

PREDICTING ACADEMIC PERFORMANCE FROM BEHAVIOURAL AND LEARNING DATA

CARLOS VILLAGR -ARNEDO, FRANCISCO J. GALLEGRO-DUR N, PATRICIA COMPA N-ROSIQUE,
FARA N LLORENS-LARGO & RAFAEL MOLINA-CARMONA
Department of Computer Science and Artificial Intelligence, University of Alicante, Spain.

ABSTRACT

The volume and quality of data, but also their relevance, are crucial when performing data analysis. In this paper, a study of the influence of different types of data is presented, particularly in the context of educational data obtained from Learning Management Systems (LMSs). These systems provide a large amount of data from the student activity but they usually do not describe the results of the learning process, i.e., they describe the behaviour but not the learning results. The starting hypothesis states that complementing behavioural data with other more relevant data (regarding learning outcomes) can lead to a better analysis of the learning process, that is, in particular it is possible to early predict the student final performance. A learning platform has been specially developed to collect data not just from the usage but also related to the way students learn and progress in training activities. Data of both types are used to build a progressive predictive system for helping in the learning process. This model is based on a classifier that uses the Support Vector Machine technique. As a result, the system obtains a weekly classification of each student as the probability of belonging to one of three classes: high, medium and low performance. The results show that, supplementing behavioural data with learning data allows us to obtain better predictions about the results of the students in a learning system. Moreover, it can be deduced that the use of heterogeneous data enriches the final performance of the prediction algorithms.

Keywords: behavioural data, learning analytics, learning data, prediction.

1 INTRODUCTION

When working with data, it is critical not just the volume and quality of data but also their relevance, that is, the ability of such data to provide information about the characteristics that are wanted to be studied. The problem is that often these data are not available and we use instead less relevant but more available data.

Within the world of education, Learning Management Systems (LMSs) provide a large amount of data from the student activity. However, these data reflect the use of the platform but they do not describe the results of the learning process of students, that is, they describe the behaviour but not the learning results.

A study of the learning progress of the students when interacting with a learning platform has been carried out. The learning platform has specially be developed to collect data not just from the use but also related to the way students learn and progress in training activities. A predictive system has also been included in the platform to early predict students' final performance. This system is based on a classifier built using the Support Vector Machine technique.

The objective of this paper is demonstrating that complementing behavioural data with other more relevant data (regarding learning outcomes) can lead to better analysis of the learning process. Specifically, two research questions are posed: which is the best configuration for the data set to obtain a better predictor of the students' performance? And, how accurate is the resulting predictive model?



This paper is part of the Proceedings of the International Conference on Big Data
(Big Data 2016)
www.witconferences.com

  2016 WIT Press, www.witpress.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), <http://www.witpress.com/journals>

DOI: 10.2495/DNE-V11-N3-239-249

This paper is organized as follows: section 2 presents the background of the proposal; section 3 describes the learning platform that allows the data collection; section 4 presents four different experiments and their results; and section 4 presents the conclusions.

2 BACKGROUND

In the era of computers and the Internet, the amount of data generated by technological systems is huge. Basing decisions on these data and building evidences from them seems logical. Data-driven decision-making is acquiring a remarkable importance in many fields, that has led to new disciplines like business intelligence, health care analytics, social media analytics, and so on [1]. In this context, a new discipline has arisen: Learning Analytics. According to the 1st International Conference on Learning Analytics and Knowledge [2], learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.

Some researchers have argued that Learning Analytics can help identify at-risk learners and provide intervention to assist learners in achieving success [3]. Kotsiantis [4] made an interesting review about Learning Analytics and prediction. An interesting example is found in [5], where the authors compare four Machine Learning models for predicting student performance. The best predicting results for individual students are obtained using a Support Vector Machine (SVM). Also, Lykourantzou *et al.* [6] achieve an accurate prediction at an early stage of an e-learning course using feed-forward neural networks. However, the proposed method failed to predict precisely the performance of certain specific students. Other authors try to predict student or group performance [7–11].

