2), RF-based models, as the most efficient ones.

Table 1. DMGT course: ROC_AUC values

Random Forest						Decision Trees					
Problem-posing		Problem-solving		Peer-assessment		Problem-posing		Problem-solving		Peer-assessment	
Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.850	0.612	0.955	0.826	0.996	0.687	0.689	0.640	0.883	0.795	0.939	0.784

Table 2. ISM course: ROC_AUC values

Random Forest						Decision Trees					
Problem-posing		Problem-solving		Peer-assessment		Problem-posing		Problem-solving		Peer-assessment	
Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.875	0.561	0.999	0.895	1.00	0.685	0.733	0.576	0.882	0.835	0.700	0.548

The DMGT course confusion matrix in Figure 1 suggests that the model is the most efficient in classifying class 1 students (lower-performing), with 9 students classified correctly, although additional 3 class 1 students were misclassified as class 3. There was some more confusion when it came to the neighboring classes 2 and 3, as 7 students were classified correctly in each of the two classes, but 4 were misclassified as belonging to the neighboring class.

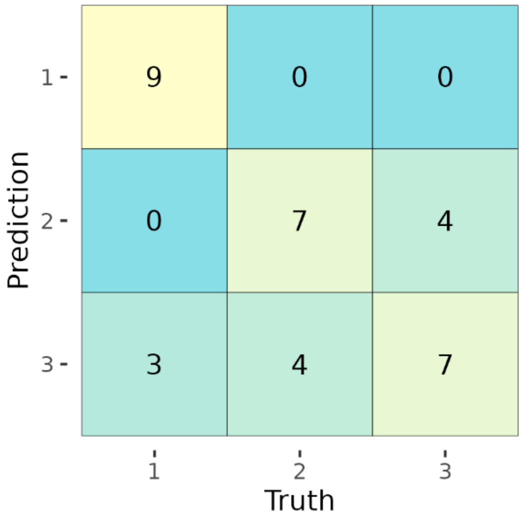

Figure 1. DMGT course: RF confusion matrix (problem-solving)

The ISM confusion matrix in Figure 2 shows the model performed best in predicting the (lower-performing) class 1, with 13 students classified correctly, but it also misclassified 3 class 1 students as the neighboring class 2. Similarly, it correctly classified 13 students as class 2, although it also misclassified 4 respective students as class 1. The model was the least successful with class 3, with 10 students classified correctly, but additional 5 students misclassified as either class 1 or 2.

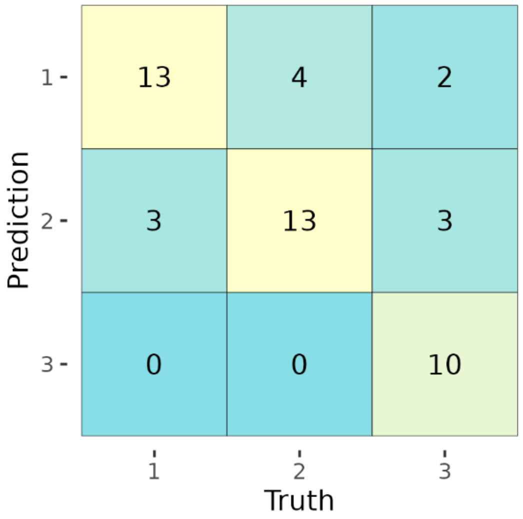

Figure 2. ISM course: RF confusion matrix (problem-solving)

4.2 Predictors

The importance of predictors was analyzed for all the models (phases of PBL). Here we present those related to *problem-solving* in DMGT and ISM, as the most efficient models.

DMGT. In the RF-based model, results in *problem-posing* turned out to be the most important predictor of the results in *problem-solving*, according to all the three VIMs (Gini, permutations, Boruta). *Problem-posing* was followed (from the most important) by:

exam 2, exam 1, quizzes (Gini, permutations) or *exam 1, quizzes, exam 2* (Boruta).

The DT-based model gave a similar picture, with *problem-posing* coming out as the most important predictor, followed by *exam 1, exam 2*, and, lastly, *quizzes*.

Looking at all the results, what comes out as the best predictor is generally *problem-posing*, followed by *exams*, and, finally, *quizzes*.

