# Using descriptive and predictive learning analytics to understand student behavior at LMS Moodle

Dijana Oreški[1]

[1] *University of Zagreb, Faculty of Organization and Informatics, Pavlinska 2, 42000 Varaždin, Croatia*

**Abstract**

Learning analytics is a data-centric field that applies machine learning algorithms in the educational domain to analyze e-Learning environment data. This study employs descriptive and predictive learning analytics approaches in order to develop descriptive and predictive models of student behavior and success. Cluster analysis, unsupervised machine learning algorithm, and decision tree, supervised machine learning algorithm, are applied on the data from a Learning Management System (LMS) Moodle. Research results indicated: (i) groups of students with similar patterns in behavior at LMS, (ii) student activities at LMS that lead to successful course completion. Such results serve as guidelines for teachers when developing courses and students when enrolling in the course. Descriptive and predictive learning analytics is an innovative approach in education that can enhance teachers and students and improve learning outcomes.

**Keywords**

Learning analytics, educational data mining, LMS data, machine learning.

## 1. Introduction

Implementation of learning management systems (LMS) has grown exponentially during COVID-19 crisis. There are numerous LMS-s developed, such as Moodle, Edmodo and Blackboard. These systems generate a huge amount of data about students' activities. Such data are a valuable source of information for students, teachers and faculty management. In order to obtain knowledge from raw data, data mining should be conducted on structured data from LMS. The use of data mining is shown to give promising results in this area [1]. Educational data mining and learning analytics are subfields designed especially for knowledge extraction from educational data. Both fields are focused on detecting patterns in educational datasets. A list of data mining and learning analytics tasks includes statistics, clustering, classification, prediction or subgroup discovery, among others [2]. In this paper clustering and prediction are combined on the Moodle LMS data set. Previous research papers showed that Moodle is LMS which is mostly used [3], [4].

The data about students' activities at the LMS invokes numerous questions regarding prediction. In this paper, we present a study on the effectiveness of descriptive and predictive learning analytics to describe student activities at LMS and predict students' success in the course. Whereas various works have used these techniques to analyze students' LMS activity, our study combines two approaches.

This paper is structured as follows. Section 2 reviews related work on the given topic. Section 3 explains data and methodology. Section 4 gives research results along with their explanation and interpretation. Section 5 concludes the paper, discusses research limitations, and gives guidelines for future research in this domain.

Proceedings 13th International Conference on eLearning (eLearning-2022), September 29–30, 2022, Belgrade, Serbia
EMAIL: dijana.oreski@foi.hr
ORCID: 0000-0002-3820-0126 (A. 1)

## 2. Related work

Students` success prediction was the subject of numerous research papers so far. Recent work in the educational scientific community strives to exploit the potential of learning management system data to develop accurate and reliable prediction models. Hereinafter, we provide an overview of existing research papers close to our approach.

In the recently published study of Feldman-Maggor et. al. [5] focus was on the characterization of students based on their learning patterns and the authors strived to identify indicators for students' success prediction in an online environment. On the data from undergraduate online chemistry courses, logistic regression and a decision tree algorithm were applied. Assignment submission and the students' video viewing are the most significant predictors. The authors emphasized the importance of students' choices they make regarding their learning process.

Gašević et. al. [6] presume that instructional conditions influence the prediction of academic success. They performed research in nine undergraduate courses offered in a blended learning model. The study illustrates the differences in predictive power and significant predictors between course-specific models and generalized predictive models. The results suggest that it is imperative for learning analytics research to account for the diverse ways technology is adopted and applied in different courses and in different domains. The authors conclude that differences in how students use the learning management system require further research examinations.

Costa et. al. [7] evaluated various educational data mining techniques for the prediction of students' failure in programming courses. Authors strived to investigate the effectiveness of these techniques in the prediction of students failing to take actions that decrease the failure rate. They evaluated four techniques (Naive Bayes, decision tree, neural network and Support Vector Machines) on two data sets from programming courses at Brazilian Public University. One dataset is from distance education and the second one is from on-campus.

Cerezo et. al. [8] examined students' learning process patterns using Moodle logs data. The authors grouped students according to similar behaviors regarding effort and time spent working. Different patterns of students' behavior at the LMS were clustered and the relationships between patterns and students' grades were analyzed. 140 undergraduate students enrolled in Moodle 2.0 course were included in the research. Results indicated variables predicting students` results. Their results could serve as a basis for the improvement of students' achievement in LMS.

Conijin et al. [9] analyzed blended courses in one institution using LMS Moodle. The authors predicted student achievement from LMS variables using multi-level and standard regressions. Their results showed that predictors vary significantly across different courses. The generalization of such prediction models is low.

