

Using access log data to predict failure-prone students in Moodle using a small dataset

Hamidullah Sokout^{1,*}, and Tsuyoshi Usagawa¹

¹Graduate School of Science and Technology, Kumamoto University, Japan

Abstract. In this paper, the authors present a predictive model for failure-prone students using access log data from two small datasets in the Moodle learning system. Although various advanced machine learning algorithms, especially supervised predictive methods, can be used with very large datasets, these tools are often not available for most initial university research, especially in developing countries, to predict learners' future outcomes. The authors examined the use of students' access patterns to track failure-prone students so that early interventions could be made to prevent failure or dropout. Real data were collected through explicit learners' actions, such as completing assignments and taking quizzes, from two compulsory blended courses, Operating System (junior level, or third year) and System Analysis and Design (sophomore level, or second year). Research methods were predominantly quantitative. The proposed models correctly predicted failure-prone students before the end of the second academic month (fifth week) for both courses (88% of the class for juniors and 86% of the class for sophomores), which made it possible to intervene early and provide required support during the semester. Similarly, the study outcomes showed the students' past performance, specifically their grade point average, could affect their final performance. The outcomes of this study can be used to analyze the behaviors of learners that lead to high success and high retention rate. Furthermore, the study results will be used to report and provide feedback to the educational parties to value the quality of teaching and learning, the improvement of course materials, and increasing learner success.

1 Introduction

Nowadays, learning analytics (LA) applications are emerging in education and are widely used by academics for early, real-time learning performance prediction [1]. These approaches can be used to predict learners' behaviors in time series to increase students' reflection and improvement. According to Wong [2], use of LA improves student retention, predicts student performance, detects undesirable learning behaviors and emotional state, and identifies students at risk and promotes their reflection and improvement. Meanwhile, LA helps institutions to make effective use of available data effectively in decision-making, increased cost-effectiveness, and timely feedback and intervention.

Due to improvements in science and technology in developed countries and the availability of huge datasets, these countries are easily benefiting and applying different approaches in their educational environments to identify students' failure or dropout early. However, the unreliability of educational data, lack of historical data, lack of students' engagement and promotion, and lack of well-defined intervention mechanisms are the greatest challenges for blended learning environments in developing countries, particularly in Afghanistan, which made it difficult to perform early prediction. In addition,

data available in developing countries and in initial university studies are often limited by smaller sets of observations than are typically preferred for models built using machine learning algorithms [3-4].

In some previous work, particularly in medical research, the constraints of small datasets have been overcome by using synthetic data or random oversampling techniques to improve model accuracy and generalization ability. However, the main drawback and consequence of using such algorithms are a high probability of overfitting the training datasets, sensitivity to noisy data, and degradation of model performance [5-6].

In this research, the authors have investigated the log patterns of students in the Moodle Learning Management System (LMS) to track the behavior of students and identify failure-prone students in a timely fashion to prevent failure or dropout. To measure success, the authors have only considered the characteristics of students extracted from log data and have used login frequency to group similar patterns related to at-risk students. In addition, this study intended to assess other major factors (beyond login frequency), including grade point average (GPA) in different semesters along with university entrance

* Corresponding author: sokout@st.cs.kumamoto-u.ac.jp

examination (Kankor^a) score, that might have influenced student outcomes. The study results can be used to provide feedback to educational parties to increase the pass rate and achieve learning objectives.

The rest of the paper is organized as follows: Section 2 provides a literature review quoting examples from existing studies. Section 3 describes the study methodology. In particular, the authors first emphasize the study data and context, followed by the data analysis. Section 3 provides the results, followed by discussions and conclusions (Section 4).

2 Literature review

In the educational environment, particularly in higher education, the lecturers are trying to accomplish course outcomes, whereas the students want to fulfill the requirements and obtain course credit. The main concern is to balance the quality of teaching and learning. On the other hand, implementing e-learning platforms in the educational environment is not particularly effective in motivating learners to achieve the course objectives. However, effective deployment of Web-based online learning and efficient use of large amounts of learning activity data, particularly log data generated by the same Web-based learning systems, has enabled higher education actors to reshape the educational environment to make it more effective and results-oriented. In the last few decades, almost all higher education institutions in developed countries, developing countries, and even in the least developed countries have started to facilitate their environment with technology to enable learners to interact with each other and share learning resources. During these interactions, large masses of fine-grained data are generated automatically. These data contain much information that could be useful in predicting future learning outcomes using analytical techniques (Sokout et al., 2018). Proper analysis of educational data can provide insights that can help improve the aims and effectiveness of the educational environment.

