Prediction of Student Success Through Analysis of Moodle Logs: Case Study

Neslihan Ademi[1](✉), Suzana Loshkovska[2], and Slobodan Kalajdziski[2]

[1] International Balkan University, Skopje, North Macedonia
neslihan@ibu.edu.mk

[2] Ss. Cyril and Methodius University, Skopje, North Macedonia
{suzana.loshkovska,slobodan.kalajdziski}@finki.ukim.mk

Abstract. Data mining together with learning analytics are emerging topics because of the huge amount of educational data coming from learning management systems. This paper presents a case study about students' grade prediction by using data mining methods. Data obtained from Moodle log files are explored to understand the trends and effects of students' activities on Moodle learning management system. Correlations of system activities with the student success are found. Data is classified and modeled by using decision tree, Bayesian Network and Support Vector Machine algorithms. After training the model with a one-year course activity data, next years' grades are predicted. We found that Decision tree classification gives the best accuracy on the test data for the prediction.

Keywords: Data Mining · Learning Analytics · LMS · E-learning · Prediction

1 Introduction

With the rapid development of e-technologies distance learning has gained a bigger role in the education. E-learning systems are used even in the formal education settings very often as they offer instructor great variety of opportunities. Learning management systems (LMSs) are one of the most popular methods as they offer the opportunities to facilitate to distribute information to students, communicate among participants in a course, produce content material, prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas. One of the most commonly used LMS is Moodle (modular object-oriented developmental learning environment), which is a free and open learning management system.

E-learning systems like Moodle contain a huge amount of data related to the students. This data gives a big opportunity to analyze students' behavior and understand the characteristics and behavior of students.

This huge amount of data, in other words Big Data in education emerged two research areas: Educational Data Mining (EDM) and Learning Analytics (LA). EDM is concerned with developing tools for the discovery of patterns in educational data [15]. LA is concerned with the measurement, collection and analysis and reporting of data

S. Gievska and G. Madjarov (Eds.): ICT Innovations 2019, CCIS 1110, pp. 27–40, 2019.
https://doi.org/10.1007/978-3-030-33110-8_3

about learners to understand, to optimize learning and the learning environment [9]. Learning analytics can be used to understand students' behaviors in learning management systems.

Especially log data in Moodle can be used as a basis for creating educational data which gives us the information about students' online behavior. In all LMSs, every student posses' digital profile, and every student can access the LMS by using their personal account. All activities performed by the students are saved in log files. Collected log data provide a descriptive overview of human behavior. Simply observing behavior at scale provides insights about how people interact with existing systems and services [4]. Generally observational log studies contain partitioning log data; by time and by user. Partitioning by time helps us to understand significant temporal features, such as periodicities (including consistent daily, weekly, and yearly patterns) and sharp changes in behavior during important events. It gives an up-to-the-minute picture of how people are behaving with a system from log data by comparing past and current behavior. It is also interesting to partition log data by user characteristics [4].

Recently there are many studies in the literature about log analysis in e-learning environments. In [14], authors explain educational data mining with a case study on Moodle logs for visualization, clustering and classification purposes. Another study about log analysis on e-learning systems demonstrates the disengagement of the students by looking their online behavior [3]. In [11], authors use learning analytics for monitoring students' online participation from Moodle logs and discover frequent navigational patterns by sequential pattern mining techniques. In study [2], clusters are used for profiling the students according to their learning styles as deep learners and surface learners with the help of Moodle log data. The study [7] aims to detect the relationship between the observed variables and students' final grades and to find out the impact of a particular activity in an LMS on the final grade considering also other data such as gender. In [10], authors used Excel macros to analyze and visualize Moodle activities based on metrics such as total page views, unique users, unique actions, IP addresses, unique pages, average session length and bounce rate. The study [8] investigates the impact of students' exploration strategies on learning and proposes a probabilistic model jointly representing student knowledge and strategies based on data collected from an interactive computer-based game. The study [13] analyzes log data obtained from various courses of an university using Moodle platform together with the demographic profiles of students and compares them with their activity level in order to find how these attributes affect students' level of activity. In [12], authors suggest an analytics package which can be integrated with Moodle, to fetch the log files produced from Moodle continuously and produce the results of learning trends of the students. Some studies like in [5], use log data for the grade prediction by using decision trees.