Macfadyen et al. [10] used LMS tracking data to explore which student online activities could predict academic success. Their analysis from a Blackboard Vista identified variables correlated with student grades. Regression was applied in data analysis resulting in a predictive model for this course which identified variables such as a number of messages posted and number of completed assessments as variables explaining the most variation in student grades. The logistic regression approach showed an accuracy of 81%. These results are useful for the extraction and visualization of LMS data on student engagement and the likelihood of success.

Matcha et. al. investigated students` learning strategies [11]. Among others, clustering was used to detect and interpret learning tactics and strategies. Recently, Saqr and Lopez-Penas [12] examined online engagement by applying clustering to reveal the clusters of students' learning strategies and engagement patterns in the courses.

Based upon previous research papers, hereinafter, we are combining descriptive and predictive analysis of LMS Moodle log data from one course taught through two generations of information technology students.

## 3. Research methodology

The goals of the research are:

(i) to identify the most important predictors of students' success among LMS Moodle logs data,
(ii) to group students of similar LMS behavior patterns,
(iii) to identify the relationship between students' clusters and student's success

To achieve the goals of this research, we address the following research questions:

(i) Which of the variables extracted from the LMS Moodle logs have the highest impact on the student's performance?
(ii) Can we create good student clusters based on their usage of the LMS?
(iii) Is there any correlation between students' clusters and students' success?

### 3.1. Data description

Data are collected from the course Knowledge discovery in data taught at the University of Zagreb, Faculty of Organization and Informatics. Dataset consists of two generations of students. The course was taught as an elective at the undergraduate study level. It was implemented as a blended course: lectures and lab exercises were held in the classroom combined with LMS Moodle usage. Data were extracted from Moodle for two generations of students, thus creating a sample of 83 students. Extracted variables considered a number of logs at specific resource and activity: File, Forum, Student Report, Folder, Choice, File submission, Overview report, Page, System, Test, and Assignment. Overall grade at the course was included in data analysis as a dependent variable.

### 3.2. Machine learning algorithms

Two machine learning algorithms were applied in data analysis: unsupervised machine learning algorithm cluster analysis and supervised machine learning algorithm decision tree. Clustering is an unsupervised machine learning algorithm used for grouping objects into clusters of similar objects [13]. In the e-learning context, clustering has been used for: finding clusters of students with similar learning characteristics [14]; discovering patterns reflecting user behaviors [15], or grouping students according to their characteristics to give them personalized learning approaches [16]. K-means clustering algorithm is applied because it is the mostly used clustering algorithm [17].

A decision tree is a supervised machine learning algorithm belonging to the information-based group of algorithms. Those algorithms develop predictive models by determining the most informative attributes for the prediction of a given task. In the e-learning context, classification and prediction have been used for: predicting students` success in the course [18], predicting students' performance and their final grade [19] as well as predicting the students' achievement along with discovering the relevance of the attributes [20].

For the prediction of student success based on the LMS activity, in this research supervised machine learning algorithm of the decision tree is used. Decision tree approach is simple to understand and interpret the results. Furthermore, previous research papers shown decision tree superiority when comparing with other approaches [21]. Finally, we study a model to analyze the influence of the LMS variables on the student's final grade in the course. Students` clusters are built within the same dataset, and finally, we investigate clusters and their correlation with students' final grades.

## 4. Research results

Data analysis consists of two parts. First, descriptive models were developed followed by predictive models. Table 1 reports CCC value (Cubic Clustering Criterion) for three different groups: 2, 3 and 4. Three clusters are optimal for a given dataset since CCC value is the lowest.

**Table 1. Descriptive model evaluation**

| Number of clusters | CCC |
|---|---|
| 2 | -3.802982345 |
| 3 | -1.666896309 |
| 4 | -2.510022785 |

Three groups of students are identified. Cluster 1 consists of 42 students, cluster 2 of 4 students, and cluster 3 of 37 students. Characteristics of each group are given in tables 2 and 3 as mean values for the groups.

**Table 2.** Students clusters characteristics

| Cluster | File | Forum | Student report | Folder | Choice |
|---|---|---|---|---|---|
| Cluster 1 | 43 | 39 | 6 | 3 | 5 |
| Cluster 2 | 101 | 130 | 14 | 8 | 6 |
| Cluster 3 | 60 | 92 | 16 | 6 | 5 |

**Table 3.** Students clusters characteristics

| Cluster | File submission | Overview report | Page | System | Test | Assignment |
|---|---|---|---|---|---|---|
| Cluster 1 | 5 | 6 | 4 | 76 | 19 | 30 |
| Cluster 2 | 24 | 7 | 9 | 387 | 33 | 122 |
| Cluster 3 | 7 | 9 | 7 | 147 | 26 | 49 |

Cluster 1 consists of students with the lowest overall activity at Moodle. Average view values for all resources and activities except files are the lowest. Cluster 2 consists of students with the highest overall activity at Moodle. Average view values for all resources and activities except student reports and view reports are the highest. Student report and view report are the highest for students in cluster 3, whereas other values are between clusters 1 and 2.

