The 10th International Conference on eLearning (eLearning-2019), 26 - 27 September 2019, Belgrade, Serbia

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR STUDENTS PERFORMANCE PREDICTION BASED ON LMS DATA

DIJANA OREŠKI
University of Zagreb, Faculty of Organization and Informatics, dijoresk@foi.hr

BOŽIDAR KLIČEK
University of Zagreb, Faculty of Organization and Informatics, bklice@foi.hr

Abstract: *Huge volumes of data created by students` activities on learning management systems (LMS) urged an opportunity to extract meaningful information from data. Development of data mining field yielded algorithms making possible to analyse data with the aim to improve quality of the educational processes. In this paper, we are comparing four data mining methods based on different machine learning algorithms to predict academic performance of IT students based on data about their activities at the LMS. Aim of the research were twofold: (i) to predict students' academic achievement and identify most important predictors of academic success, (ii) to compare different machine learning algorithms and identify which fits the best on LMS data. Research results indicated frequency of students` discussions as the most important predictor of success. Neural networks provided most accurate and reliable model.*

Keywords: *LMS data, data mining, CRISP DM, neural networks*

1. INTRODUCTION

Learning Management Systems today play a significant role in higher education learning models. The large amounts of data generated by such systems have opened up new research pathaways, especiall in an analysis of student behavior. The aim of such analyzes is to identify behavioral patterns for the purpose of improving the learning process and consecentually, enhance students academic performance. The large amounts of data and characetristics of such data also require new methodological approaches in analysis [1]. Educational Data Mining (EDM) is an emerging field developed with the aim of machine learning and data mining techniques application in educational domain. Results of such analysis are beneficial to both, students and teachers because they help in improvement of learning processes. Recent research trends in the EDM field are focused on massive open online courses (MOOC) [2]. This is in the focus of our research. Hereinafter, we used data for different students activities on one blended course using the Moodle platform. Machine learning techniques shown to be useful in the Moodle data analysis [3] . Majority of previous research papers analyzed activity logs [4-7] . Their research results indicated accurate predictive models of dropout based on the Moodle data.

This study aims to analyze activity logs of third year students at the course Knowledge discovery in data, which is tought at the IT faculty in Croatia. Data are collected from three generations in order to provide information regarding students' performance to the teachers and study programme management to help them in improving the programme and to help students improving the learning process. Furthermore, this research also focuses on the methodological aspects of activity logs analysis and tried to identified the best machine learning approach for LMS data anaylsis.

This study is structured as follows. In the next section we present relevant and recent research results regarding the research topic. Section 3 presentes research methodology: describes used data set and data mining standard used in the research: CRISP DM. Section 3 presents research results and section 4 concludes the paper.

2. RELATED RESEARCH

Academic performance of students serves as indicator of education system quality, so it is interesting research topic among scientific community. Nowadays, academic performance predictions are based on LMS data. LMS platforms offer different channels for teachers and students to communicate in a course, store files, share information, upload assignments, do the tests. LMS saves log data of the students' activities [8]. Romero and Ventura [9] were among first scientists to apply data mining techniques to LMS. They explored benefits of data mining techniques for LMS data analysis. Our research seeks to further advance their approach in terms of investigating which of the data mining techniques based on machine learning provides the best predictive models on LMS data. Other investigated topics are related towards students' satisfaction with the LMS [10] and gender differences in attitudes [11-13]. Yukselturk and Top [13] found out that females` participation was more intensive then males`. These results are in line with previous research results of Kimbrough et al. [12] which emphasized high females` engagement in communication on LMS. However, there are different evidence from Prinsen, Volman, and Terwel [14]. They found out high participation of male students. There is a need to investigate the impact of such participation on student course performance measured by students` grade. This paper presents a case study based on an undergraduate IT course. Hereinafter, we present the results of our research on this topic.

3. RESEARCH METHODOLOGY

The participants of this study were 105 undergraduate students who enrolled in a course Knowledge discovery in data at the University of Zagreb. Participants in this study used online learning to access course materials (a content page, web links, self-reports), to download presentations, to upload their assignments, to ask questions to the teacher, to solve tests. The online learning environment was developed in a Learning Management System (LMS), Moodle. This study used data about students` activities presented in Table 1.