In our study we focus on the engagement of the students by evaluating number of their activities and we make grade prediction upon these activities. The purpose of this study is to predict students' success in terms of grade and also to figure out the students who are about to drop out. Differently from the above mentioned studies, in this paper we use one years' data to predict the next year. The paper presents overall steps of the analysis of Moodle log data by using several data mining methods and tools as a case study. The second section defines the used methodology for the analysis, while the third section gives results and discussion, finally the last section is the conclusion of the paper.

2 Methodology

Generally e-learning data mining process consists of four steps like in the general data mining process: (1) collect data, (2) pre-process the data, (3) apply data mining, (4) interpret, evaluate and deploy the results [14]. When the log files are considered for data mining pre-processing part is more complicated because of the structure of the log files. Log files contain one row for each activity in the system, so for the user based analysis logs should be filtered and the data should be transformed into a data frame based on user and number of each activity of the user. In Fig. 1 we give the methodology describing the work flow of the steps used for the case study.

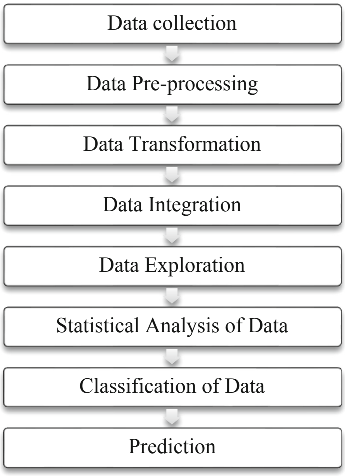


Fig. 1. Flow chart of the used methodology.

2.1 Data Collection

For the study, two log files are taken from Moodle which is installed and used at the Faculty of Computer Science and Engineering at the University of Ss. Cyril and Methodius in Skopje. Log files were extracted in .csv format and they contained all activities of the students from a one semester bachelor's degree course of User Interfaces at two academic years; 2016–2017 and 2017–2018. Teaching process was designed as blended learning. Moodle was used to support classroom teaching to distribute course material, lectures, homework, laboratory exercises and to provide discussion through the forums. A total of 260 students registered with Moodle for the course of User Interfaces for the first year and 206 students for the second year.

The standard retrieved fields in the log files are: Time, User full name, Affected user, Event context, Component, Event name, Description, Origin, IP address. The retrieved data for the academic year 2016–2017 was composed of 161.007 rows, for the academic year 2017–2018 was composed of 165.000 rows each with a filled value in every of the above-mentioned column fields.

Separately, another file is used which contained scores and grades of the students from the course for both academic years.

2.2 Data Pre-processing

Several pre-processing steps are applied to the log files, to keep on relevant and correct information. The first step was to convert all Cyrillic letters found in the original log files into Latin letters, as they were not recognized by R packages. The second step was to remove the actions logged by instructors and administrators selectively by filtering them using *sqldf* package [6], as we want to analyze only the students' actions. Log data produced by the system is also removed by filtering the data where component field is system. Required fields are extracted and duplicate records are removed. The number of remaining rows after filtering is 144.522 and 146.711 for the first and the second file respectively.

As the last step, *userID* and *moduleID* from the description field is filtered so we generated new columns for *userID* to be used instead of user full name to provide anonymity of the data.

2.3 Data Transformation and Integration

For data transformation *sqldf* package which allows complex database queries is used in RStudio. Raw data was consisting of Time, User full name, Affected user, Event context, Component, Event name, Description, Origin, IP address. After filtering and transformation, we created a table with the fields given in Table 1.

Table 1. Attributes after data transformation.