Some authors have used learning indicators (for instance, Standard Admission Test (SAT) exams scores) while others prefer to use behavioural LMS data (such a LMS logins) to predict. For instance, Schalk *et al.* [9] built a Machine-Learning-based predictive system to determine which students were at risk of failing introductory courses in mathematics and physics. The system used Random Forests technique to model data coming from previous years of SAT. While their results are good, the method designed was neither thought to be maintained over time, nor to do progressive predictions based on incremental information. Wang & Newlin [12, 13] proposed the use of data from a web-based LMS about student online activity and provided an early indicator of student academic performance. Other works [3] conclude that student tracking data obtained from LMS can provide with pedagogically meaningful information about engagement and likelihood of success. Several authors [14, 15] have concluded that LMS usage patterns and student performance are related, and the combination of SAT scores and LMS logins has an important predictive power. They state that the best results are obtained when data about learning results (such as the SAT scores) are combined with data about LMS usage.

Behavioural data captured from LMSs are well structured and easy to be obtained. Nevertheless, learning is a very complex human activity, which cannot be simplified. Long & Siemens [1] consider that there is a risk to return to behaviourism as a learning theory if we confine analytics to behavioural data. The current challenge of Learning Analytics is introducing not just behavioural information, but also information about interaction between learners and other participants, attitudes, skills and learning results, that is, about the learning process itself.

3 LEARNING PLATFORM

Most LMSs only provide behavioural data about the use of the platform by the students. Since the objective of this work is studying the potential of learning data to explain and predict the students' learning outcomes, a new platform with special features has been developed.

The learning platform is based on an interactive web where different exercises are delivered to the students [16]. The students can upload the solutions of the exercises to the system, so that they are

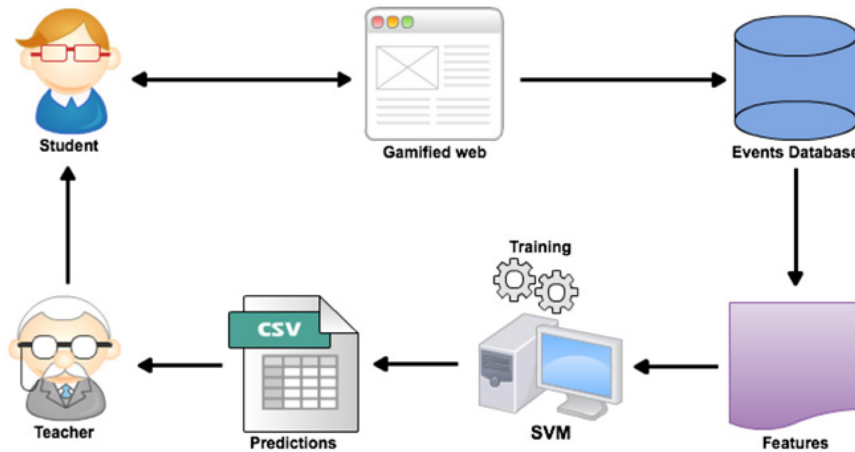


Figure 1: Elements of the proposed platform.

automatically marked. The students get immediate feedback of their results and can redo the exercises to get a better mark. During this process, a large amount of data is collected both from the student interaction (usage or behavioural data) and from the student progression in the exercise resolution (learning data). All these data are stored in a database to be used in further analysis. Specifically, there is a selection of the relevant data (the meaningful characteristics are extracted), and they are introduced in a predictive system based on a SVM to obtain a prediction of the student performance at the end of the term. Finally, the predictions are offered to the teachers in the form of graphs that present a progressive view of performance predictions, producing valuable information about the evolution of students learning activity and expected performance. These elements make up the architecture of the learning platform that is illustrated in Fig. 1. In the following sections, a deeper explanation of the most important modules is presented.

3.1 Interactive web and exercises

All the learning process is performed using a web site that allows the interaction of students and teachers. Other than habitual user management, the web provides several available functions for the students (download exercise formulation, upload exercise solution, obtaining exercise mark, access their learning progress...) and for the teachers (upload new exercise formulations, monitor the progress of their students...).