Following descriptive models, predictive models were developed for clusters 1 and 3. A decision tree algorithm is applied for predictive models' development. Evaluation of two models is given in table 4 in terms of RSquare values (metric for model reliability) and RASE (metric for model error).

**Table 4.** Predictive model evaluation

| Model | RSquare | RASE |
|---|---|---|
| A predictive model for Cluster 1 | 0.547 | 0.969 |
| A predictive model for Cluster 3 | 0.425 | 0.973 |

Both models indicate good quality results. Results interpretability assumes conducting sensitivity analysis. Figures 1 and 2 show sensitivity analysis results for models of cluster 1 (figure 1) and cluster 3 (figure 2).

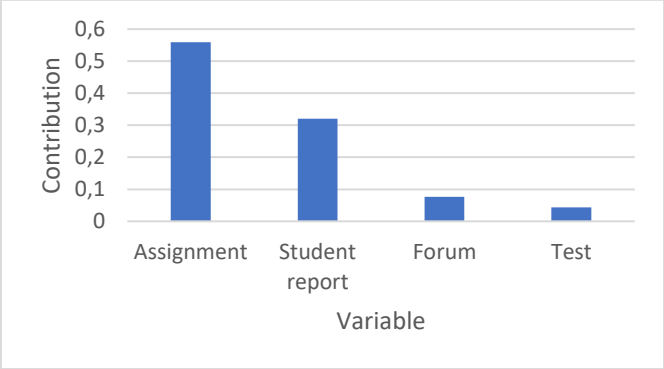


**Figure 1**: Sensitivity analysis for Cluster 1

Assignments and student report views are the most important predictors of grade for cluster 1. Choice and test logs are the most important predictors of grade for cluster 3.

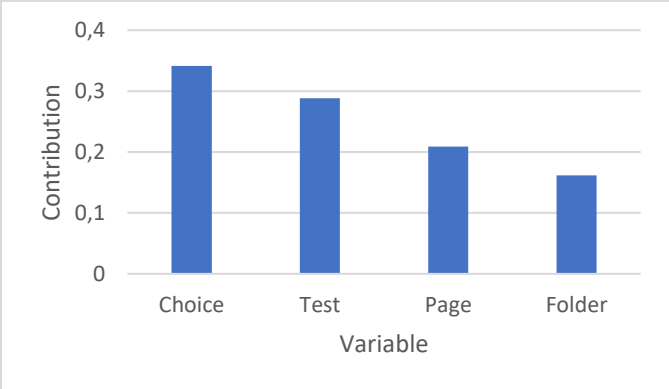


**Figure 2**: Sensitivity analysis for Cluster 3

We can see that there are differences in student success predictors among groups of students with different levels of activities at LMS. That's why it is justified to combine descriptive and predictive learning analytics approaches and hybrid approaches to provide complete information.

One of the advantages of decision tree application is that decision tree results can be presented in the form of rules. Prediction rules for cluster 1 are given in table 5 and prediction rules for cluster 3 in table 6.

**Table 5.** Prediction rules for Cluster 1

| Leaf report | Average grade |
|---|---|
| Assignment<31&Student report<4 | 1.5 |
| Assignment<31&Student report>=4&Assignment>=26 | 2.2857142857 |
| Assignment<31&Student report>=4&Assignment<26&Assignment<21 | 3.4 |
| Assignment<31&Student report>=4&Assignment<26&Assignment>=21 | 4.2 |
| Assignment>=31&Forum>=17&Test<22 | 3.1666666667 |
| Assignment>=31&Forum>=17&Test>=22 | 4 |
| Assignment>=31&Forum<17 | 4.6 |

**Table 7.** Prediction rules for Cluster 3

| Leaf report | Average grade |
|---|---|
| Choice<2 | 1.6 |
| Choice>=2&Page>=7&Folder<7 | 2 |
| Choice>=2&Page>=7&Folder>=7 | 3.2 |
| Choice>=2&Page<7&Test<23 | 2.4 |
| Choice>=2&Page<7&Test>=23&Test>=28 | 3.375 |
| Choice>=2&Page<7&Test>=23&Test<28 | 4 |
| Choice<2 | 1.6 |
| Choice>=2&Page>=7&Folder<7 | 2 |

To answer the third research question, correlation analysis was performed. The correlation coefficient of $r = 0.2037$ indicates there is no relationship between students' grades and grouping based on Moodle activity.

## 5. Conclusion

Earlier research papers have demonstrated that higher education institutions could use the predictive power of LMS data in combination with machine learning algorithms to develop models' tools that identify successful students and at-risk students and allow interventions. In this paper, we have proved that a combination of unsupervised and supervised machine learning algorithms on LMS data results in useful models for explaining and predicting students' behavior at LMS.