Name	Description
UserID	ID number of the student
Visits	Total number of visits by the student
Quizzes	Number of quizzes taken by the student
Assignments	Number of submitted assignments by the student
ForumCreated	Number of forum creations by the student
ForumView	Number of forum views by the student
CourseView	Number of course views by the student
FileSubmission	Number of file submissions by the student
GradeView	Number of grade views by the student

Transformed data is integrated with the data containing course grades. Table 1 is updated by adding one more attribute "Grade". Grades are in the range of 5–10 where 5 represent failure of the student, while 10 is the highest score.

2.4 Data Exploration and Statistical Analysis

The results of this step are obtained by using the *sqldf* package and the lattice package in R Studio and also with WEKA software.

In a previous study [1], we explored 2016–2017 data in details by visualizing views of lectures for each week, views of lab exercises, total visit frequency, distribution of the grades, quiz, assignment, forum, file submission, grade view frequencies and distribution of weekly visits. Here we just give the summary statistics of both years' data and correlation matrices of visits, quizzes, assignments, forum creations, forum views, file submissions, grade views with the course grade as they show us the similarity of the activities and correlations of both years. Also, we tried to use those summary statistics (5-point summary of data) for classification when labeling the classes. Standard discretization/binning methods in WEKA could not be used, because the data of the next year has different summary statistics. This is why 5 point summary of the data is important for both of the years' data.

Correlations are found by using Pearson correlation test. Equation 1 gives the formula for Pearson correlation coefficient which calculates the correlation between two variables X and Y. The value of the Pearson coefficient is always between –1 and +1, where r = –1 or r = +1 indicates a perfect linear relationship, where sign indicates the direction, while r = 0 indicates no linear relationship.

$$r = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sqrt{\left[\sum (x - \bar{X})^2\right]\left[\sum (y - \bar{Y})^2\right]}} \tag{1}$$

2.5 Classification of Data

Classification is a supervised learning method which uses labels to group the data. The attributes of both years shown in Table 1 are labeled in terms of LOW, MEDIUM, HIGH; or in terms of YES and NO. For the labeling process 5 point summary statistics of the data is used. The attributes Visits, Assignments, Course views and Forum views are labeled as LOW if the value is between min to Q1, MEDIUM if the value is in between Q1-Q3 and HIGH if the value is in between Q3-max, where Q1 represents the first quartile and Q3 represents the third quartile of the data attributes in 5-point summary statistics. The attributes Forum created and Grade view are labeled in terms of YES or NO as their number is very low and their first quartile values are 0. So, all activities of the students are labeled with in their study year.

The grade attribute in the first data file are left in form of numbers 5, 6, 7, 8, 9 and 10 just changed from numeric to nominal by using WEKA's NumericToNominal function so that the text values of the grades are considered. The grade attribute is set as the target attribute to be predicted.

We also tried the model with the labeling of the grades in terms of FAIL, GOOD and EXCELLENT, and in terms of PASS and FAIL to compare the accuracy of the model in different number of labels.

2.6 Prediction

The data from 2016–2017 is used as training data to generate a model and this model is used to predict the grades of the next year. In this step WEKA is used. For the classification, J48 Decision Tree (DT), Bayesian Network (BN) and Support Vector Machine (SVM) algorithms are applied. In all cases the model is trained with 10 fold cross-validation. After training the first data file with above mentioned classification algorithms, it is saved as a model in WEKA; so this model can be loaded and used any time on a test data set. The second file for 2017–2018 data also contained grade attribute. The content of this attribute is changed by a "?" and this file is used as a test set. So, the created model from the first-year data is used to predict the grades of the second year by using DT, BN and SVM. In the end the grades generated by the model and the real grades are compared.

2.7 Evaluation of the Predictions

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 DT, BN and SVM. After the models are created and applied on the test data in WEKA, results are given in form of confusion matrices for the predicted data vs. real data. Confusion matrices contain Recall and Precision parameters. In the end three classification models are compared in terms of overall accuracy and Kappa Index values.

