Predicting Student Progress from Peer-Assessment Data

Michael Mogessie Ashenafi
University of Trento
via Sommarive 9, 38123
Trento, Italy
+390461285251
michael.mogessie@unitn.it

Marco Ronchetti
University of Trento
via Sommarive 9, 38123
Trento, Italy
+390461282033
marco.ronchetti@unitn.it

Giuseppe Riccardi
University of Trento
via Sommarive 9, 38123
Trento, Italy
+390461282087
giuseppe.riccardi@unitn.it

ABSTRACT

Predicting overall student performance and monitoring progress have attracted more attention in the past five years than before. Demographic data, high school grades and test result constitute much of the data used for building prediction models. This study demonstrates how data from a peer-assessment environment can be used to build student progress prediction models. The possibility for automating tasks, coupled with minimal teacher intervention, make peer-assessment an efficient platform for gathering student activity data in a continuous manner. The performances of the prediction models are comparable with those trained using other educational data. Considering the fact that the student performance data do not include any teacher assessments, the results are more than encouraging and shall convince the reader that peer-assessment has yet another advantage to offer in the realm of automated student progress monitoring and supervision.

Keywords

Progress prediction; peer-assessment; learning analytics.

1. INTRODUCTION

Common examples of traditional student assessment methods are end-of-course examinations that constitute a very high proportion of final scores and other standardised and high stakes tests.

There are, however, other student-centric, yet less practiced, forms of assessment. Formative assessment is a fitting example [7]. It is designed with the goal of helping students meet specified learning goals through continuous discussion, gauging and reporting of their performance.

Peer-assessment is another form of assessment, which may be designed with summative or formative goals. It is a form of assessment where students evaluate the academic products of their peers [15].

Automated peer-assessment provides a rich platform for gathering data that can be used to monitor student progress. In such context, another dimension of peer-assessment emerges – its potential to serve as a foundation for building prediction models on top of.

In this study, we demonstrate how this potential can be exploited by building linear regression models for predicting students' weekly progress and overall performance for two undergraduate-level computer science courses that utilised an automated peer-assessment.

The rest of this paper is organised as follows. The next section discusses recent advances in student performance prediction. Section 3 presents a brief overview of the web-based peer-assessment platform using which the data was collected. Section 4 discusses details of the data and the features that were selected to build the prediction models. Section 5 provides two interpretations of student progress and details how these interpretations determine which data shall be used for building the models. Section 6 introduces the reader to how the prediction models are trained and provides details of the prediction performance evaluation metrics reported. Section 7 discusses the first interpretation of progress prediction and demonstrate the respective prediction models. Section 8 builds upon the second interpretation and follows the same procedure as section 7. Section 9 provides a short discussion and conclusion of the study.

2. PREVIOUS WORK IN PREDICTING STUDENT PERFORMANCE

Earlier studies in student performance prediction investigated the correlation between high school grades and student demographic data and success in college education as evidenced by successful completion of studies [1, 6].

Unsurprisingly, many of these studies were conducted by scholars in the social sciences and involved the use of common correlation investigation methods such as linear and logistic regression. The large majority of recent studies have, however, been conducted in the computer science discipline. These studies use data from courses administered as part of either computer science or engineering programmes at the undergraduate level. Of these, many focus on predicting performance of freshman and second year students enrolled in introductory level courses.

A generic approach to student performance prediction is to predict overall outcome such as passing or failing a course or even forecasting successful completion of college as marked by graduation [9, 13, 14]. A further step in such an approach may include predicting the classification of the degree or achievement [8].

More fine-grained and sophisticated approaches involve predicting actual scores for tests and assignments as well as final scores and grades for an entire course.

Due to the varying nature of the courses and classes in which such experiments are conducted and advanced machine learning techniques that are readily available as parts of scientific software packages, the number distinct, yet comparable, studies in performance prediction has been growing steadily. Another factor, the proliferation of MOOCs, has fuelled this growth with the immense amount of student activity data generated by these platforms.

Examples of studies that utilise information from students' activities in online learning and assessment platforms in predicting performance include [2, 10, 11].

Apart from predicting end-of-course or end-of-programme performance, prediction models may be used to provide continuous predictions that help monitor student progress. When used in this manner, such prediction models could serve as instruments for early detection of at-risk students. Information provided by these models could then serve the formative needs of both students and teachers. Studies that demonstrate how prediction models can be used to provide continuous predictions and may serve as tools of early intervention include [5, 10].