Early prediction of failure-prone students or those at risk of dropout is an interesting and timely topic that has been quite widely addressed in the literature, where higher education institutions are believed not only to be widening their outreach and impact, but also to be improving their learning outcomes, with a significant retention rate and high learner engagement. As Afghanistan proceeds to improve the quality of teaching and learning through technology and strives to provide better higher education for its citizens, it still must tackle the major issue of high failure and dropout rates in its educational institutions. According to achievement documents for 2019, among all the courses conducted in the Faculty of Computer Science at Kabul Polytechnic University (KPU), 45% of the students had to take a

second or third round of examinations. Among the 45% who failed during 2019, 14% of students lost an academic year [7]. This was not because the learners did not deserve higher education, but rather because they were not motivated and autonomous when dealing with technology. Therefore, there is a dire need for real-time monitoring of students' learning activity during the semester to intervene with at-risk students early and reduce the failure rate.

According to Owen et al. [8], use of learning analytics enables educational institutions to better analyze and predict students' outcomes and provide timely intervention based on student learning activity. The authors highlighted the early predictions for 59 students in a blended learning course. They obtained 83.5% accuracy by the sixth week of the semester using applied principal component regression. In addition, they highlighted four major online factors and three traditional factors as critical influences affecting students' academic outcomes. Similarly, Fungai et al. [9] reported 75% accuracy in a compulsory second-year course using an unsupervised method with 88 students. The authors used ratios of weekly quiz attempts to login frequency and applied k-means clustering algorithms.

Meanwhile, Romero et al. [10] claimed 65% prediction accuracy in their studies with 438 students. The authors extracted different features like assignments, quizzes, and forum activities from the Moodle system, compared multiple algorithms, and concluded that the fuzzy rule learning algorithm and decision trees performed better than statistical methods. Furthermore, Sukhbaatar [11] achieved early identification of at-risk students with 42%–73% prediction accuracy in the middle of the semester using a supervised method. Last but not least, Milne et al. [12] used LMS-logged students' usage over the semester for nine courses involving 703 students. The authors characterized active and non-active students by generating a weekly graph for each course. The authors concluded that at-risk status in the courses was associated with less LMS usage, whereas excellent grades were associated with more use of LMS.

In these few studies and some other studies with small datasets, researchers have shown the impact of LA and the effect of monitoring learners' online activities on student performance. However, most current studies identified student performance of students at the middle or end of the semester, which is too late to intervene and provide necessary feedback to at-risk students. In addition, recent studies have mainly focused on log data and the lack of some other factors, specifically course structure and student differences, that may influence overall student outcomes. Similarly, to the best of the authors' knowledge, few studies have reported the online learning behavior of students and effectively tracked their learning experiences to better analyze the details of their online learning activities. Moreover, no effective steps have been taken to analyze the use of ICT tools (LMS) in the educational environment in developing countries, particularly in Afghanistan. Therefore, this study aimed to find an appropriate strategy to better

^aKankor is a proficiency-based university entrance exam in Afghanistan that is taken by school graduates to be admitted into higher education institutions. The maximum score for the Kankor exam is 360.

analyze the online student behaviors as early as possible in the semester. This would make it possible to intervene with at-risk students, as well as paving the way for effective use of ICT tools in the educational environment.