The core elements in the learning system are the exercises. Although the exercise type depends on the subject, the system can be adapted to manage any type of computational problem that can be automatically marked (e.g. programming problems or mathematical questions). Specifically, the system in its current version deals with logics problems, consisting on the solution of mazes of a game called PLMan [16]. In this game, students program the artificial intelligence of a Pac-Man like character, using a Prolog knowledge base. For each maze, the students have to program a set of rules so that the character can get all the dots, dodging the perils. In Fig. 2, a typical maze of PLMan and its solution as a set of Prolog rules is presented.

Mazes in PLMan are organized into six main stages (stages from 0 to 4 and a final checkpoint) and up to 5 levels of difficulty per stage. Students have to beat the 6 stages to get the maximum grade. At



Figure 2: PLMan maze and its solution.

each stage, students have to solve 1 to 5 different mazes (5 mazes at stage 0; 3 mazes at stages 1 and 2; 2 mazes at stage 3; 1 maze at stages 4 and checkpoint). To get each new maze for solving, students start by pick up their desired difficulty level among valid levels for the stage in which they are. Then, the system presents them with a random maze from their selected level of difficulty. Mazes can only be assigned once, so the same maze cannot be assigned to two different students.

Every possible solution can be delivered to the web system. The automated system runs the solution and evaluates its score and mark, mainly based on the total proportion of dots their PLMan achieves to eat.

3.2 Data collection and feature extraction

While designing, building, redesigning and maintaining a system like this, it becomes clear that the flow of data that go in and out the system is invaluable from the teaching/learning point of view. It is crucial to transform this data into valuable information that can give advice for teachers about progress of the students. Building some informative characteristics well correlated with student performance and evolution is crucial to get good prediction results, so the first step is designing which information would be valuable to log. Although the meaningful data depend on the specific problem, we consider that this feature set must consist of both behavioural data (data related to the system usage) and learning outcomes (data related to the exercise results).

In the particular case presented in this paper, a list of concrete events that occur during the interaction between students, and the system was considered. All these events (see Table 1) are logged in an event database, with their appropriate timestamp and related information. They made up the disposable behavioural data. Moreover, to make up the set of learning data, all learning results from these events where also considered: submission results, submission attempts at each maze, mazes solved (percentage obtained > 75%), time to solve each maze, etc. These learning outcomes are qualified as potential predictors because of their intuitively logical relationship with the final performance.

All this information about the student activity is the primary matter to construct the set of features that make up the input of the predictive system, based on a Machine Learning algorithm. In order to analyse the effect of behavioural and learning data in the prediction system, four experiments have been designed. In Section 4, a complete description of the experiments is presented, so that some conclusions about the best data to use can be extracted.

Table 1: Events stored in the database.

Event	The student ...
show_frontpage	enters in the system and sees the main status page
show_results	sees their results on a concrete maze
select_difficulty	picks up the difficulty they want for a new maze
maze_download	downloads a maze
solution_submission_ok	submits a solution which compiles and executes ok
download_logs	downloads execution logs (used to replicate AI bugs)
solution_submission_error	submits a solution that does not work properly

3.3 Predictive system

The function of the predictive system is to make predictions of the students' marks at the end of the term, based on the data collected from the interactive web. These data, selected and organized into features, are normalized and given to a machine learning algorithm as input.

The predictive system classifies the student expected performance (measured as a mark in percentage) in three possible classes: high performance (expected mark $> 80.5\%$), medium performance ($57.5\% < \text{expected mark} < 80.5\%$) or low performance (expected mark $< 57.5\%$). The reason to split the output in three classes is to get an adequate performance out of the classification algorithm, adjusted to the size of the sample (the sample has only 336 students, so three equilibrated classes of 112 individuals are proposed). The balanced distribution of the students in the mark range explains the selected intervals for each class.