3 Results and Discussion

In the complete process we used two software; RStudio and WEKA which are both open source under GNU. RStudio is a development interface for R language which allows easy manipulation of data. WEKA is useful in terms of its ready interface, which contains machine learning algorithms for data mining tasks.

3.1 Exploratory Data Analysis of Student Activities

Tables 2 and 3 gives the summary statistics of data in years 2016–2017 and 2017–2018 respectively. Summary statistics contain 5-point summary of the data which is minimum value, first quartile, mean, third quartile and maximum value; and additionally median value.

Table 2. Summary statistics of 2016–2017 data.

	Visits	Quizzes	Assignment	Forum created	Forum view	Course view	File submission	Grade view	Grade
Min	1	0	0	0	0	0	0	0	5
Q1	206.8	1	7	0	3	45.75	7	0	5
Median	281.5	2	11	0	7	65.50	11	1	6
Mean	301.9	1.938	10.87	0.1769	11.7	71.8	10.87	2.85	6.77
Q3	390.2	3	15	0	14	93.50	15.00	3	8
Max	1188	5	25	5	77	256	25.00	74	10

Table 3. Summary statistics of 2017–2018 data.

	Visits	Quizzes	Assignment	Forum created	Forum view	Course view	File submission	Grade view	Grade
Min	1.0	0	0	0	0	0	0	0	5
Q1	258.2	7	7	0	2	66.25	7	0	6
Median	362.5	11	12	0	5	95.00	12	1	4
Mean	402.5	9.53	11.35	0.01	8.28	106.03	11.35	3.64	7.3
Q3	494.8	13	16	0	11	127.75	16	4	9
Max	2853	15	31	1	63	531.00	31	52	10

Figure 2 shows the activities of the students in terms of grades for the year 2016–2017. To generate this graph grades are grouped in terms of PASS and FAIL. The blue color represents the failed students while the red color represents passed students. Figure 3 shows the data when grades are classified in terms of numbers from 5 to 10.

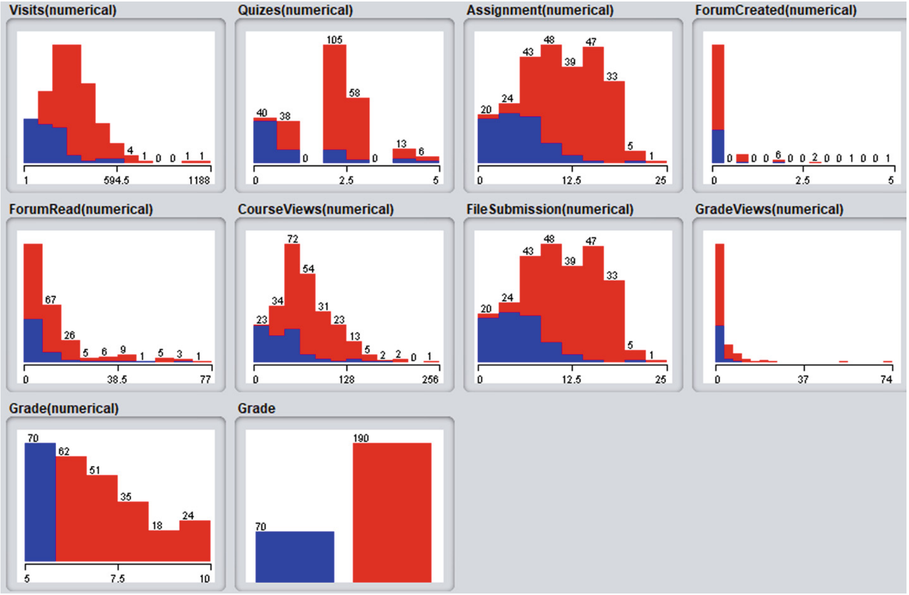


Fig. 2. WEKA output of 2016–2017 data with two classes (Pass/Fail).