3 Methodology

3.1 Data and study context

This study used a quantitative approach. The preliminary data were collected through explicit learners' actions, including completing assignments and taking weekly quizzes based on actual data in two blended courses, namely "Operating System" (OS) and "System Analysis and Design (SAD)" in 2018. Both courses are compulsory credit courses taught in the Faculty of Computer Science at the University of Kabul Polytechnic. The total numbers of enrolled students for OS and SAD were 114 and 70 respectively. Among the enrolled students, only 106 students from the OS course with a total of 24,605 "course view log entries" and 62 students with a total of 17,416 "course view log entries" from the SAD course were retained for classification. Figure 1 indicates the proportions of failure-prone and non-risk students for both courses. The semester lasted for 15 weeks and included both online and face-to-face learning activities. The course content, quizzes, assignments, and discussions were scheduled for release every week on Moodle. Both courses also included on-campus activities, including weekly lecture seminars, laboratory sessions, and mid-term and final examinations.

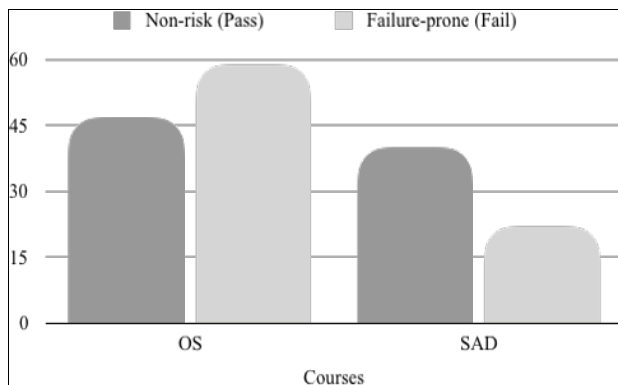


Fig. 1. Proportions of failure-prone and non-risk students in courses.

All on-campus and online activities contributed to the final grade of each student. The average final score to earn credit defined 60% as a minimum score; in the Afghan credit system, the passing score for each subject is 55%, and the attainment of a 60% overall average per semester is considered necessary. In addition, 40% of each subject's grades represent class activities, including assignments and quizzes. Therefore, those students who miss class activities must obtain the required score from the paper-based examination. Table 1 provides a statistical summary of online activities for both courses.

Table 1. Statistical summary of online activities for both courses.

Courses	All Semester Quizzes and Assignments	Number of Enrolled	classified students after cleaning dataset	Mean Score	Standard Deviation
SAD	Quizzes	70	62	84.4	5.9
	Assignments			74.8	15.4
OS	Quizzes	114	106	83.0	8.0
	Assignments			59.0	18.0

3.2 Data analysis

In this study, a moving threshold was used as a statistical method and was calculated weekly on aggregate data to model the recurring patterns of students falling behind the rest of the class. The collected data were first accumulated on a week-by-week basis and classified in two steps. For the first step, the collected data were classified according to a defined "access log" threshold A_t and a "mean score of quizzes and assignments" threshold Q_t for any given week n . For each week, the average "access log" was calculated, and 30% weekly access was defined as a minimum threshold for each student because students access online materials on-campus. Meanwhile, the "mean score of quizzes and assignments" threshold was set to 0.45 as a minimum score for non-risk students. Q_t was arbitrarily set to 0.45 based on high academic pressure on students, the volume of students' online and practical activities during the semester, and the availability of essential content before mid-semester for passing the courses. Equation (1) defines the status of the students in any week n . The itemized list below describes the variables used in the classifier equation:

I : total number of students after cleaning the dataset (106 for OS and 62 for SAD)

N : number of weeks (15 weeks in a semester)

T : failure-prone status threshold

f_i : cumulative weekly election based on the number of online activities for the i^{th} student

k_n^i : predictor variable, for the i^{th} student in the n^{th} week

x_n^i : score of the i^{th} student in the n^{th} week

y_n^i : access log frequency of the i^{th} student in the n^{th} week

A_t : Access log threshold

Q_t : Quizzes and assignments score threshold

$$k_n^i = \begin{cases} 0, & x_n^i < Q_t \ \&\& \ y_n^i < A_t, \ i=1,2,3,\dots,I \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

$$f_i = \begin{cases} 0, & \sum_{n=1}^N k_n^i < T \\ 1, & \text{otherwise} \end{cases}, \ n = 1,2,3,\dots,N \quad (2)$$

As expressed in Eq. (1), the target condition k was satisfied when the score and the access pattern of a student were less than the defined thresholds. In this case, the function assigned the student as inactive (failure-prone), otherwise as active (non-risk). With the help of Eq. (1), a model has been trained that can be used to test the model for prediction of failure-prone students.