The prediction system is based on a standard C-parameterized margin, SVM classifier (C-SVC) [17] with Radial Basis Function (RBF) kernel [18, 19], adding probability estimates using Pairwise Coupling [20]. This is a very effective and efficient Machine Learning algorithm that works very well for general datasets like the one under analysis, as Huang & Fang [5] pointed out.

4 EXPERIMENTS

4.1 Experiment design

The proposed system has been implemented in the past term for a first year university subject about Computational Logic and Prolog programming. There were around 400 students enrolled, 336 of which finally participated in the experiments. Although the complete term consists of 15 weeks, only 11 weekly lessons are usable since the first 4 lessons are introductory. Consequently, 10 weeks of predictions are given, and then the system closes giving final marks to students in week 11.

For each week, all the information in the event database described in Section 3.2 is processed to get up-to-date features describing students. A specific set of features is selected for every experiment we present in the following sections but, in general, the features are scaled to $[0,1]$ range before passing them as a vector to the SVM, preserving the many values that are already 0 and maintaining matrix scarcity. So, the SVM gets this feature vector as input, and a vector of 3 probabilities as output, which effectively gives the probability of a student being classified in each one of the 3 classes (high, medium or low performance).

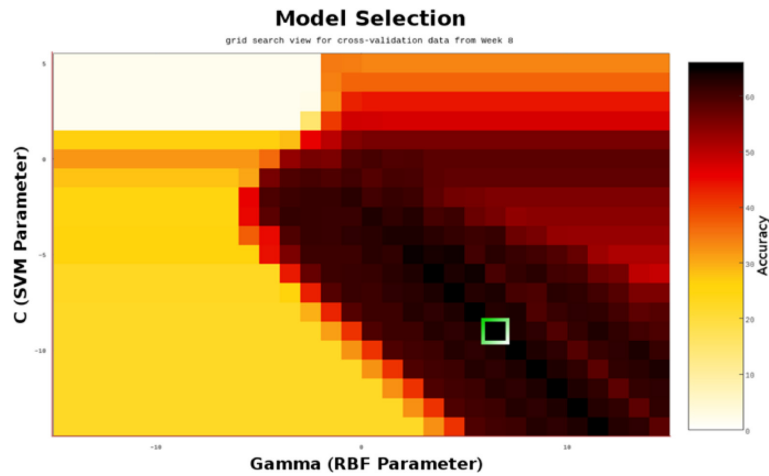


Figure 3: Model selection using 2D grid-search.

With the scaled features, a model selection is conducted to boost the performance of the SVM by picking up the best set of parameters (see Fig. 3). Since two parameters have to be selected (C regularization parameter, and γ amplitude of RBF), a 2D grid-search is performed, with parameter values varying exponentially in powers of 2:

$$C \in \{2^i : i \in [-15, 15] \cap \mathbb{Z}\}, \gamma \in \{2^j : j \in [-15, 5] \cap \mathbb{Z}\}$$

For each pair of parameters, a leave-one-out Cross-Validation training gives its accuracy estimation. Finally, the two parameters corresponding to the best accuracy estimation are selected.

After selecting the best set of parameters, 10 SVMs are trained, one for each week being considered. This improves expected performance by having specialized SVMs for each week. It is important to clarify that data used to train SVMs has to be from past, closed terms, because knowing the actual final performance achieved by students is required to train SVMs. Once SVMs are trained, they can be used to do real-time predictions for an undergoing term. These predictions are added to the prediction stack, where all predictions for passed weeks are stored. Finished this prediction step, graphs and information are elaborated and presented to teachers for student evaluation [21]. The last value in the weekly prediction stack always correspond to the week in course, whose predictions are made using its corresponding SVM as if the week had passed.

The aim of this paper is comparing the accuracy of the prediction depending on the type of data used for classifying, as well as determining which is the best configuration of features for this particular problem. To compare the results of the experiments, we will calculate the accuracy of the weekly predictions. Accuracy is defined as number of correctly classified students out of all the 336 students (i.e. the system classified them to the same group they finally got).