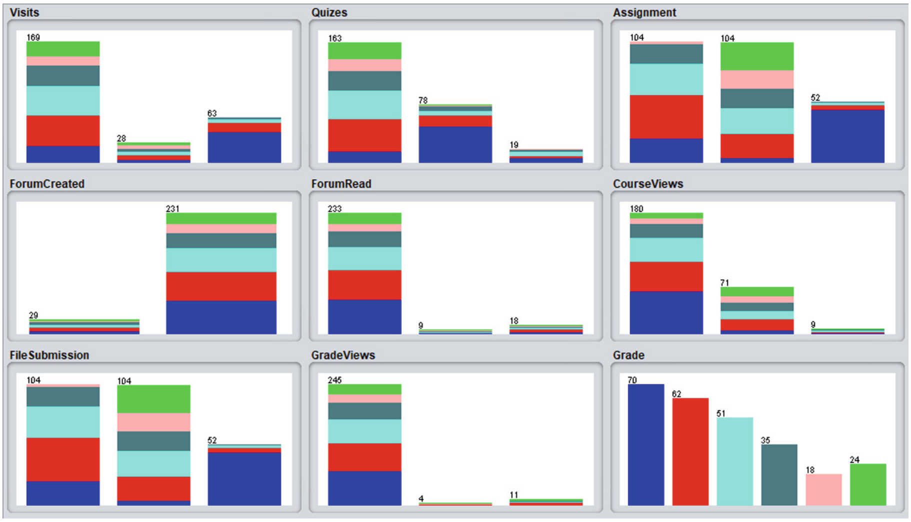
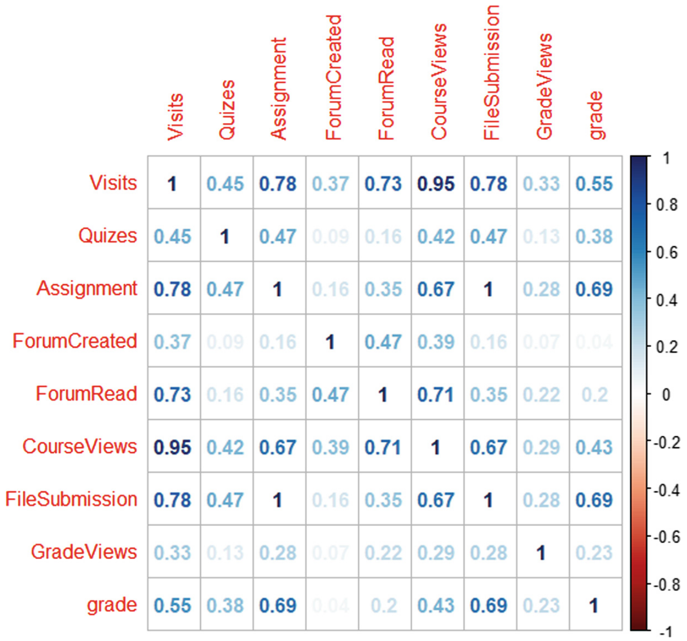


Fig. 3. WEKA output of 2016–2017 data with six classes (5, 6, 7, 8, 9, 10).

Figures 4 and 5 gives the correlation matrices of 2016–2017 and 2017–2018 data respectively.



	Visits	Quizes	Assignment	ForumCreated	ForumRead	CourseViews	FileSubmission	GradeViews	grade
Visits	1	0.45	0.78	0.37	0.73	0.95	0.78	0.33	0.55
Quizes	0.45	1	0.47	0.09	0.16	0.42	0.47	0.13	0.38
Assignment	0.78	0.47	1	0.16	0.35	0.67	1	0.28	0.69
ForumCreated	0.37	0.09	0.16	1	0.47	0.39	0.16	0.07	0.04
ForumRead	0.73	0.16	0.35	0.47	1	0.71	0.35	0.22	0.2
CourseViews	0.95	0.42	0.67	0.39	0.71	1	0.67	0.29	0.43
FileSubmission	0.78	0.47	1	0.16	0.35	0.67	1	0.28	0.69
GradeViews	0.33	0.13	0.28	0.07	0.22	0.29	0.28	1	0.23
grade	0.55	0.38	0.69	0.04	0.2	0.43	0.69	0.23	1

Fig. 4. Correlation analysis of 2016–2017 data.

