©2019 Published in 7th International Symposium on Innovative Technologies in Engineering and Science 22-24 November 2019 (ISITES2019 SanliUrfa - Turkey)
https://doi.org/10.33793/acperpro.02.03.112

Early Detection of Drop Outs in E-Learning Systems

*[1]Neslihan Ademi and [2]Suzana Loshkovska
*[1] Faculty of Engineering, Department of Computer Engineering, International Balkan University, Skopje, North Macedonia
[2] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia

Abstract

After the popularity of Learning Management Systems, Data Mining and Learning Analytics have become emerging topics. Learning Management Systems such as Moodle, provide big amount of data to be used in analyzing students' online behavior. This paper represents a method for early detection of drop outs from a Bachelor degree course using data mining methods. Data is collected through Moodle logs. For early detection, event logs till the first exam is taken into consideration. Decision Tree (DT) and Bayesian Network (BN) algorithms are used for the prediction. In the end it is shown that DT algorithm gives a higher over-all accuracy but BN is better for discovering fail cases as it has higher specificity.

Key words: Learning Management Systems, Educational Data Mining, Learning Analytics, Drop outs

1. Introduction

Learning Management Systems (LMSs) are used in both distance and blended learning settings thanks to their capability of managing courses and opportunities such as easy sharing course material, opening forums, communicating with the course participants, preparing assignments and tests, etc. One of the most popular LMSs is Moodle which is free and open. Learning Systems like Moodle can store huge amount of data which opens a door for the Educational Data Mining (EDM). Analyzing this data gives the opportunity for understanding users' behaviors and their characteristic.

Most analyses of log data collected provide a descriptive overview of human behavior. Simply observing behavior at scale provides insights about how people interact with existing systems and services [1]. Recently there are many studies in the literature about log analysis in e-learning environments [2]–[7].

The main objective of this study is to understand user engagement level at the early stages of the course period and to detect drop outs earlier to avoid them. In our previous study [8], complete course period was analyzed. In this study, we consider the period before the first exam up to 7th week of lectures out of 14 weeks of complete lecture in the semester. So from an earlier time period students who have possibility for drop out can be discovered and several precautions can be planned to avoid this situation.

*Corresponding author: Address: Faculty of Engineering, Department of Computer Engineering, International Balkan University, Skopje, North Macedonia. E-mail address: neslihan@ibu.edu.mk

In this study we compare the prediction by using J48 Decision Tree and Bayesian Network Algorithms. Decision Tree algorithm is one of the most commonly used classification approaches in EDM and user modeling [9]. Bayesian networks are also widely used for student modeling in intelligent learning systems [10].

2. Materials and Method

In a previous study[8] , we analyzed online behavior of the students and their engagement levels during the course and we found positive correlations of grade with each of the activities on the learning system. In this study we used the similar pre-processing steps but this time the period analyzed on the logs were limited to 7 weeks and particularly starting from the first week of the lectures and until the first exam week is taken into consideration to be able to detect possible failure of the students. The methodology describing the work flow of the steps used for the case study is given in Figure 1.

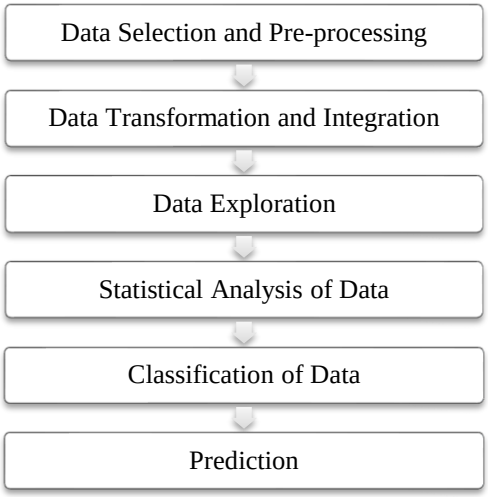

Figure 1 Steps used in data analysis

2.1. Data selection and pre-processing

For this study, log files of a blended course are taken from Moodle which is used at the Faculty of Computer Science and Engineering at the University of Ss. Cyril and Methodius in Skopje, North Macedonia. Log files were extracted in .csv format and they contained all activities of the students from a one semester bachelor's degree course at the academic year 2016-2017. Moodle was used to support classroom teaching to distribute course material, lectures, homework, and laboratory exercises and to provide discussion through the forums. A total of 260 students registered with Moodle for the course. After cleaning and deleting the data with missing information; remaining number of students analyzed in the system was 255.