In addition, in the second classification step, the final target variable was determined by the number of weekly online activities. The students were supposed to submit quizzes and assignments each week. The number of submitted assignments and attempted quizzes was accumulated on a weekly basis, and a threshold of 70 percent for the numbers of submitted assignments and attempted quizzes was set for early dropout detection for each week. The cumulative weekly selection was based on the number of online activities and was calculated using Eq. (2). The target condition was satisfied whenever Eq. (2) became less than the arbitrarily defined threshold, at which point the function classified the student as at-risk, or otherwise as non-risk. These formulas made it possible to track a given student's performance easily, not only relative to the moving class average in course activities, but also based on the number of submitted assignments and quiz attempts. Figure 2 indicates the process of the failure-proneness prediction model.

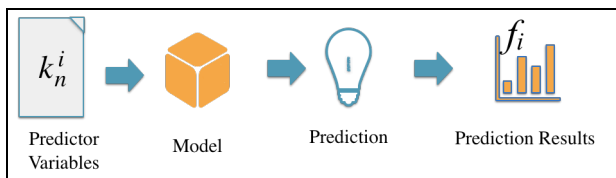


Fig. 2. Failure-proneness prediction model.

Model training and testing was based on the number of weeks, when the extracted variables for each week were accumulated. For the first week, only the variables extracted from the first-week log data were considered. In the same way, in the second week, variables from the log data for the first and second weeks were accumulated. Finally, for the last week, data from the beginning of the semester until the end of the 14th week were accumulated and analyzed. Due to differences in course setup, the authors considered each course separately and divided the dataset for each course into four equal subsets ($k = 4$ -fold cross validation); a random split was used for training and test set formation. Each time, one of the k subsets was used as the test set and the other $k-1$ subsets together formed the training set. Finally, after the average error had been calculated across all k trials, the average accuracy metric for the four test sets was obtained. Table 2 gives the number of

samples for every fold, and Table 3 provides a summary of the course structure for both courses.

Table 2. Number of samples for training and validation

Fold #	OS: n=106		SAD: n=62	
	Test set	Training set	Test set	Training set
Fold1	26	80	15	47
Fold2	26	80	15	47
Fold3	27	79	16	46
Fold4	27	79	16	46

Table 3. Summary of course structure for both courses

Courses	Course Type	# of Credits	Implementation Type
SAD	Compulsory Sophomore (2nd year)	3	Conducted as blended learning (BL) for one department and taught by one lecturer.
OS	Compulsory Junior (3rd year)	3	Conducted as BL for two departments and traditional learning (TL) for one department, taught by one lecturer

There were no major differences between the structures of the two courses. However, the difference in course implementation type was considerable. The OS course was conducted in three different departments of the computer science faculty at KPU, whereas the SAD course was conducted in only one department. This could have been the reason that the authors considered each course separately to identify the factors influencing student outcomes.

4 Results

Based on overall experimental results, there was a statistically significant ($p < 0.001$) correlation with student academic performance based on online activities in the LMS for both courses (Table 4). Most of the students who regularly participated in online activities (with a high access pattern) obtained better scores ($(M=75.8, sd=9.3)$ for the SAD course and $(M=65.9, sd=6.4)$ for OS) compared to inactive students ($(M=49.2, sd=11.4)$ for SAD and $(M=44.5, sd=13.9)$ for OS). To better identify failure-prone students through log patterns, the authors performed an individual pattern comparison for each student, as well as an overall log pattern comparison for each group of students (failure-prone and non-risk). The line plot in Fig. 3 shows a whole-semester daily login frequency comparison of non-risk (left) and failure-prone (right) students in the SAD course, and Fig. 4 does the same for the OS course. As shown in both figures, from week to week, the login frequency increased almost constantly during weeks 4 to 8, particularly for the SAD course, but the values on the vertical axis went down for week 8 (the fourth week of October), which was the mid-term exam for the semester.