4.2 Experiment 1: behavioural data as performance predictors

From the event database, only the behavioural features have been used for this experiment. Specifically, the set of significant behavioural features is presented in Table 2. A total of eight behavioural features have been used in this experiment.

The accuracy results for this experiment are presented in Section 4.6, together with the other experiments to illustrate the comparison.

4.3 Experiment 2: learning data as performance predictors

Experiment 2 uses learning features for prediction. The learning features (up to 53 different values) are presented in Table 3.

The accuracy results for this experiment are presented in section 4.6, together with the other experiments to illustrate the comparison.

4.4 Experiment 3: behavioural and learning data as performance predictors

In this experiment, all the features (both behavioural and learning) are used for training the system and predicting the performance, summing up 61 features. The whole set of features are the ones presented in Tables 2 and 3. The values of accuracy for every week are presented altogether in Section 4.6.

4.5 Experiment 4: selecting the most correlated data as performance predictors

A possible improvement over the previous experiments has also been considered. It is possible that some data may contain noise because of the low correlation with the final mark. To isolate the potentially noisy data, a new experiment is presented. We have calculated the lineal correlation between every feature and the final mark, so that we have selected the most correlated ones, assuming that a higher correlation implies a higher significance of the feature.

Table 2: Behavioural features.

Features	Values
Number of frontpage visits	1 value
Number of maze downloads	1 value
Number of submissions per stage	6 value (1 for each stage)
Total	8 values

Table 3: Learning features.

Features	Values
Accumulated mark	1 value
Average marks (percentage) per stage	6 values (1 for each stage)
Mark (percentage) per maze and stage	15 values (1 for each maze of each stage)
Time (seconds) to finish each stage	6 values (1 for each stage)
Time (seconds) to finish each maze and stage	15 values (1 for each maze of each stage)
Difficulty levels selected per maze and stage	10 values (1 for each maze of each stage but the first one)
Total	53 values

There are several coefficients to measure the correlation degree. In this case, the most suitable is that of Spearman [22], due to the fact that some variables are discrete and the continuous ones may not be normal. The Spearman coefficient is calculated every week between every feature of experiment 3 and the final mark of the students. From the study, it can be concluded that, in general, learning features have a higher correlation than behavioural features. In fact, there are 7 learning features and only 1 behavioural feature with a high correlation coefficient (Spearman coefficient > 0.7).

For this experiment, every week we have selected the 23 features with the higher correlation out of the total of 61, which can be different for every week. The number of features (23) has been chosen so that no feature with at least moderate correlation (Spearman coefficient > 0.4) is out of the sample. The accuracy results for this experiment as well as the discussion about them are presented in Section 4.6.

4.6 Experimental results and discussion

As a measure of performance for the classifier, a pure classification technique has been used. Every classifier (one for each week), for every test sample gives three probabilities values as a result, that is, it gives the probability of the student of belonging to one of the three possible classes (high, medium or low performance). To compare the results of the experiments, the average accuracy of the classifiers of every week is obtained as the proportion of well-classified samples, considering that the sample is finally classified in the class with a higher probability. The accuracy values for every experiment are shown in Table 4 and Fig. 4.

As expected, the worst results are obtained when using only behavioural data (experiment 1). This is caused by the fact that behavioural data are only describing the use of the platform instead of the actual progression of the students in the learning process. This fact was detected in experiment 4, when obtaining the correlation between the features and the final mark. Indeed, behavioural features are the less correlated ones.

Experiment 2, using learning data to predict, improves the results of experiment 1. This result is also logical, since the learning features are describing the progress of students in the learning process (confirmed by the correlation study, where the learning features are much more correlated with the final mark than the behavioural ones).

Table 4: Accuracy results.