To achieve the goals of this research, we have answered the following research questions:

(i) Which of the variables extracted from the LMS Moodle logs have the highest impact on the student's performance?

Assignments and student report views have the highest impact on grades for students with lower LMS activity. Choice and test logs have the highest impact on grades for students with higher LMS activity.

(ii) Can we create good student clusters based on their usage of the LMS?
Based on the CCC value, we can conclude that good student clusters are created.

(iii) Is there any correlation between students' clusters and students' success?
There is no correlation between students' clusters and students' success.

Research results contribute to the personalization of learning and teaching approach, especially for online environment. In future research, we will upgrade the study with the following aspects. First, more courses will be included in the analysis, and students of different study programs. Secondly, different machine learning algorithms will be applied to the LMS data and compared to see if are there differences in performance between different algorithms.

## 6. References


[1] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, Jan. 2013, doi: 10.1002/WIDM.1075.

[2] L. Mainon, Oded; Rokach, *Data mining and knowledge discovery handbook*. .

[3] B. Alexander and B. Alexander, "Web 2.0: A New Wave of Innovation for Teaching and Learning?," *Educ. Rev.*, vol. 41, no. 2, pp. 33–34, 2006.

[4] N. Cavus and A. M. Momani, "Computer aided evaluation of learning management systems," *Procedia - Soc. Behav. Sci.*, vol. 1, no. 1, pp. 426–430, Jan. 2009, doi: 10.1016/J.SBSPRO.2009.01.076.

[5] Y. Feldman-Maggor, R. Blonder, and I. Tuvi-Arad, "Let them choose: Optional assignments and online learning patterns as predictors of success in online general chemistry courses," *Internet High. Educ.*, vol. 55, p. 100867, Oct. 2022, doi: 10.1016/J.IHEDUC.2022.100867.

[6] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *Internet High. Educ.*, vol. 28, pp. 68–84, Jan. 2016, doi: 10.1016/J.IHEDUC.2015.10.002.

[7] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, Aug. 2017, doi: 10.1016/J.CHB.2017.01.047.

[8] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education," *Comput. Educ.*, vol. 96, pp. 42–54, May 2016, doi: 10.1016/J.COMPEDU.2016.02.006.

[9] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, Jan. 2017, doi: 10.1109/TLT.2016.2616312.

[10] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: A proof of concept," *Comput. Educ.*, vol. 54, no. 2, pp. 588–599, 2010, doi: 10.1016/j.compedu.2009.09.008.

[11] W. Matcha *et al.*, "Analytics of learning strategies: Associations with academic performance and feedback," *ACM Int. Conf. Proceeding Ser.*, pp. 461–470, Mar. 2019, doi: 10.1145/3303772.3303787.

[12] M. Saqr and S. López-Pernas, "The longitudinal trajectories of online engagement over a full program," *Comput. Educ.*, vol. 175, p. 104325, Dec. 2021, doi: 10.1016/J.COMPEDU.2021.104325.

[13] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.

[14] T. Tang, T. Tang, and G. McCalla, "Smart Recommendation for an Evolving E-Learning System: Architecture and...," *Int. J. E-Learning*, vol. 4, no. 1, pp. 105–129, 2005.

[15] E. Gaudioso and L. Talavera, “Data mining to support tutoring in virtual learning communities: experiences and challenges.”
[16] W. Hämäläinen, T. H. Laine, and E. Sutinen, “Data Mining in Personalizing Distance Education Courses.”
[17] Bhise RB, Thorat SS, and Supekar AK, “‘Importance of Data Mining in Higher Education System,’” *IOSR J. Humanit. Soc. Sci. (IOSR-JHSS*, vol. 6, no. 6, pp. 18–21, Accessed: Sep. 20, 2022. [Online]. Available: www.Iosrjournals.Org.
[18] W. Hämäläinen and M. Vinni, “Classifiers for educational data mining.”
[19] B. Minaei-Bidgoli and W. F. Punch, “Using genetic algorithms for data mining optimization in an educational web-based system,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2724, pp. 2252–2263, 2003, doi: 10.1007/3-540-45110-2_119/COVER.
[20] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “PREDICTING STUDENTS’ PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES,” *http://dx.doi.org/10.1080/08839510490442058*, vol. 18, no. 5, pp. 411–426, May 2010, doi: 10.1080/08839510490442058.
[21] G. Aksu and C. Reyhanlioglu Keceoglu, “Comparison of Results Obtained from Logistic Regression, CHAID Analysis and Decision Tree Methods,” *Eurasian J. Educ. Res.*, vol. 19, no. 84, pp. 115–134, Nov. 2019, doi: 10.14689/EJER.2019.84.6.