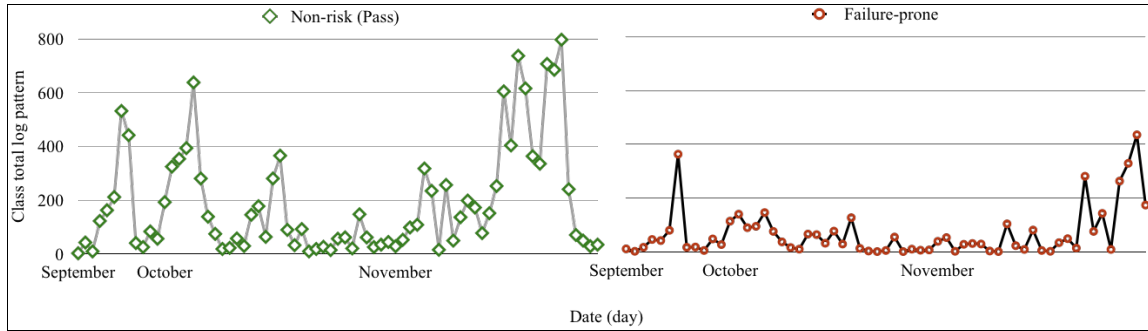


Fig. 3. All-semester access patterns for passing and failing students in the SAD class.

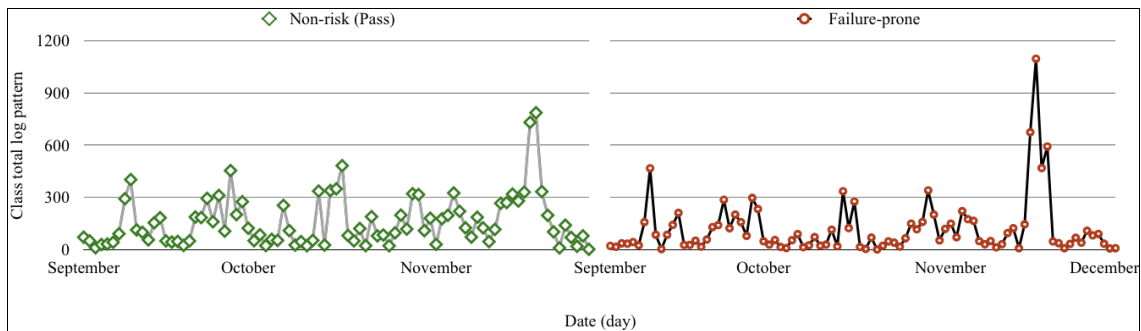


Fig. 4. All-semester access patterns for passing and failing students in the OS class.

The findings seem to indicate a very significant difference between non-risk and failure-prone students' access log patterns for the SAD course, with slightly lesser differences for the OS students' overall log patterns. This could have occurred because of differences in course setting and/or in the contribution of online activities to final student scores. As shown in Table 3, the OS course was conducted for three departments and taught by one lecturer, whereas the SAD course was conducted for one department and managed by one lecturer. This difference (class size) in course implementation could have had a major impact on the quality of teaching and learning. However, teaching methodology can be an appropriate solution to maintain quality and achieve course objectives. Therefore, it can be concluded that class size along with teaching

methodology may explain the better performance of SAD students compared to OS students.

Generally, non-risk students who performed better in the courses showed a more frequent access pattern ($p < .001$) during the semester. Their academic outcomes were positively correlated with their access profile ($M=347, sd=100.2$ for SAD and $M=297.6, sd=129.3$ for OS) compared to failure-prone students ($M=160.8, sd=153.2$ for SAD and $M=180.0, sd=92.8$ for OS). Furthermore, the individual access logs revealed that the non-risk profile characteristics were similar to the overall non-risk class and that the individual failure-prone student log access profiles resembled the overall failure-prone class for both courses. Figure 5 shows a comparison of the individual log patterns of non-risk and failure-prone students.