The most common algorithms in recent literature that are used for making performance predictions are Linear Regression, Neural Networks, Support Vector Machines, Naïve Bayes Classifier, and Decision Trees.

Studies that follow less common approaches include those that use smartphone data to investigate the correlation between students' social and study behaviour and academic performance [16] and those that perform Sentiment Analysis of discussion form posts in MOOCs [4].

Two studies that present algorithms developed for the sole purpose of student performance prediction are [12] and [17].

3. THE PEER-ASSESSMENT PLATFORM

In 2012, an experimental web-based peer-assessment system was introduced into a number of undergraduate level courses at an Italian university. Using this peer-assessment system, students completed three sets of tasks during each week of the course. The weekly cycle started with students using the online platform to submit questions about topics that were recently discussed in class. These questions were then reviewed by the teacher, who would select a subset and assign them to students, asking them to provide answers. The assignment of the questions to students was automatically randomised by the system, which guaranteed anonymity of both students who asked the questions and those who answered them. Once this task was completed, the teacher would assign students the last task of the cycle, in which they would rate the answers provided by their peers and evaluate the questions in terms of their perceived difficulty, relevance and interestingness.

Eight cycles of peer-assessment were carried out in two undergraduate-level computer science courses, IG1 and PR2. Participation in peer-assessment activities was not mandatory. However, an effort to engage students in these tasks was made by awarding students with bonus points at the end of the course in accordance with their level of participation and the total number of peer-assigned marks they had earned for their answers. The design and development of the peer-assessment platform and the theoretical motivations for it are discussed in [3].

4. THE DATA

Because participation in peer-assessment tasks was not mandatory, there was an apparent decline in the number of participants towards the end of both courses. In order to minimise noise in the resulting prediction models, only peer-assessment activity data of those students who completed at least a third of the total number of tasks and for whom final grades were available were selected for building the models. This led to the inclusion of 115 student records for IG1 and 114 for PR2.

In a previous study [2], a linear regression model for predicting final scores of students using the same data was discussed. Experiments in that study revealed that predicting the range within which a score would fall was more accurate than predicting actual scores. Indeed, this is tantamount to predicting grades. During the experiments in that study, although attempts were made to build classification models that predicted grades in a multiclass classification manner, the results were found to be much better when actual scores were predicted using linear regression and those scores were mapped to grades according to mappings which were specified beforehand. Hence, the authors decided to apply those techniques in this study as well.

Grades are arguably the ideal approach to judging the performance levels of students because they usually span a wider range of scores, within which a student's scores are likely to fall if the student sits the same exam in relatively quick successions. Consequently, scores predicted by the linear regression models were transformed into grades.

The parameters used to build the linear regression models are:

Tasks Assigned (TA) – The number of tasks that were assigned to the student

Tasks Completed (TC) – The number of tasks that the student completed

Questions Asked (QAS) – The number of 'Ask a Question' tasks the student completed

Questions Answered (QAN) – The number of 'Answer a Question' tasks the student completed

Votes Cast (VC) – The number of 'Rate Answers' tasks the student completed

Questions picked for answering (QP) – The number of the student's questions that were selected by the teacher to be used in 'Answer a Question' tasks

Votes Earned (VE) – The number of votes the student earned for their answers

Votes Earned Total Difficulty (VED) – The sum of the products of the votes earned for an answer and the difficulty level of the question, as rated by students themselves, for all answers submitted by the student

Votes Earned Total Relevance (VER) – The sum of the products of the votes earned for an answer and the relevance level of the question, as rated by students themselves, for all answers submitted by the student

Votes Earned Total Interestingness (VEI) – The sum of the products of the votes earned for an answer and the interestingness level of the question, as rated by students themselves, for all answers submitted by the student

Selected Q total difficulty (SQD) – The sum of the difficulty levels of the student's questions, as rated by students themselves, which were selected to be used in subsequent tasks

Selected Q total relevance (SQR) – The sum of the relevance levels of the student's questions, as rated by students themselves, which were selected to be used in subsequent tasks

Selected Q total interestingness (SQI) – The sum of the interestingness levels of the student's questions, as rated by students themselves, which were selected to be used in subsequent tasks

Details of the linear regression model, possible justifications for its prediction errors and experiments comparing its performance to baseline predictors are provided in [2].