The DT (Fig. 3) gives a visual representation of criteria-based decisions about student classification. For example, students with less than 4.5 points in *problem-posing* are likely to end up in class 1. If their result in *problem-posing* was still equal to or above 1 point, they can end up in either class 1 or 2 depending on their *quiz* results. Interestingly, if their *problem-posing* results were below 1 point, they could still end up in class 3, provided that their *exam 1* results were above 16.3 points. Other paths through the DT can be interpreted in a similar way. Looking at the leaves of the DT, we can interpret the structure of the three classes.

ISM. In the RF-based model, similarly to DMGT, *problem-posing* was found to be the most significant predictor according to the tree VIMs. The order of

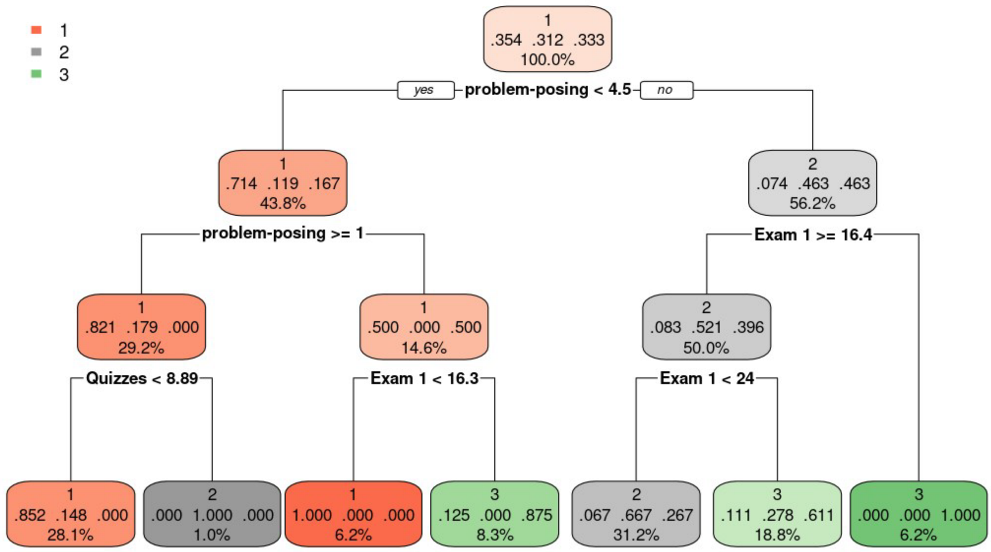

Figure 3. DMGT course: decision tree

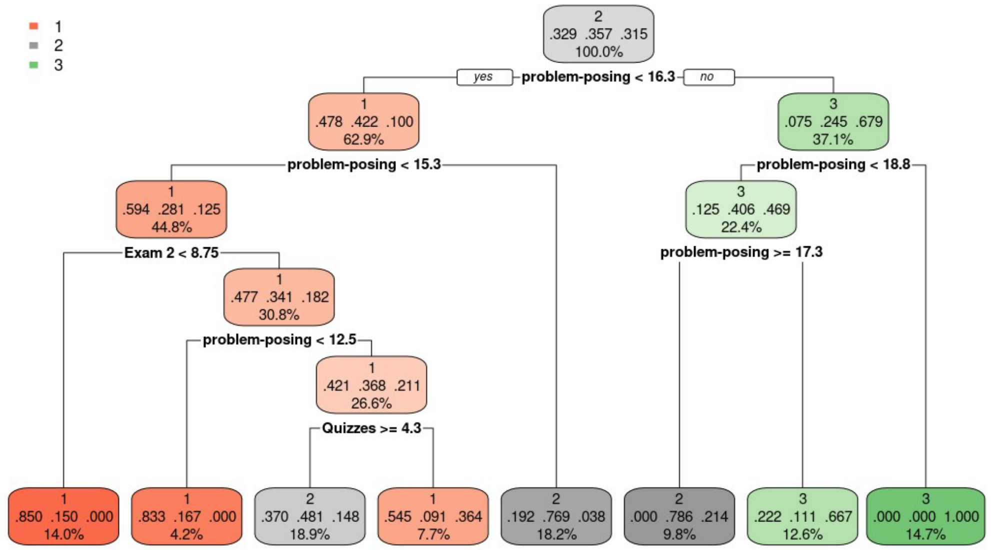

Figure 4. ISM course: decision tree

other predictors varied depending on the VIM, as follows (from the most important): *quizzes, exam 2, exam 1* (Gini); *exam 1, quizzes, exam 2* (permutation); *quizzes, exam 1, exam 2* (Boruta).