Week	Experiment 1	Experiment 2	Experiment 3	Experiment 4
1	0.452381	0.464286	0.464286	0.443452
2	0.428571	0.482143	0.47619	0.446429
3	0.497024	0.526786	0.517857	0.544643
4	0.479167	0.553571	0.553571	0.547619
5	0.479167	0.53869	0.556548	0.5625
6	0.580357	0.613095	0.604167	0.544643
7	0.583333	0.657738	0.636905	0.595238
8	0.613095	0.64881	0.660714	0.627976
9	0.60119	0.672619	0.675595	0.669643
10	0.627976	0.723214	0.741071	0.72619

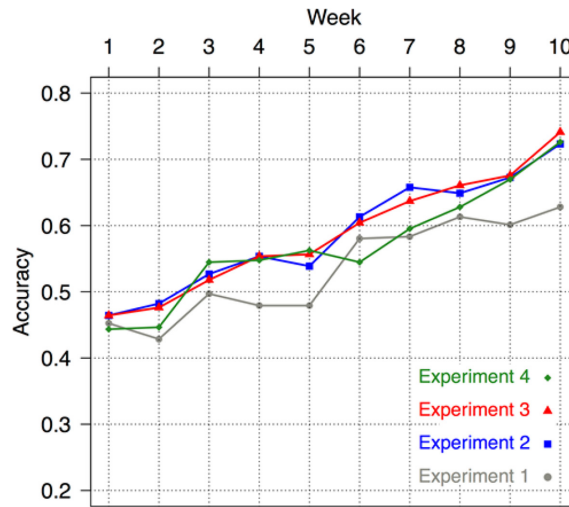


Figure 4: Accuracy results.

From the accuracy results, it can be deduced that the best results are obtained for experiment 3, when all the features are used to predict. Although experiments 2, 3 and 4 have very similar results, experiment 3 is the one showing a lower variance (the curve is smoother, so the classifier is more robust). This is an evidence that the use of both behavioural and learning data offers better results. In fact, the accuracy of the classifier of experiment 4 is slightly lower than that of experiment 3. Therefore, reducing the number of features used by the SVM prediction algorithm does not improve the results of prediction, even in the case of eliminating the features that are supposed to have no significant correlation with the final grade. This result is due probably to the fact that the SVM algorithm is detecting non-linear relationships, which are lost by reducing the number of input variables.

5 CONCLUSIONS

In this paper, it is studied the use of different sets of data to predict the student performance in a subject. As a result, the system obtains a weekly classification of each student as the probability of belonging to one of three classes: high, medium and low performance. The classifiers are based on Support Vector Machine technique.

Two research questions were posed: which is the best configuration for the data set to obtain a better predictor of the students' performance? And, how accurate is the resulting predictive model? The results show that, supplementing behavioural data (describing the system usage) with learning data (describing the learners progression during their training) allows us to obtain better and more stable predictions about the results of the students in a learning system. So, it can be deduced that the use of heterogeneous significant data enriches the final performance of the prediction algorithms. Moreover, the levels of accuracy increase as the course progresses, reaching a value around 73% at the last week. Even in the first weeks, the accuracy values are around 50% (significantly higher than a random classification of 33%), which can be considered a good result considering that there is little information about the learning progress in this moment. These results can be improved in the future when more data are available but they are enough to detect learning problems and tendencies.