The standard fields in Moodle log files are: Time, User full name, Affected user, Event context,

Component, Event name, Description, Origin, IP address. The raw data retrieved for the academic year 2016-2017 was composed of 161.007 rows. Separately, another file is used which contained scores of the students from the course.

In the pre-processing step to keep only relevant information, the actions logged by instructors and administrators, log data produced by the system is removed by filtering the data. Filtering is done by using RStudio which an open source software supporting R language [11]. Especially Sqldf [12] package is used as it contains SQL like commands. Data was also filtered according to the given certain period of time which is in between the first week of lectures and first exam week. After filtering the remaining number of rows was 48.883.

2.2. Data transformation and integration

The raw data collected was consist of Time, User full name, Affected user, Event context, Component, Event name, Description, Origin, IP address attributes. Event name contains what kind of activity is performed by the user. By filtering event name Visits, Quizzes, Assignments, FileSubmissions, ForumView and CourseView attributes and how many times each of the activity was performed by the users are extracted from the raw data. After this transformation, the file with the scores of the users is integrated.

2.3. Data exploration and statistical analysis of data

Data Exploration consists of following steps; (1) calculating summary statistics of the data and correlations between the grade and activities, (2) visualizing total visit frequency, distribution of the grades, quizzes, assignments, forum reads and file submissions. The first step is performed by using RStudio, and the second step is performed by using WEKA [13] software.

2.4. Classification of data and prediction

This step contains classification of the data in a supervised manner, as the labels of the grades are certain. They were given in a 10 base system, where grade “5” represents failure, grades “6” to “10” represent pass. The grades are labeled into two classes, such as FAIL and PASS. Decision Tree and Bayesian Network algorithms are used for the classification and prediction with a 10 folds cross validation.

2.5. Evaluation

Evaluation of the classification models are done in terms of True Positive (TP) rate, False Positive (FP) rate, Precision, Recall and F-Measure parameters for the algorithms Decision Tree, Bayesian Network. 10 folds of cross validation is used for the evaluation of the both prediction methods. Evaluation of the methods are also given in terms of Specificity given in Equation1, as it is the measure how effectively a classifier identifies negative labels[14].

$$\text{Specificity} = \frac{TN}{FP+TN} \quad (1)$$

3. Results

Results are divided into two main parts; (1) exploratory data analysis which tries to understand the trends and correlations in data and (2) predictions which shows the accuracies of the used methods.

3.1. Exploratory Data Analysis

Table 1 gives 5-point summary statistics of the data for 7 weeks period. Figure 2 shows the correlation matrix for students' activities and their grade taken from the course. There are positive correlations between the visits, assignments, course views and the grade.

Table 1 Summary statistics of 7 weeks data

	Visits	Assignment	ForumRead	CourseViews	Grade
Min	3.00	0.00	0.00	0.00	5.00
Q1	76.50	3.00	0.00	17.00	5.00
Median	110.00	5.00	1.00	24.00	6.00
Mean	111.60	4.61	1.55	26.62	6.80
Q3	142.50	6.00	2.00	32.00	8.00
Max	463.00	10.00	10.00	115.00	10.00

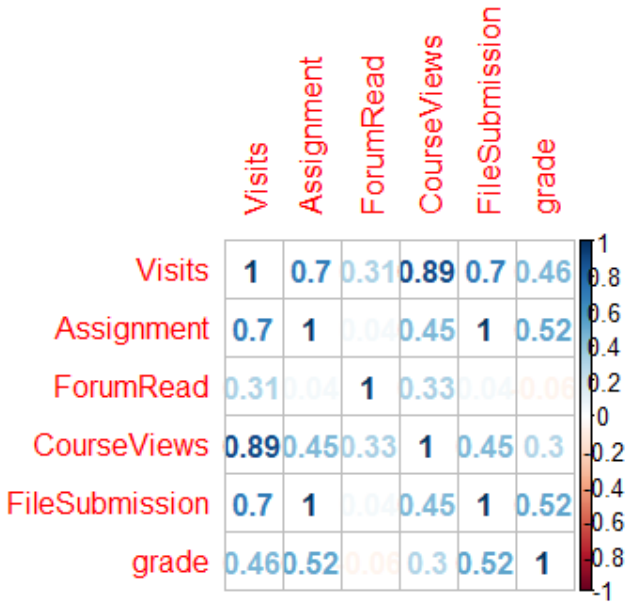

Figure 2 Correlation Analysis of 2016-2017 data

Figure 3 shows the total visit frequency, distribution of the grades, quizzes, assignments, forum reads and file submissions. Where the color red represents PASS and the color blue represents FAIL cases.