The DT-based model confirmed *problem-posing* as the most important predictor, followed by *exam 1, quizzes,* and *exam 2*.

Overall, the results undoubtedly point to *problem-posing* as the most important predictor, followed by *quizzes* and *exam 1*, with *exam 2* seeming the least important.

The DT (Fig. 4) shows that students with less than 16.3 points in *problem-posing* are likely to end up in class 1. If their result in *problem-posing* is above 15.3 (and still less than 16.3) they would end up in class 2, regardless of their results in other activities. If the student's result in *problem-posing* is above 16.3, they are most likely to end up in class 3. Only in the case when the result in *problem-posing* is above or equal to 17.3., students would end up in class 2.

5 Discussion

The comparison of prediction models based on the two algorithms, expectedly, suggests a higher efficiency of the RF compared to the DT. However, when it comes to problem-solving, both algorithms gave acceptable results. As for the other two phases of PBL, namely problem-posing and peer-assessment, the lower efficiency of the models may be related to the nature of these two PBL phases. For example, problem-posing is a type of activity in which students are generally less experienced, and which differs in terms of required competences from other parts of the assessment program. In fact, some research has pointed out that problem-posing is generally harder for students (Divjak, 2015). For example, identifying project requirements in real-life situations does not necessarily require the same competences as solving exams. Therefore, it makes sense that exams might not be good predictors of performance in problem-posing. On the other hand, this is not the case for problem-solving, as in this phase students necessarily use techniques, theory and models covered by exams. Moreover, problem-solving builds on problem-posing, so it is expected that problem-posing is the most important predictor. Although the potential of problem-posing in enhancing learning and understanding of concepts is clear (Cai, 2022), problem-posing, as a less researched part of PBL, remains a mystery. In that respect, our research is in line with previous studies (Cai & Rott, 2024). Furthermore, when it comes to peer-assessment, the better efficiency of the predictive models related to the graduate-level course may suggest that graduate-level students' performance in peer-assessment is more consistent with other parts of the assessment program. This may also be related to the fact that in the DMGT course, students peer-assessed the projects solving the problems they had posed themselves, allowing them to better understand the problem and feel more ownership.

As for other predictors, their importance may depend on the level of LOs and the learning context. For example, in the DMGT course, problem-solving tasks require students to use the algorithms and approaches which are covered by exams, but are more complex than those included in quizzes. Therefore, exams are a better predictor of PBL than quizzes. On the contrary, in the ISM course quizzes are more related to PBL, which is reflected in the importance of the predictors. Previous research (Divjak et al., 2024) showed that both formative and summative assessment are important for the prediction of final summative results. Here we showed that similar applies to PBL.

If we look at the confusion matrices, we notice that the models perform best when it comes to lower-performing (class 1) students. This has an important implication for identification and advising of students at risk from failing the course.

Moreover, the confusion among higher classes can be explained if we consider the nature of PBL, which includes teamwork. Namely, the data reflect the grades obtained by the students in teams. However, this does not reflect the exact contribution of each of the students within a team, whose performance in other parts of the assessment program might differ significantly. This, however, does not mean that teamwork is not valuable for peer-learning.

Furthermore, when it comes to explainability, DTs offer clear visual representations of how decisions are made in classification. However, DTs should be interpreted carefully, as they do not consider the timeline of activities, and do not present activities which are not nodes (vortices) of a given DT. In that sense, it is advisable to combine the DT results with other representations of students' progression, like the braided graph (Sankey diagram), which offer insights in the fluctuations of students between classes.

In terms of practical implications, the tested predictive models could be used by practitioners to inform learning design and innovate assessment.

6 Limitations and further research

The main limitation of this study is a relatively small sample and a limited number of courses. Data collection through several years and in several courses is needed to obtain more comprehensive and conclusive results. Furthermore, the use of other prediction methods and algorithms, like Artificial Neural Networks, Gradient Boost or Support Vector Machine, may elicit different results, which could shed a different light. Finally, including digital trace data, as well as multimodal data, in LA, is important in order to build predictive LA models which could be used as the basis for recommendation systems.