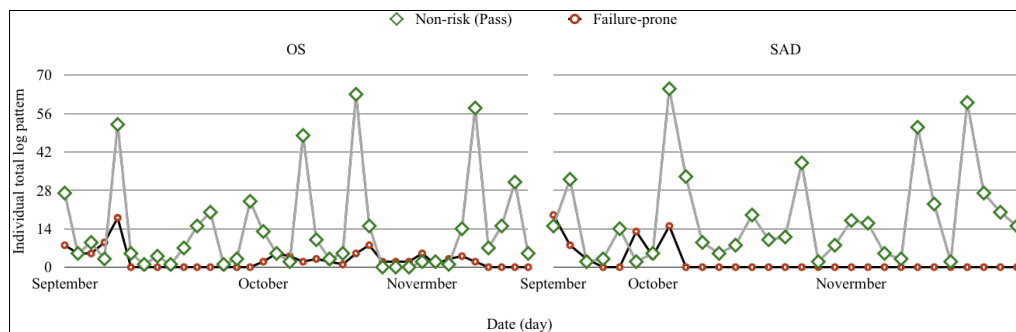


Fig. 5. Individual profiles of non-risk and failure-prone students for both courses.

As shown in Fig. 5, “failure-prone” students did not regularly participate in online activities, and therefore their access patterns were not constantly compared during the semester to those of peers who did perform

well. It is also obvious from Fig. 5 that relatively early, during weeks 3 to 6, students can be tracked and individuals singled out for support. However, there were some students with frequent, but fluctuating access patterns, and these students could easily be identified

relatively early (during the third to sixth weeks) as posing an academic risk. Figure 6 shows a comparison of failure-prone students with high and low access patterns.

Similarly, the authors considered certain other factors (including former grade point average (FGPA), semester grade point average (SGPA), and Kankor score) beyond login frequency to gain more insight into their influence on overall student performance. For this part of the analysis, the authors used *Welch's t-test* to compare the

two groups of students (failure-prone and non-risk) and determine the differences in their outcomes. As shown in Table 4, Welch's *t-test* gave a p -value <0.01 for all factors except the last one (Kankor). The results showed a significant difference between failure-prone and non-risk students for both courses. This proved that among the other factors beyond login frequency, the past performance of students could have had a significant effect on their final outcomes.

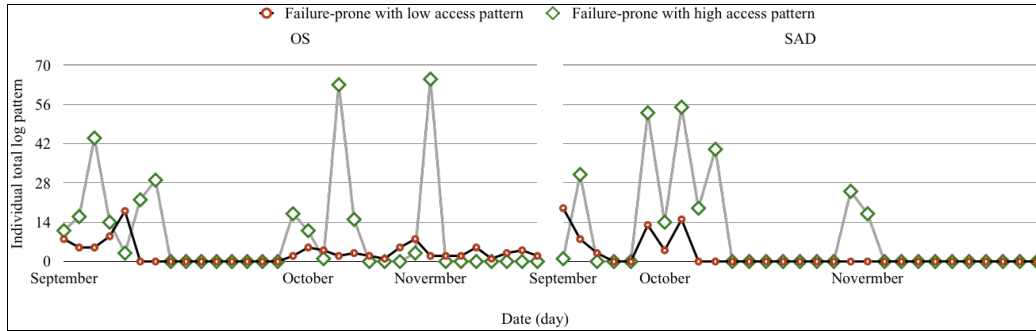


Fig. 6. Comparison of individual profiles of failure-prone students with high and low access patterns.

As shown in Fig. 6, students with high and low access pattern could be identified very early, with similar

at-risk characteristics; hence, it would be possible to provide support during the semester.

Table 4. Comparative results between failure-prone and non-risk students for both courses

Factors [Mean (sd)]	OS		<i>p</i> -value	SAD		<i>p</i> -value
	Failure-prone	Non-risk		Failure-prone	Non-risk	
Total access log	180.0 (92.8)	297.6 (129.3)	≤ 0.01	160.8 (153.2)	347 (100.2)	≤ 0.01
Final results	44.5 (13.9)	65.9 (6.4)	≤ 0.01	49.2 (11.4)	75.8 (9.3)	≤ 0.01
FGPA	1.4 (0.4)	2.1 (0.7)	≤ 0.01	1.8 (0.4)	2.2 (0.6)	≤ 0.01
SGPA	1.5 (0.3)	2.2 (0.6)	≤ 0.01	1.4 (0.4)	2.04 (0.6)	≤ 0.01
Kankor	285.4 (15.9)	282.3 (15.7)	0.3	286.9 (12.1)	289.0 (11.3)	0.5

In conclusion, the study results indicated that log patterns could identify similar characteristics of failure-prone students who may be candidates for failure or dropout; hence, it was possible to provide support after they have given up.