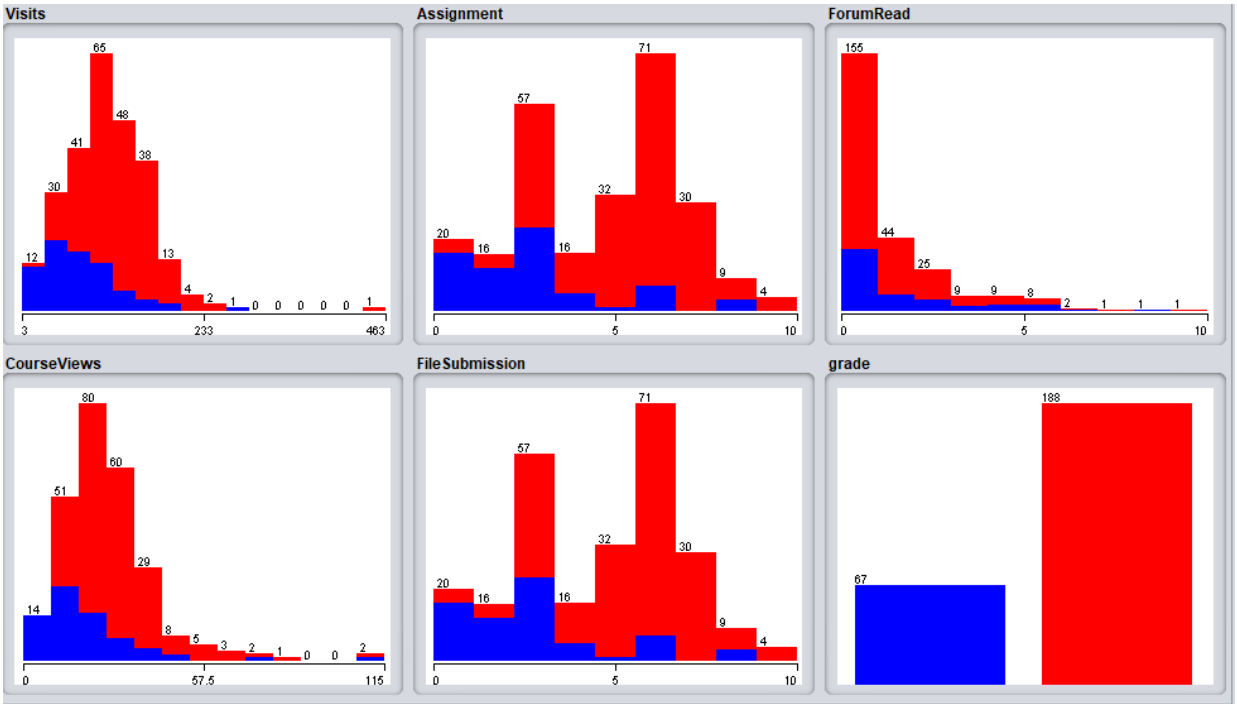

Figure 3 WEKA classification based on two classes PASS and FAIL

3.2. Prediction

Correctly classified instances were 207 out of 255 which means 81.18 % when a Decision Tree (DT) algorithm was used. With Bayesian Network (BN) algorithm correctly classified instances were 190 out of 255 which means 74.51 %. Table 4 shows the confusion matrix for J48 DT algorithm. Table 5 gives the detailed accuracy measure of J48 DT algorithm. In both of the algorithms 10 folds of cross validation is used.

Table 2 Confusion matrix for Decision Tree algorithm

Real /Predicted	PASS	FAIL
PASS	181	7
FAIL	41	26

Table 3 Detailed accuracy by class using J48 Decision Tree

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
FAIL	0.388	0.037	0.788	0.388	0.520
PASS	0.963	0.612	0.815	0.963	0.883
Weighted Avg.	0.812	0.461	0.808	0.812	0.788

Table 6 shows the confusion matrix for BN algorithm, when Table 6 shows detailed accuracy measures of BN algorithm for the prediction.