7 Conclusion

This paper contributes to the growing body of research on the use of learning analytics to understand and enhance problem- and project-based learning (PBL). We presented predictive analytics models based on the Random Forest (RF) and Decision Tree (DT) algorithms. We confirmed that while RF enables better efficiency, DTs are more interpretable. The models provided the best predictions for the lower performing students in PBL. Importantly, the study found that students' performance in problem-posing is an important predictor of problem-solving. Other predictors from the assessment program, including both formative and summative assessment, can be used as predictors, but fine-tuned according to the learning context. Finally, problem-posing cannot be easily predicted based on formative and summative assessment results.

Acknowledgments

This work has been supported in part by the Croatian Science Foundation under the project IP-2022-10-2854 (*Trustworthy Learning Analytics and Artificial Intelligence for Sound Learning Design* – TRUELA).

References

Brassler, M., & Dettmers, J. (2017). How to Enhance Interdisciplinary Competence—Interdisciplinary Problem-Based Learning versus Interdisciplinary Project-Based Learning. *Interdisciplinary Journal of Problem-Based Learning*, *11*(2). https://doi.org/10.7771/1541-5015.1686

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., Lai, K. W., & Shin, J. (2023). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology*, *54*(1), 19–39. https://doi.org/10.1111/bjet.13276

Cai, J. (2022). What Research Says About Teaching Mathematics Through Problem Posing. *Éducation et Didactique*, *16*, 31–50. https://doi.org/10.4000/educationdidactique.10642

Cai, J., & Rott, B. (2024). On understanding mathematical problem-posing processes. *ZDM – Mathematics Education*, *56*(1), 61–71. https://doi.org/10.1007/s11858-023-01536-w

Divjak, B. (2015). Assessment of complex, non-structured mathematical problems. *IMA International Conference on Barriers and Enablers to Learning Maths: Enhancing Learning and Teaching for All Learners*.

Divjak, B., Svetec, B., & Horvat, D. (2024). How can valid and reliable automatic formative assessment predict the acquisition of learning outcomes? *Journal of Computer Assisted Learning*, *January*, 1–17. https://doi.org/10.1111/jcal.12953

Dole, S., Bloom, L., & Kowalske, K. (2015). Transforming Pedagogy: Changing Perspectives from Teacher-Centered to Learner-Centered. *Interdisciplinary Journal of Problem-Based Learning*, *10*(1). https://doi.org/10.7771/1541-5015.1538

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, *45*, 171-186Agasisti, T., Bowers, A. J. (2017). Dat.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley. https://doi.org/10.1002/0471722146

Kabathova, J., & Drlik, M. (2021). Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Applied Sciences*, *11*(7), 3130. https://doi.org/10.3390/app11073130

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11). https://doi.org/10.18637/jss.v036.i11

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, *34*(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Prinsloo, P. (2020). Of 'black boxes' and algorithmic decision-making in (higher) education – A commentary. *Big Data & Society*, *7*(1), 205395172093399. https://doi.org/10.1177/2053951720933994

Rokach, L., & Maimon, O. (2014). *Data Mining with Decision Trees* (Vol. 81). WORLD SCIENTIFIC. https://doi.org/10.1142/9097

Savery, J. R. (2006). Overview of Problem-based Learning: Definitions and Distinctions. *Interdisciplinary Journal of Problem-Based Learning*, *1*(1). https://doi.org/10.7771/1541-5015.1002

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, *28*(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Susnjak, T., Ramaswami, G. S., & Mathrani, A. (2022). Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, *19*(1), 12. https://doi.org/10.1186/s41239-021-00313-7

Tempelaar, D., Rienties, B., & Giesbers, B. (2015). Stability and Sensitivity of Learning Analytics based Prediction Models. *Proceedings of the 7th International Conference on Computer Supported Education*, 156–166. https://doi.org/10.5220/0005497001560166

Wang, X., Sun, D., Cheng, G., & Luo, H. (2023). Key factors predicting problem-based learning in online environments: Evidence from multimodal learning analytics. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1080294