In addition, the results indicated a significant difference in students' effort to solve problems, especially for the OS course. As shown in Fig. 4, despite the weak difference between the access profiles of the two groups of students (failure-prone and non-risk), there was still a large difference in student problem-solving attempts.

For better model accuracy and predictive ability, the authors used two accuracy metrics, accuracy and sensitivity. The overall accuracy metrics were calculated by means of Eqs. (3) and (4):

a) Accuracy

- Indicates the number of successful predictions and can be calculated using Eq. (3):

$$accuracy = \frac{TP+TN}{N} \quad (3)$$

The itemized list below describes the variables used in the equations:

- TP*- (True Positive): predicted as failure-prone and actually do fail
- TN*- (True Negative): predicted as non-risk and are actually non-risk
- FP*- (False Positive): predicted as failure-prone, but are actually non-risk
- FN*- (False Negative): predicted as non-risk, but are not actually non-risk
- N*: number of samples.

b) Sensitivity

- Indicates the accuracy of determination of failure-prone students and is calculated as in Eq. (4):

$$sensitivity = \frac{TP}{TP+FN} \quad (4)$$

The overall results of the weekly classification of students are summarized in Table 5.

Table 5. Classification percentages and classifier accuracy for both courses

Week	SAD			OS		
	% of class posing academic risk	Accuracy	Sensitivity	% of class posing academic risk	Accuracy	Sensitivity
3, 4	26	0.63	0.73	47	0.58	0.83
5	31	0.6	0.86	50	0.62	0.88
6	31	0.69	0.86	48	0.67	0.85
7	31	0.66	0.86	47	0.69	0.83
9, 10	31	0.74	0.86	45	0.65	0.8
11	32	0.76	0.87	45	0.72	0.8
12	31	0.74	0.86	44	0.72	0.78
13, 14	29	0.79	0.82	52	0.79	0.92

As shown in Table 5, despite the small dataset, the overall accuracy and sensitivity for both courses were good enough. For the SAD course, an accuracy of 63% was acquired in the first week of the semester and reached 79% in the 14th week. Similarly, for the OS course, an overall accuracy of 58% was acquired in the first week and gradually reached 79% in the last week of the semester. High accuracy alone cannot evaluate predictive ability or demonstrate the goodness of the evaluated method because it includes true positive and true negative rates. Therefore, the sensitivity metric was also examined to evaluate model performance. The sensitivity prediction for both courses was promising, and the overall classification results were considered better, with high accuracy and sensitivity.

In conclusion, the models correctly predicted 88% of the class for OS and 86% of the class for SAD as having academic risk after one-third of the course period had been completed. This result is in accordance with the findings of similar studies except that the maximum was attained relatively earlier in this study, as well as differences in the classifier used.

5 Discussion

The authors found that an early prediction of failure-prone or dropout students in LMS was possible through the students' log patterns. Analysis showed significant ($p < 0.001$) differences in students' patterns of activity, which were large enough to identify similar characteristics among failure-prone students who may be candidates for failure or dropout. In addition, based on experimental results, the proposed models correctly predicted 88% of the failing students for the junior class and 86% for the sophomore class before the end of the second academic month (the fifth week). This opens up the possibility of early intervention with failure-prone students and of working to give them the ability to change their behavior and improve their chances of academic success. This could be a great opportunity for actors in the education system to minimize the percentages of failing students and increase retention rates by providing the required feedback and academic support to a group of at-risk students. Furthermore, this finding provides a basis for additional exploration and

research to be performed with other small datasets and fewer historical data points.