5. TWO INTERPRETATIONS OF PROGRESS PREDICTION

Monitoring student progress using prediction models requires making predictions using evolving student data at several intervals. Continuous peer-assessment data are the ideal candidate for building such prediction models.

Through years of experience, teachers are usually able to make educated guesses about how student are likely to perform at end-of-course exams by studying their activities throughout the course. Prediction models that use data from previous editions of the same course adopt and formalise such experience with greater efficacy.

Indeed, prediction models can be used not only to make one-off predictions of student performance at the end of a course, but also at several intervals throughout the course. While continuous predictions focus on determining student progress by evaluating performance at different stages, one-off predictions put more importance on whether a student would finally pass a course on not.

This study focuses on the former, making continuous predictions to measure student progress and provides two interpretations of student *progress*.

One interpretation compares a student's standing at any point in the course to the standings of students at the same point but from previous editions of the course. For instance, in a previous edition of a course, if student performance data at every week of the course were collected and if these data were complemented with end-of-course grades, in subsequent editions of the course, a student's performance at any week would be compared to the performances of students at that specific week in the previous edition of the course and the respective grade for the student's level of performance could be predicted. In favour of brevity, this interpretation of progress will be referred to as ***Progress Type A***.

The other interpretation focuses on evaluating how far a student is from achieving goals that they are expected to achieve at the end of a course. In a fairly simplified manner, this evaluation may be made by comparing the expected final grade of student at any point during the course to what is deemed to be a desirable outcome at the end of the course. For instance, predicting a student's end-of-course grade in the second week of an eight-week course and comparing that predicted grade to what is considered to be a favourable grade at the end of the course, which is usually in the range A+ to B-, can provide information about how far the student is from achieving goals that are set out at the beginning of the course. In favour of brevity, this interpretation of progress will be referred to as ***Progress Type B***.

6. TRAINING AND MEASURING THE PERFORMANCE OF THE PREDICTION MODELS

Peer-assessment data collected during the course were divided into weekly data according to the three sets of tasks completed every week. The final score of each student for the course was then converted into one of four letter grades.

The data for each week incorporate the data from all previous weeks. In this manner, the prediction model for any one week is built using more performance data than its predecessors. Naturally, the data used to build the model for the first week would be modest and the data for the final week model would be complete. In general, the performances of models from consecutive weeks were expected to be better.

A common metric used in measuring the performance of linear regression prediction models is the Root Mean Squared Error (RMSE). While RMSE provides information about the average error of the model in making predictions, the conversion of numerical scores to letter grades enables using more informative performance evaluation metrics.

The conversion of numerical scores to letter grades transforms this prediction into a classification problem, with grades treated as class labels. Although multiclass classification algorithms were not applied due to their relatively low performance for this specific task, transformation of predicted scores into grades permitted the application of any of the classification performance evaluation metrics. Therefore, performance is reported in terms of precision, recall, F1, False Positive Rates (FPR) and True Negative Rates (TNR).

When evaluating student performance prediction models, the two questions that are more critical than others are:

- How many of the students the model predicted not to be at-risk were actually at-risk and eventually performed poorly (False Positives) and
- How many of the students that the model predicted to be at-risk of failing were indeed at-risk (True Negatives).

A prediction model with a high FPR largely fails to identify students who are at risk of failing. Conversely, a model with a high TNR identifies the majority of at-risk students. The ideal prediction model would have a very low FPR and, consequently, a very high TNR.

The prediction models are evaluated at two levels. The first level is their performance in making exact prediction of grades. The second is their performance in making a prediction that is within a one grade-point range of the actual grade.

For the purpose of this study, the performance metrics are defined as follows.

Grade – Any of the letters A, B, C, D – A and B denote high performance levels and C and D, otherwise. Although C is usually a pass grade, it is generally not favourable and considered to be a low grade.

Positive – A prediction that is either A or B

Negative – A prediction that is either C or D

True – A prediction that is either the exact outcome or falls within a one grade-point range of the actual outcome

False – A prediction that is not True

Any combination of positive or negative predictions with true or false predictions yields one of the following counts – True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Important statistics that use these counts are Precision (P), Recall (R) and, inherently, F1 scores.