REFERENCES

- [1] Long, P. & Siemens, G., Penetrating the fog: analytics in learning and education. *Educase Review*, 2011.
- [2] Siemens, G., Learning analytics: envisioning a research discipline and a domain of practice. *International Conference on Learning Analytics and Knowledge LAK*, pp. 4–8, 2012.
<http://dx.doi.org/10.1145/2330601.2330605>
- [3] Macfadyen, L.P. & Dawson, S., Mining LMS data to develop an ‘early warning system’ for educators: a proof of concept. *Computers & Education*, **54**(2), pp. 588–599, 2010.
<http://dx.doi.org/10.1016/j.compedu.2009.09.008>
- [4] Kotsiantis, S.B., Use of machine learning techniques for educational proposes: a decision support system for forecasting students’ grades. *Artificial Intelligence Review*, **37**(4), pp. 331–344, 2012.
<http://dx.doi.org/10.1007/s10462-011-9234-x>
- [5] Huang, S. & Fang, N., Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. *Computers & Education*, **61**(0), pp. 133–145, 2013.
<http://dx.doi.org/10.1016/j.compedu.2012.08.015>
- [6] Lykourantzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V. & Loumos, V., Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology*, **60**(2), pp. 372–380, 2009.
<http://dx.doi.org/10.1002/asi.20970>
- [7] Illanas Vila, A.I., Calvo Ferrer, J.R., Gallego-Dur n, F.J. & Llorens Largo, F., Predicting student performance in translating foreign languages with a serious game. In *INTED2013 Proceedings*, Valencia, Spain, pp. 52–59, 2013.
- [8] Yoo, J. & Kim, J., Can online discussion participation predict group project performance? investigating the roles of linguistic features and participation patterns. *International Journal of Artificial Intelligence in Education*, **24**(1), pp. 8–32, 2014.
<http://dx.doi.org/10.1007/s40593-013-0010-8>
- [9] Schalk, P.D., Wick, D.P., Turner, P.R. & Ramsdell, M.W., Predictive assessment of student performance for early strategic guidance. In *Frontiers in Education Conference (FIE)*, pp. S2H–1–S2H–5, 2011.
<http://dx.doi.org/10.1109/fie.2011.6143086>
- [10] Petkovic, D., Okada, K., Sosnick, M., Iyer, A., Zhu, S., Todtenhoefer, R. & Huang, S., Work in progress: a machine learning approach for assessment and prediction of teamwork effectiveness in software engineering education. In *Frontiers in Education Conference (FIE)*, pp. 1–3, 2012.
<http://dx.doi.org/10.1109/fie.2012.6462205>
- [11] Ley, T. & Kump, B., Which user interactions predict levels of expertise in work-integrated learning? In *EC-TEL*, 8095, pp. 178–190, 2013.
http://dx.doi.org/10.1007/978-3-642-40814-4_15
- [12] Wang, A.Y. & Newlin, M.H., Characteristics of students who enroll and succeed in psychology web-based classes. *Journal of Educational Psychology*, **92**(1), p. 137, 2000.
<http://dx.doi.org/10.1037/0022-0663.92.1.137>
- [13] Wang, A.Y. & Newlin, M.H., Predictors of performance in the virtual classroom: identifying and helping at-risk cyber-students. *The Journal of Higher Education*, **29**(10), pp. 21–25, 2002.
- [14] Campbell, J.B. & Oblinger, D.G., Academic analytics. *Educase*, 2007.

- [15] Goldstein, P.J. & Katz, R.N., Academic analytics: the uses of management information and technology in higher education. *Educase*, 2005.
- [16] Villagr -Arnedo, C., Castel De Haro, M.J., Gallego-Dur n, F.J., Pomares Puig, C., Suau P rez, P. & Cort s Va llo, S., Real-time evaluation. In *EDULEARN09 Proceedings*, Barcelona, Spain, pp. 3361–3369, 2009.
- [17] Cortes, C. & Vapnik, V., Support-vector networks. *Machine Learning*, **20**(3), pp. 273–297, 1995.
<http://dx.doi.org/10.1007/BF00994018>
- [18] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc.: New York, NY, USA, 1995.
- [19] Vapnik, V.N., *Statistical Learning Theory*, 1st edn., Wiley, 1998.
- [20] Wu, T.F., Lin, C-J. & Weng, R.C., Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, **5**, pp. 975–1005, 2004.
- [21] Villagr -Arnedo, C., Gallego-Dur n, F.J., Molina-Carmona, R. & Llorens-Largo, F., Boosting the learning process with progressive performance prediction. In *Design for Teaching and Learning in a Networked World*, vol. 9307, eds. G. Conole, T. Klobu ar, C. Rensing, J. Konert &  . Lavou e, Springer International Publishing: Cham, pp. 638–641, 2015.
http://dx.doi.org/10.1007/978-3-319-24258-3_77
- [22] Szczepa nska, A., Research design and statistical analysis, third edition by Jerome L. Myers, Arnold D. Well, Robert F. Lorch, Jr. *International Statistical Review*, **79**(3), pp. 491–492, 2011.