Table 4 Confusion matrix for Bayesian Network algorithm

Real /Predicted	PASS	FAIL
PASS	149	39
FAIL	26	41

Table 5 Detailed accuracy by class using Bayesian Network

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
FAIL	0.612	0.207	0.513	0.612	0.558
PASS	0.793	0.388	0.851	0.793	0.821
Weighted Avg.	0.745	0.341	0.762	0.745	0.752

4. Discussion

Correlation analysis of the grades with the number of activities shows that there are positive correlations. This allows us to draw an opinion the more engagement with the system the better grade. By following 7 weeks of activities of the students, we saw that the students which are more active are about to pass and get higher grades (Figure 3).

Classification and prediction results show that in terms of correctly classified instances (weighted averages for True Positives in Table 5 and 7), Decision Tree algorithm performs better with 81.18%, when Bayesian Network algorithm performs 74.51%.

Here we are using a binary classification where we have FAIL or PASS options. And the purpose of the study is to discover drop outs; meaning that number of FAIL cases are more significant rather than PASS cases. This is why specificity should be used for the performance evaluation of two classification methods. Specificity is the measure of how effectively a classifier identifies negative labels, in our case FAIL cases. Basically TP rate of FAIL cases in Table 5 and 7 gives the specificity of the two methods. DT has specificity of 0.388 (from Table 5) and BN has 0.612

(from Table 7). In this case we suggest BN to be used for predicting drop outs.

Conclusions

In this paper we define a methodology for the early detection of drop outs in e-learning systems, - basically in Moodle LMS by using Educational Data Mining steps. Decision Tree and Bayesian Network methods are used for classification of the data. It is concluded that overall accuracy of DT is higher than BN method, but as the focus of the study is to discover the drop outs; specificity of predictions should be considered. It is shown that BN performs better in finding the failing students.

In the future different algorithms can be used and their performances in discovering drop puts can be compared. We hope that this study will give insights to the future designs of adaptive learning systems in defining the students under risk of drop outs; so that remedial activities or other engaging activities will be applied before the failure of the students.

References

[1] S. Dumais, R. Jeffries, D. M. Russell, D. Tang, and J. Teevan, "Understanding User Behavior Through Log Data and Analysis," in *Ways of Knowing in HCI*, New York, NY: Springer New York, 2014, pp. 349–372.

[2] G. Akçapınar, "Profiling Students' Approaches to Learning through Moodle Logs," *Proceedings of Multidisciplinary Academic Conference on Education, Teaching and E-learning in Prague 2015, Czech Republic (MAC-ETeL 2015)*, no. December, p. 7, 2015.

[3] A. Konstantinidis and C. Grafton, "Using Excel Macros to Analyse Moodle Logs," *UK Research.Moodle.Net*, no. September, pp. 4–6, 2013.

[4] Á. Figueira and Álvaro, "Mining Moodle Logs for Grade Prediction," in *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM 2017*, 2017, pp. 1–8.

[5] Á. Figueira, "Mining Moodle Logs for Grade Prediction: A Methodology Walk-through," in *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2017, pp. 44:1-44:8.

[6] T. Käser, N. R. Hallinen, and D. L. Schwartz, "Modeling exploration strategies to predict student performance within a learning environment and beyond," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 2017.

[7] M. Cocea and S. Weibelzahl, "Log file analysis for disengagement detection in e-Learning environments," *User Modeling and User-Adapted Interaction*, 2009.

[8] N. Ademi and S. Loshkovska, "Exploratory Analysis of Student Activities and Success Based On Moodle Log Data," in *16th International Conference on Informatics and Information Technologies*, 2019.(in press)

[9] K. Almohammadi, H. Hagras, D. Alghazzawi, and G. Aldabbagh, "A Survey of Artificial Intelligence Techniques Employed for Adaptive Educational Systems Within E-Learning," vol. 7, no. 1, pp. 47–64, 2017.

[10] T. Kaser, S. Klingler, A. G. Schwing, and M. Gross, "Dynamic Bayesian Networks for

Student Modeling," *IEEE Transactions on Learning Technologies*, vol. 10, no. 4, pp. 450–462, Oct. 2017.

[11] "RStudio - RStudio." [Online]. Available: https://rstudio.com/. [Accessed: 13-Oct-2019].

[12] G. Grothendieck, "CRAN - Package sqldf." [Online]. Available: https://cran.r-project.org/web/packages/sqldf/index.html. [Accessed: 12-Jun-2019].

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and W. I. H., "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, 2009.