It should be noted that FPR and TNR provide two interpretations of the same outcome and that they are inversely proportional. Indeed, FPR = 1 – TNR.

7. MODELLING PROGRESS TYPE A

This type of progress monitoring compares a student's current progress at any week during the course to the progresses of past students at the same week of the course. The question that such an approach aims to answer is: 'Compared to how other students were doing at this stage in the past, how well is this student doing now?' 'How well' the student is doing is predicted as follows. First, a linear regression model is built using data collected from the first week to the week of interest. This data comes from a previous edition of the course and the predicted variable is the final score or grade, which is already available. Then, the student's performance at the week in question, represented using the parameters in section 4, is fed to the model to make a prediction. Such weekly information shall provide insight into whether the student is likely to fall behind other students or not.

The prediction errors for the course PR2 gradually decreased for successive weeks, as expected. For IG1, however, early decreases were followed by increases and a slight decrease in the final week. The average RMSE for PR2 for the eight models was 3.4 while it was 3.6 for IG1. The scores predicted were in the range 18 to 30 Figure 1 shows the weekly prediction errors for each course.

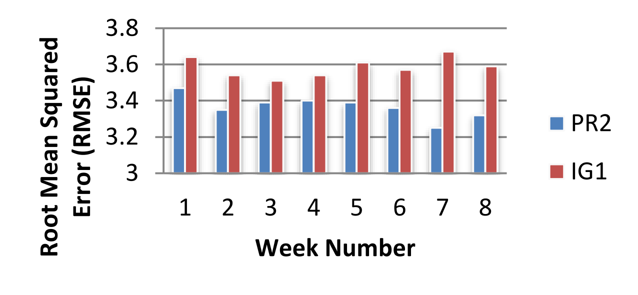

Figure 1. Prediction Errors for the models of each course over eight weeks

Low performance levels were recorded for exact grade prediction of the models for both courses. Specifically, High false positive rates persisted throughout the eight-week period.

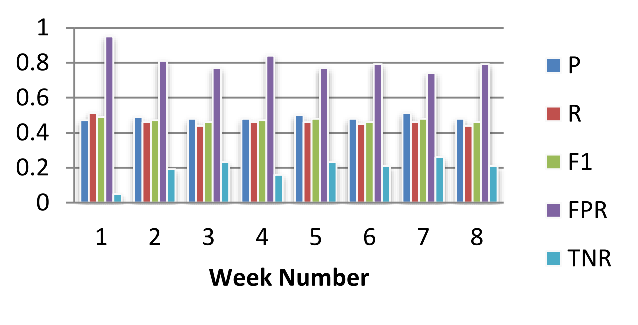

Figure 2. Exact grade prediction performance for PR2

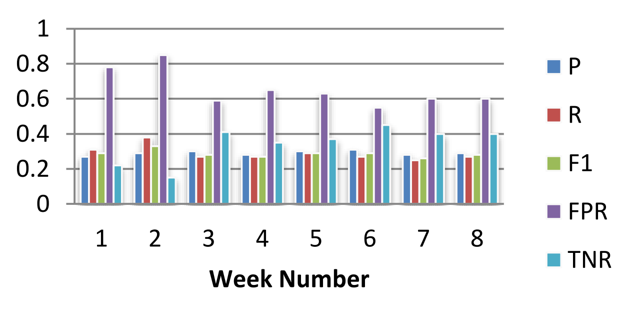

Figure 3. Exact grade prediction performance for IG1

As expected, performance levels of the models for both courses significantly increased for within one grade-point predictions. Low FPR and, consequently, high TNR were recorded even in the first week and performance increased gradually for both courses over the eight-week period.

The models that made within-one-grade-point predictions performed well from the very first week of the course. Although predictions are not made on exact grades, the wider range helps lower the rate of false positives and increase true positives. The same consideration may lead to an increase in false negatives, and hence, a decrease in true positives. However, the high precision and recall values for these models attest that this is not so in this case.