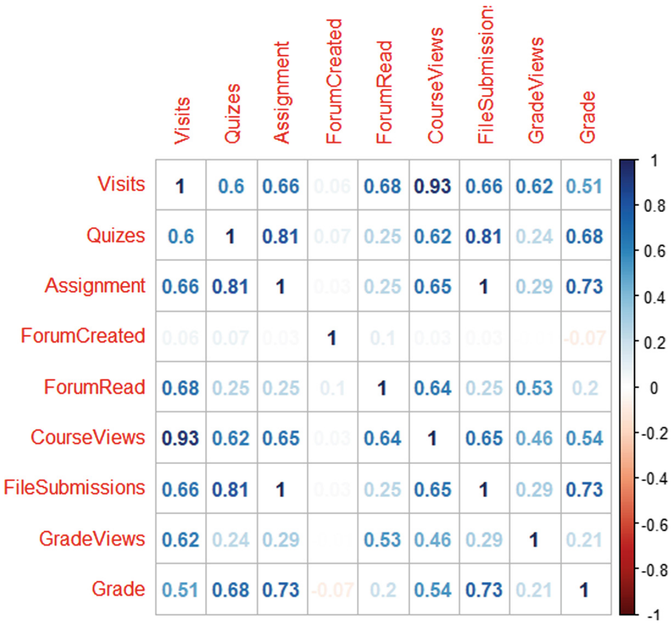


Fig. 5. Correlation analysis of 2017–2018 data.

Comparison of the correlations is given in Table 4. The higher correlation is found in Assignment, Total visits and Course views. Correlations in first and second year data are close to each other. This is why previous year's data can be used for predicting next year's grades. We can assume that students behaving similarly will receive similar grades.

Table 4. Correlation of attributes with the grade.

Attributes	Correlation coefficients	
	2016–2017 data	2017–2018 data
Total visits	0.55	0.51
Course view	0.43	0.54
Forum view	0.20	0.19
Forum created	0.04	−0.07
Quizzes	0.38	0.68
Assignment	0.69	0.72
File submission	0.69	0.72
Grade view	0.23	0.21

3.2 Prediction

As we found positive correlations between all Moodle activities except the forum creations in second data set; and the course grade as a success measure, we can use this data for modeling. The first file with 2016–2017 data is selected as training set to create the model.

At the first step the model is created by using J48 decision tree algorithm when grades are scaled from 5 to 10. The model is trained with 10 fold cross-validation. Table 5 shows the evaluation of the model with 6 classes for the grade. Overall accuracy of the model is 42.7% while TP rate for the failed students is 68.6%. The model with 6 classes works well for two classes: 5 and 10 only, and could not predict the grade 9 at all.

Table 5. Evaluation of training data set - J48 decision tree with 5 classes.

Class	TP rate	FP rate	Precision	Recall	F-Measure
5	0.686	0.179	0.585	0.686	0.632
6	0.290	0.217	0.295	0.290	0.293
7	0.529	0.230	0.360	0.529	0.429
8	0.000	0.040	0.000	0.000	0.000
9	0.000	0.000	?	0.000	?
10	0.750	0.067	0.545	0.750	0.632
Weighted Avg.	0.427	0.156	?	0.427	?

When the grades are labeled into three classes, such as FAIL for grade 5, GOOD for grades 6–7–8 and EXCELLENT for the grades 9–10; overall accuracy of the model increases to 71.5%. Table 6 shows the evaluation of the model for 3 classes, but TP rate for the FAIL class decreases to 67.1%.