Similarly, the study revealed a significant ($p < 0.01$) difference between failure-prone and non-risk students' past performance using Welch's *t*-test. This proved that beyond login frequency, students' past performance could also be a solid indicator that influences and was clearly associated with their final outcomes. In addition, this study found that class size and teaching methodology may also play an effective role in student performance. In contrast, the study found no significant difference between the Kankor results of failure-prone and non-risk students. This information could be valuable for educational actors to examine the factors influencing student performance. Furthermore, the findings of this study point to the importance of LA and monitoring students' online activities, which may improve the academic experience and learners' degree of engagement. Monitoring of learning patterns could also help lecturers tailor-make a syllabus for a set of learners according to their capabilities and needs.

6 Conclusions

In conclusion, the experimental results demonstrated that indeed there were patterns hidden within log activity that indicated the likelihood of failure-prone or at-risk students. The models proposed here predicted 88% (sensitivity metric) for the junior class and 86% for the sophomore class as having academic risk just before the end of the second academic month (the fifth week), at which point early intervention was still possible. In addition, the results clearly pointed out differences in terms of student engagement and the pedagogical aspects of online activities. Therefore, it was concluded that the online activities used for the courses were generally well-received by the non-risk students, but not necessarily by the failure-prone students. Likewise, these results could not only provide a great opportunity to identify failure-prone students at an early stage of the semester, but they could also be useful and effective in minimizing the number of failures and creating a better learning and teaching environment.

In further research, the authors plan to incorporate such data to predict student performance in real time

using supervised and unsupervised methods to move towards achieving better accuracy and reaching the potential of LA in optimizing teaching and learning. Similarly, in future work, it is recommended to extend the method to increase accuracy and balance the overall results by detecting high-performing students.

Acknowledgments

Part of this work was supported by a Grant-in-Aid for Scientific Research [19K12251 and 15H02795]. The authors would like to thank the JICA PEACE project for providing a chance to conduct this research.

References

1. N. Sclater, *Learning Analytics Explained*. Florence: Taylor and Francis, pp.35-43, New York: Routledge (2017).
2. B. Wong, Learning analytics in higher education: an analysis of case studies, AAOUJ, pp. 21-40 (2017).
3. H. Sokout, and T. Usagawa, Analyzing the Current Situation of E-learning at Kabul Polytechnic University. In *Proceedings of the 2nd International Conference on Education and Multimedia Technology*, ACM, Okinawa, Japan, (2018). DOI: [10.1145/3206129.3239428](https://doi.org/10.1145/3206129.3239428).
4. H. Sokout, N. Ramaki, & T. Usagawa, Prospects of learning analytics for higher education in developing countries. A case of Kabul Polytechnic University, Afghanistan. In *Proceedings of IEICE Technical Report, Educational Technology* (pp. 43–48), Nagaoka, (2018).
5. U. Upasana, *How to handle Imbalanced classification problems in machine learning*, [online] *Analytics Vidhya*. Retrieved from <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>, (2019).
6. G. Lina, Data sampling improvement by developing SMOTE technique in SAS. Paper 3483 (2015).
7. M. Qais Sherzai, Students' General Results, Kabul, (2020).
8. H. Owen, Y. Anna, C. Jeff, J. Q. Albert, H. Ogata, and J. H. Yang, Applying Learning Analytics for the Early Predicting of Students' Academic Performance in Blended Learning, *ET&S*, **21** (2018).
9. A. FUNGAI, and T. Usagawa, Isolating hidden recurring patterns in unlabeled access log data on LMS to identify dropout risk students, In *The 11th International Student Conference on Advanced Science and Technology*, Kumamoto, Japan, (2016).
10. C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, and S. Ventura, Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, Vol. 21, No. 5, pp. 135–146 (2013).
11. O. Sukhbaatar and T. Usagawa, A prediction of failure-prone students in web-based educational environment, *TCET*, **119**, pp. 17-20 (2019).
12. J.D. Milne, L.M. Jeffrey, G. Suddaby, and A. Higgins. Early identification of students at risk of failing. In M. Brown, M. Hartnett, & T. Stewart (Eds.), *Future challenges, sustainable futures. Proceedings Ascilite Wellington 2012*, pp. 657–661 (2012).