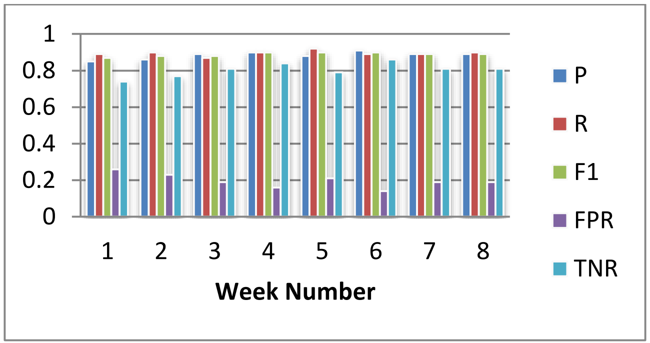

Figure 4. Within-one-grade-point prediction performance for PR2

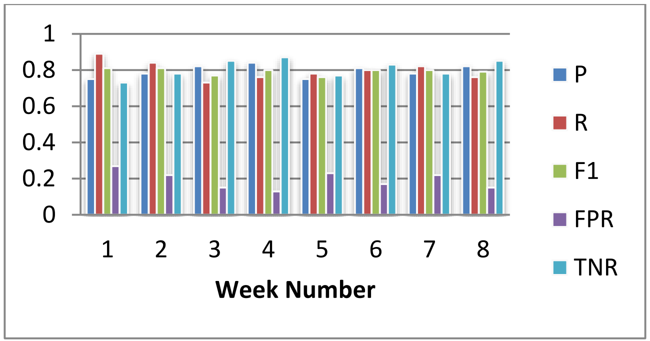

Figure 5. Within-one-grade-point prediction performance for IG1

8. MODELLING PROGRESS TYPE B

The focus of this type of measuring progress can be informally described as *measuring the gap* between a student's performance *now* and what it is expected to be *at the end* of the course. Modelling this type of progress only requires building a single linear regression model using the entire data from previous editions of the same course. Then, a student's performance data at any week, which is represented by an instance of the values for the parameters discussed in section 4, is fed to the linear regression equation to compute the expected score of the student. This score is then transformed to a grade. Such weekly information would help keep track of a student's progress towards closing this gap and achieving the desired goals.

The prediction errors of this model for the eight weeks are reported in Figure 6. The prediction errors for both courses were significantly lower than those for Progress Type A, with the model for PR2 having an average RMSE of 3.0 and the model for IG1 scoring a higher average RMSE of 3.5. Moreover, prediction errors for both courses consistently decreased throughout the eight weeks.

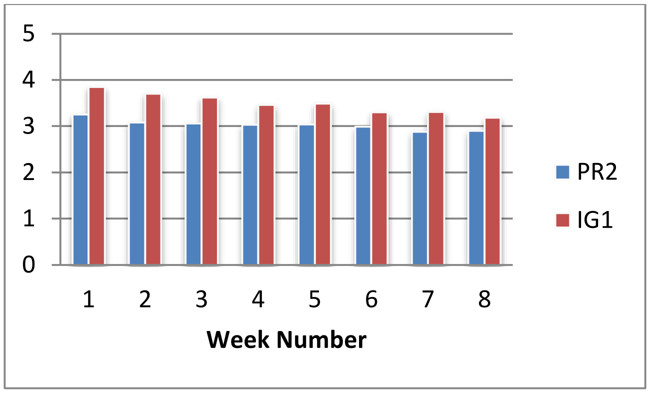

Figure 6. Prediction errors of the model of the two courses over an eight-week period

Exact grade prediction performance, although better than that of Progress Type A, was still low for both courses.