Table 1: Variables description

Variable	Description
Files view	*Numeric* Frequency of files opening (Power point presentations, PDF files,…)
Forum view	*Numeric* Frequency of forum reading and posting
Student report view	*Numeric* Frequency of report opening
Folder view	*Numeric* Frequency of folder opening
Choice	*Numeric* Frequency of choice opening
File upload	*Numeric* Number of files uploading
System login	*Numeric* Frequency of system login
Test	*Numeric* Optional tests points
Assignment	*Numeric* Points for solutions of assignments
Gender	*Categorical/* Female or male
Grade	*Categorical* *Overall grade achieved at the course*

CRISP DM process model is applied in this research to analyze the data. CRISP DM stands for CRoss-Industry Standard Process for Data Mining [15]. CRISP DM implies six steps for data analysis (see figure 1).

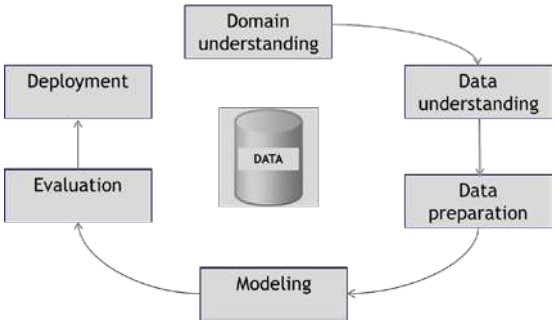

Image 1: CRISP DM process

First, data exploration and data preparation were performed. Modeling phase is in the focus of this research. Modeling consists of building and assessing the predictive models. Here, we are performing students academic performance prediction based on activity logs at the LMS. Prior, modeling technique should be selected. Four modeling techniques belonging to different machine learning approaches were selected and applied and their parameters are calibrated to optimal values. Those are:
(i) Classification and regression trees (CART) categorized as information based machine learning approach,
(ii) Neural networks categorized as error based machine learning approach,
(iii) Naive Bayes classifier categorized as probability based machine learning approach,
(iv) k-nearest neighbours categorized as similarity based machine learning approach.

Evaluation was performed and four models were compared based on two measures: accuracy of the model (Root mean square error, RMSE) and reliability of the model (RSquare).

4. RESEARCH RESULTS

To develop predictive models, four machine learning techniques were applied using the same data set. Results of accuracy and reliability for each model are given in table 2.

Table 2: Performance of ML models

ML Algorithm	Accuracy	Reliability
Neural network	86,53%	0,64
CART	83,21%	0,55
Naïve Bayes classifier	79,54 %	0,51
kNN	78,44%	0,48

As results demonstrate, neural networks yielded predictive model of highest accuracy and reliability, followed by Classification and regression trees. K-nearest neighbours, technique based on the similarity produced the worst results. T-test was further performed on the results in order

to see are differences among machine learning models statistically significant.

Table 3: T-test results

ML Algorithm	p-value
Neural network	-
CART	*0,003*
Naïve Bayes classifier	*0,01*
kNN	*0,01*

Based on pairwise t-test shown in Table 3, the overall accuracy of neural network algorithm is better than the overall average accuracy of all the other algorithms and the difference is statistically significant at 95% confidence level. It is interesting to note that neural networks and kNN are two approaches requiring only numerical inputs. Since all of our input variables were numerical, those results are not surprising. CART proven to work better on categorical inputs.

Since neural networks model is the best, we will interpret and discuss results of that model. First, architecture of neural network is presented at image 2.

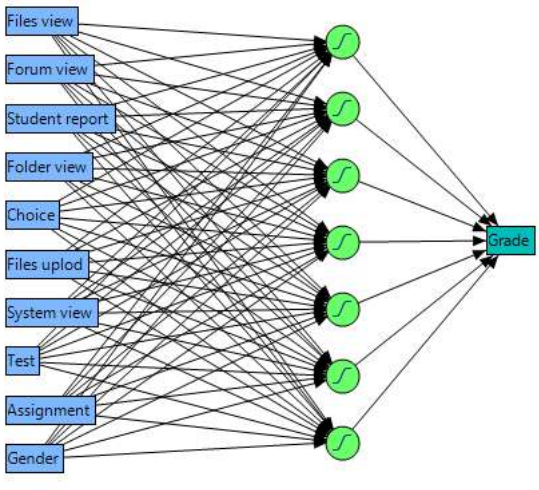

Image 2: Neural network architecture

Neural network architecture consists of 10 variables (neurons) in the input layer, 7 neurons in the hidden layer processing the data, and one neuron (grade) at the output layer of the neural network.

The results of sensitivity analysis based on NN predictive model is presented in Table 2. As shown in Table 3, from the total 10 variables listed for investigation, 2 variables emerged as the strongest predictors: Forum View and Test.

The predictive model tells us that having more forum views or having better results on optional tests will lead to the students achieving a better overall grade at the course.