Table 6. Evaluation of training data set - J48 decision tree with 3 classes.

Class	TP rate	FP rate	Precision	Recall	F-Measure
FAIL	0.671	0.105	0.701	0.671	0.686
GOOD	0.824	0.428	0.718	0.824	0.767
EXCELLENT	0.405	0.028	0.739	0.405	0.706
Weighted Avg.	0.715	0.277	0.717	0.715	0.706

When the grades are labeled into two classes, such as FAIL and PASS, overall accuracy of the model increases to 86.5%. TP rate for the FAIL class decreases to 65.7%. Table 7 shows the evaluation of the model for 2 classes.

Table 7. Evaluation of training data set - J48 decision tree with 2 classes.

Class	TP rate	FP rate	Precision	Recall	F-Measure
FAIL	0.657	0.058	0.807	0.657	0.724
PASS	0.942	0.343	0.882	0.942	0.911
Weighted Avg.	0.865	0.266	0.862	0.865	0.861

As it can be seen from Tables 5, 6 and 7; having less classes increase the overall accuracy of the model and precision of the failed grades as well, but the TP value of FAIL class decreases slightly.

We can conclude here that the activities in the learning system are not giving the complete picture of the effort and the grade for the students who passed. But on the other hand it is good measure for the prediction of failed students. For the prediction of real grades, other parameters from the course and the students should be considered, not only online activities. To provide higher overall accuracy for the model, we decided to use 2 classes and we trained the J48 decision tree model with two classes, and got the predictions for the second-year data. Confusion matrix of the test data predictions is given in Table 8. Overall accuracy of the predictions is calculated 89.32% with Kappa index 0.701 for the test data when we compared with the real grade results.

Table 8. Confusion matrix for the predicted data vs. real data with Decision tree algorithm.

Real/Predicted	PASS	FAIL	Classification overall	Precision
PASS	147	10	157	93.63%
FAIL	12	37	49	75.51%
Truth overall	159	47	206	
Recall	92.45%	78.72%		

We applied the same procedure by using Bayesian Network and Support Vector Machine classification algorithms. Tables 9 and 10 show the confusion matrices for the prediction on the test set.

Table 9. Confusion matrix for the predicted data vs. real data with Bayesian Network algorithm.

Real/Predicted	PASS	FAIL	Classification overall	Precision
PASS	146	10	156	93.59%
FAIL	13	37	50	74.00%
Truth overall	159	47	206	
Recall	91.82%	78.72%		

Table 10. Confusion matrix for the predicted data vs. real data with SVM algorithm.

Real/Predicted	PASS	FAIL	Classification overall	Precision
PASS	155	33	188	93.59%
FAIL	4	14	18	77.78%
Truth overall	159	47	206	
Recall	97.48%	29.79%		

In the end Table 11 gives the comparison of the accuracies of three different classification algorithms on the training set and on the test set. The best accuracy is achieved with J48 Decision Tree algorithm (89.32%) on the test set.

Table 11. Comparison of classification methods.

Classification method	Evaluation of the model on the training set		Evaluation of the prediction on the test set	
	Accuracy	Kappa Index	Accuracy	Kappa Index
J48 Decision Tree	86.50%	0.693	89.32%	0.701
Bayesian Network	85.00%	0.624	88.83%	0.69
Support Vector Machine (SVM)	86.15%	0.64	82.04%	0.348

4 Conclusion

This study explains all steps of Moodle log data analysis from raw data to the prediction of new data by using different classification algorithm. Data is taken from a bachelor degree course for two different years. The results show that activities in learning management system have positive correlations with the student success in terms of grade.

The main purpose of the study is to predict students' success and figure out the students who are about to drop out. Differently from many existing studies, we used one year's data to predict the next year's success; so that the drop outs of the students could be recognized in advance.