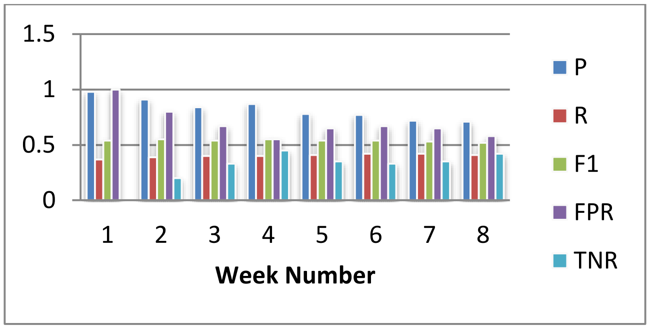

Figure 7. Exact grade prediction performance for PR2

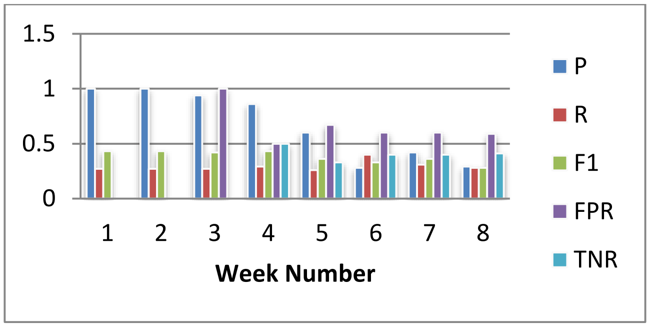

Figure 8. Exact grade prediction performance for IG1

Similar to the models of Progress Type A, this model had very high levels of performance in predicting grades that fell within one grade-point of the actual grades. Prediction performance was very high in the first week and consistently increased, albeit by small amounts, throughout the remaining weeks for both courses.

Missing FPR and TNR values for both courses in the beginning weeks imply that predictions of the model were distributed over TP and FN values. However, high precision values during those weeks indicate that FN values were very low.

Overall, the model for Progress Type B outperformed the models that from Progress Type B, for both courses.

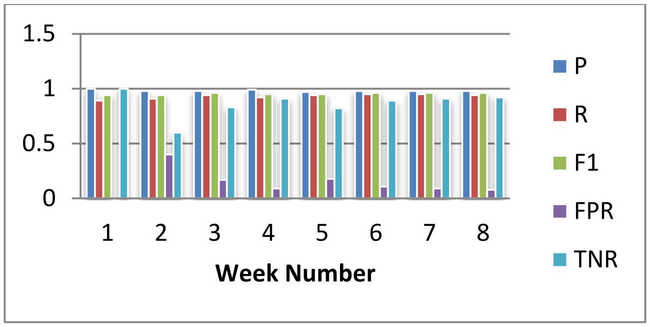

Figure 9. Within-one-grade-point prediction performance for PR2

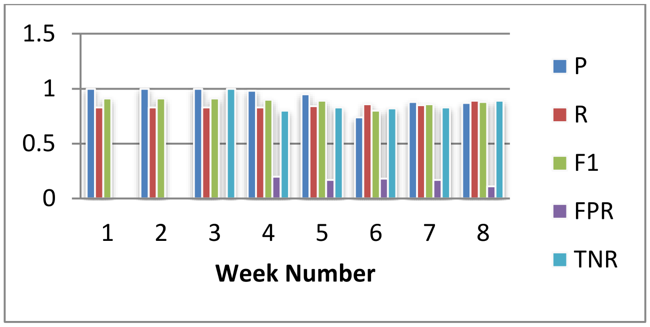

Figure 10. Within-one-grade-point prediction performance for IG1

9. DISCUSSION AND CONCLUSION

From peer-assessment tasks that were conducted over an eight-week period in two courses, data were used to build several prediction models according to two distinct interpretation of performance prediction. While the first interpretation focused on comparing the performance of a student at any week during the course to those of past students' performance levels obtained in the same week, the second focused on measuring how far a student is from achieving the desired level of performance at the end of a course.

The approach of using data from previous editions of the same course may raise doubts as to whether different editions of the same course are necessarily comparable. However, the extents to which the prediction models discussed here performed should convince the reader that this is indeed possible. Performance of the models is in fact expected to improve with increase in the number of previous editions of the course used as input for making predictions. Indeed, the long-term consistency in the number of below-average, average and above average students over many editions of a course is how many teachers usually measure the overall difficulty level of questions that they include in exams.

Although exact grade predictions did not produce satisfactory levels of performances for either approach, high levels of performance were obtained for both interpretations of student progress when making within-one-grade-point predictions. This signifies the promising potential of carefully designed peer-assessment and the prediction models built using data generated from it as tools of early intervention.

While the statement that a student's performance at the end of a course can be fairly predicted as early as the first weeks of the course from their peer-assessment activity may be construed as simplistic, it is worth noting that the experiments were carried out in two computer science courses and that the results suggest otherwise.

While a comparison between the performances of the models for the two courses may be made, the reasons behind one model outperforming the other may be latent at this stage and require detailed investigation. Hence, the authors decided to defer making such comparisons until a later stage.