Table 3: Sensitivity analysis results

Column	Main Effect	Total Effect
Forum view	0.107	0.437
Test	0.074	0.405
Gender	0.067	0.317
Assignment	0.076	0.252
Choice	0.034	0.217
System view	0.036	0.193
Files upload	0.032	0.189
Student report	0.024	0.144
Folder view	0.02	0.103
Files view	0.011	0.058

Neural network model indicated gender differences in academic performance. Variable Gender is third highest predictor of academic performance. This finding is in line with previous research which proved differences in attitudes and behaviour on LMS between male and female students. It is interesting to note that frequency of files and folder views isn't significant predictor of academic success.

5. CONCLUSION

Results of this research contributes to existing research on educational data mining in LMS data by demonstrating that machine learning approaches are effective for LMS data analysis. We have demonstrated that neural networks show greatest potential for LMS data analysis application, so we assume that if the characteristics of other data-sets are similar, then the performances of the machine learning approaches on these data-sets are similar as well.

The purpose of this study was also to examine based on which LMS activity student performance could be predicted? Neural network model has pointed out forum interaction as most important predictor. Students engagement into reading forum posts and interacting into discussions found to be indicator of overall grade at the course. Solving optional tests is the second highest predictor.

The results of this study could help students to focus on the LMS activities which would lead them to achieving academic success and to help teachers to identify students who could struggle. It will also give teachers the way to provide early guidelines for online learning activities configuration.

This study was limited by small data used in the analysis However, it was reliable enough to develop accurate predictive models and to give insights into students activities and their relation with academic success. In the future research we will include students of different courses and study programmes to see are there differences in patterns.

REFERENCES

[1] Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a Big Data framework. Future Generation Computer Systems, 90, 262-272.

[2] Koedinger KR , D'Mello S , McLaughlin EA , Pardos ZA , RoséCP . Data mining and education. WIREs Cogn Sci 2015(6):333–53.

[3] Romero C , Ventura S , Hervás C , Gonzales P . Data mining algorithms to classify students. In: Proceedings of the international conference on educational data mining; 2008. p. 8–17 .

[4] Lara JA , Lizcano D , Martínez MA , Pazos J , Riera T . A system for knowledge discovery in e-learning environments within the European higher education area – application to student data from Open University of Madrid, UDIMA. Comput Educ 2014;72:23–36 . [4] Baker R , Yacef K . The state of educational data mining in 2009: a review and future visions. J Educ Data Mining

[5] Mclaren BM , Koedinger KR , Schneider M , Harrer A , Lollen L .Bootstrapping novice data: semi-automated tutor authoring using student log files. In: Proceedings of the workshop analyzing student-tutor interaction logs improve educational outcomes; 2004. p. 1–10 .

[6] Yu CH , Jannasch-Pennell A , Digangi S , Wasson B .Using onlineinteractive statistics for evaluating web-based instruction. J Educ Media Int 1999;35:157–61 .

[7] Hübscher R , Puntambekar S , Nye A . Domain specific interactive data mining. In: Proceedings of the 11th international conference on user model. Workshop data mining user model; 2007. p. 81–90 .

[8] Mazza, R., & Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education.

[9] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

[10] Martínez-Caro, E., & Campuzano-Bolarín, F. (2011). Factors affecting students' satisfaction in engineering disciplines: traditional vs. blended approaches. European Journal of Engineering Education, 36(5), 473–483.

[11] Huang, W. H. D., Hood, D. W., & Yoo, S. J. (2012). Gender divide and acceptance of collaborative Web 2.0 applications for learning in higher education. Internet and Higher Education, http://dx.doi.org/10.1016/j.iheduc.2012.02.001.

[12] Kimbrough, A. M., Guadagno, R. E., Muscanell, N. L., & Dill, J. (2013). Gender differences in mediated communication: women connect more than do men. Computers in Human Behavior, 29(3), 896–900. http://dx.doi.org/10.1016/j.chb.2012.12.005.

[13] Yukselturk, E., & Top, E. (2012). Exploring the link among entry characteristics, participation behaviors and course outcomes of online learners: an examination of learner profile using cluster analysis. British Journal of Educational Technology, http://dx.doi.org/10.1111/j.1467-8535

[14] Prinsen, F. R., Volman, M. L. L., & Terwel, J. (2007). Gender-related differences in computer-mediated communication and computer-supported collaborative learning. Journal of Computer Assisted Learning, 23, 393–409. http://dx.doi.org/10.1111/j.1365- 2729.2007.00224.x.

[15] Wirth, R., Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Location: New York, USA.