When the DT model is used with all grades from 5 to 10, it could not predict all the grades accurately. But the most sensitive group of the students (failing ones) with grade 5 were predicted with a high accuracy. Predicting the students who are about to drop out would help instructors to take precautions. Accuracies with passing grades are quite small it means system activity level is not significant on the exam score but it is significant on passing or failing the course. Because of that for predicting all grades accurately learning styles, cognitive styles and study habits of the passing students could be considered in the future studies.

We applied DT model also with three and two labels, and we saw that two labels give the best accuracy overall. By using these two labels we also created BN and SVM models for classification. Accuracy results on the test sets show that DT had the best accuracy on the prediction,

Additionally this study gave us opportunity to try two different tools: RStudio and WEKA which are very popular in data science and to see advantages and disadvantages. We preferred RStudio in pre-processing and visualization as it gives great support for executing SQL queries for data transformation. For the application of decision tree algorithm and prediction we used WEKA as its interface is very simple to set training and test data sets.

In the future, to obtain a model which can predict the grades in a complete scale; different modeling methods can be used. In this study the models were created by using 9 data attributes, in the future more data attributes can be included in the analysis. Also analyzing online activities in a weekly manner and measuring the impact of activities in the first few weeks of the semester could be done as a future work, so the early prediction of drop outs might be useful for further prompt actions.

References

1. Ademi, N., Loshkovska, S.: Exploratory analysis of student activities and success based on Moodle log data. In: 16th International Conference on Informatics and Information Technologies (2019)
2. Akçapınar, G.: Profiling students approaches to learning through Moodle logs. In: Multidisciplinary Academic Conference on Education, Teaching and Learning (MAC-ETL 2015) (2015)
3. Cocea, M., Weibelzahl, S.: Log file analysis for disengagement detection in e-learning environments. User Model. User Adap. Inter. **19**(4), 341–385 (2009)
4. Dumais, S., Jeffries, R., Russell, D.M., Tang, D., Teevan, J.: Understanding user behavior through log data and analysis. In: Olson, J.S., Kellogg, W.A. (eds.) Ways of Knowing in HCI, pp. 349–372. Springer, New York (2014). https://doi.org/10.1007/978-1-4939-0378-8_14
5. Figueira, Á.: Mining Moodle logs for grade prediction: a methodology walk-through. In: Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, p. 44. ACM (2017)
6. Grothendieck, G.: CRAN - Package sqldf. https://cran.r-project.org/web/packages/sqldf/index.html
7. Kadoić, N., Oreški, D.: Analysis of student behavior and success based on logs in Moodle. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0654–0659. IEEE (2018)
8. Käser, T., Hallinen, N.R., Schwartz, D.L.: Modeling exploration strategies to predict student performance within a learning environment and beyond. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 31–40. ACM (2017)
9. Klašnja-Milićević, A., Ivanovic, M.: Learning analytics–new flavor and benefits for educational environments. Inform. Educ. **17**(2), 285–300 (2018)
10. Konstantinidis, A., Grafton, C.: Using Excel macros to analyse Moodle logs. In: 2nd Moodle Research Conference, pp. 33–39 (2013)

11. Poon, L.K.M., Kong, S.-C., Yau, T.S.H., Wong, M., Ling, M.H.: Learning analytics for monitoring students participation online: visualizing navigational patterns on learning management system. In: Cheung, S.K.S., Kwok, L.-F., Ma, W.W.K., Lee, L.-K., Yang, H. (eds.) ICBL 2017. LNCS, vol. 10309, pp. 166–176. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59360-9_15
12. Rachel, V., Sudhamathy, G., Parthasarathy, M.: Analytics on moodle data using R package for enhanced learning management. Int. J. Appl. Eng. Res. **13**(22), 15580–15610 (2018)
13. Estacio, R.R., Raga Jr., R.C.: Analyzing students online learning behavior in blended courses using Moodle. Asian Assoc. Open Univ. J. **12**(1), 52–68 (2017)
14. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Comput. Educ. **51**(1), 368–384 (2008)
15. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**(6), 601–618 (2010)