10. REFERENCES

[1] Al-Hammadi, A. S., and Milne, R. H. (2004). A neuro-fuzzy classification approach to the assessment of student performance. In Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on (Vol. 2, pp. 837–841 vol.2). http://doi.org/10.1109/FUZZY.2004.1375511

[2] Ashenafi, M. M., Riccardi, G., and Ronchetti, M. (2015). Predicting students’ final exam scores from their course activities. In Frontiers in Education Conference (FIE), 2015 IEEE (pp. 1–9). http://doi.org/10.1109/FIE.2015.7344081

[3] Ashenafi, M.M., Riccardi, G., & Ronchetti, M. (2014, June). A Web-Based Peer Interaction Framework for Improved Assessment and Supervision of Students. In World Conference on Educational Multimedia, Hypermedia and Telecommunications (Vol. 2014, No. 1, pp. 1371-1380).

[4] Chaplot, D. S., Rhim, E., and Kim, J. (2015). Predicting student attrition in moocs using sentiment analysis and neural networks. In Proceedings of AIED 2015 Fourth Workshop on Intelligent Support for Learning in Groups.

[5] Coleman, C. A., Seaton, D. T., and Chuang, I. (2015). Probabilistic Use Cases: Discovering Behavioral Patterns for Predicting Certification. In Proceedings of the Second (2015) ACM Conference on Learning @ Scale (pp. 141–148). New York, NY, USA: ACM. http://doi.org/10.1145/2724660.2724662

[6] Evans, G. E., and Simkin, M. G. (1989). What best predicts computer proficiency?. Communications of the ACM, 32(11), 1322-1327. http://doi.org/10.1145/68814.68817

[7] Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. Assessment in Education, 4(3), 365-379. http://doi.org/10.1080/0969594970040304

[8] Jiang, S., Williams, A., Schenke, K., Warschauer, M., and O'dowd, D. (2014). Predicting MOOC performance with week 1 behavior. In Educational Data Mining 2014.

[9] Karamouzis, S. T., and Vrettos, A. (2009). Sensitivity Analysis of Neural Network Parameters for Identifying the Factors for College Student Success. In Computer Science and Information Engineering, 2009 WRI World Congress on (Vol. 5, pp. 671–675). http://doi.org/10.1109/CSIE.2009.592

[10] Koprinska, I., Stretton, J., and Yacef, K. (2015). Predicting Student Performance from Multiple Data Sources. In C. Conati, N. Heffernan, A. Mitrovic, and M. F. Verdejo (Eds.), Artificial Intelligence in Education SE - 90 (Vol. 9112, pp. 678–681). Springer International Publishing. http://doi.org/10.1007/978-3-319-19773-9_90

[11] Manhães, L. M. B., da Cruz, S. M. S., and Zimbrão, G. (2014). WAVE: An Architecture for Predicting Dropout in Undergraduate Courses Using EDM. In Proceedings of the 29th Annual ACM Symposium on Applied Computing (pp. 243–247). New York, NY, USA: ACM. http://doi.org/10.1145/2554850.2555135

[12] Meier, Y., Xu, J., Atan, O., and van der Schaar, M. (2015). Predicting Grades. Signal Processing, IEEE Transactions on, PP (99), 1. http://doi.org/10.1109/TSP.2015.2496278

[13] Nghe, N. T., Janecek, P., and Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports, 2007. FIE ’07. 37th Annual (pp. T2G–7–T2G–12). http://doi.org/10.1109/FIE.2007.4417993

[14] Plagge, M. (2013). Using Artificial Neural Networks to Predict First-year Traditional Students Second Year Retention Rates. In Proceedings of the 51st ACM Southeast Conference (pp. 17:1–17:5). New York, NY, USA: ACM. http://doi.org/10.1145/2498328.2500061

[15] Topping, K.J. (1998). Peer Assessment Between Students in Colleges and Universities. Review of educational Research, 68(3), 249-276. http://doi.org/10.3102/00346543068003249

[16] Wang, R., Harari, G., Hao, P., Zhou, X., and Campbell, A. T. (2015). SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 295–306). New York, NY, USA: ACM. http://doi.org/10.1145/2750858.2804251

[17] Watson, C., Li, F. W. B., and Godwin, J. L. (2013). Predicting Performance in an Introductory Programming Course by Logging and Analyzing Student Programming Behavior. In Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on (pp. 319–323). http://doi.org/10.1109/ICALT.2013.99