Development and comparison of predictive models based on learning management system data

DIJANA OREŠKI, GORAN HAJDIN
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42 000 Varaždin Address
CROATIA
dijoresk@foi.hr, gohajdi@foi.hr

Abstract: - This study introduces and evaluates different feature reduction and supervised machine learning approaches to predict the students` academic success. Our analysis is based on data about students` activities in the learning management system. We found that neural networks based approach with principal component analysis in feature reduction is the most effective for learning management system data.

Key-Words: data preparation, feature reduction, machine learning, neural networks, academic performance, student success.

1 Introduction

Online systems such as Learning Management Systems (LMS), Course Management Systems (CMS), Massive Open Online Courses (MOOCS), Virtual Learning Environments (VLE), Intelligent Tutoring systems (ITS) and other web-based educational systems collect huge amounts of data about learners` activities. These datasets represent valuable tool for investigation of learners` behavior by analyzing their activities and providing suggestions based on learner's activities. Due to the huge volume of data, students` performance prediction becomes very challenging. In the recent years machine learning has been used and explored, especially in the field of higher education [1;2;3;4]. Multiple machine learning approaches exist which offer different nature of classification and regression models [5]. The behavior of these algorithms is largely data dependent [6]. Several empirical studies [7; 8] have been carried out to analyze performance of different classification algorithms. Some studies [7] analyzed behavior of algorithms for a wide spectrum of problems and found weak performance without analysis of relationship between algorithms performance and specific domains and datasets characteristics. With the increasing popularity of machine learning algorithms, it is becoming imperative to identify which specific algorithm will perform better for a particular domain and dataset.This kind of studies received our attention since it is important to consider descriptions and domain specific characteristics of datasets. The idea is to explore capabilities of machine learning approaches on educational data. Educational data is generated as a result of interaction between learners and instructors. As such, educational data analysis is multidisciplinary field of research [9]. In this paper we examine how different machine learning approaches deal with LMS data and investigate implications of predictive models for education policy. We have tested various machine learning approaches and found the one that gives reliable and accurate results with small error on the LMS dataset. The paper is organized through several sections, where the section 2 discusses the previous research papers related to this topic. Section 3 explains used methodology and describes a chosen dataset. Section 4 lists the results and provides a discussion about research results. Conclusions of our analysis and suggestions for future works are explained in the final section of this paper.

2 Previous work

In the study which focused on the application of machine learning [1] results indicated that

specific machine learning techniques could provide solution for specific learning problems. In another study [10] researchers applied learning analytics in information technology context to explore the process of students' teamwork and to improve students` success, while researchers of the next study [11] tried to identify student behavior patterns in their interaction with online learning environments and draw conclusions to improve student academic performance [11]. Various machine learning techniques were deployed on educational datasets to predict students' academic success [12;13;14;15]. The algorithms they have used are the linear regression (LR), logistic regression (LGR), neural network (NN), decision trees (DT), k-nearest neighbors (k-NN), Support Vector Machine (SVM), Bayesian networks (BN) and Naïve Bayes classifiers (NB) for supervised learning, and factor analysis and K-Means for unsupervised learning. Most of the papers claim that machine learning models are superior to classical statistical-learning-based mechanism in prediction accuracy. All techniques are employed in order to improve data-driven decision making in educational domain [16]. As seen in literature review, bulk of machine learning algorithms in educational domain has been applied. However, little is known about the effect of different approaches on the domain specific datasets. Thus, we are focusing on this issue and we are investigating which of the five main machine learning approaches is the most effective for students` success prediction based on the LMS data. To do so, factor analysis is used as feature reduction method and neural networks [17], decision tree [18], k-nearest neighbours [19] and Naïve Bayes classifier [20] are used as methods for developing predictive models. In the next section we are explaining research design and methodology.

3 Methodology

This chapter gives details of dataset characteristics and an overview of machine learning approaches employed.

3.1 Data Set Description

The dataset is retrieved from students' log data at one course which is a part of the Information and Business Systems study program at the University of Zagreb, Faculty of Organization and Informatics in Croatia. Thus, analysis is performed in the specific, information technology context.

Two datasets are used. One is Moodle log files of accessing the course materials and the other consists of grades which students achieved in course sections. Two datasets are integrated into one file. Description of the used variables is shown in Table 1.

TABLE I. VARIABLE DESCRIPTION

Variable	Description
Pass	Target variable. It tells if student has passed the course.
additional_grades	Exam for additional grades.
blitz_1	First unannounced exam.
blitz_2	Second unannounced exam.
blitz_3	Third unannounced exam.
blitz_4	Fourth unannounced exam.
exam_1_excel	First exam. Topic is MS excel.
exam_2_accessdb	Second exam. Topic is MS

	access.
retry_full_exam	Exam for whole course.
attend_lab	Attendance for laboratory practice.
attend_lectures	Attendance for lectures.
self_check_1	Self-assesment quiz 1.
self_check_2	Self-assesment quiz 2.
self_check_3	Self-assesment quiz 3.
self_check_4	Self-assesment quiz 4.
self_check_5	Self-assesment quiz 5.
self_check_6	Self-assesment quiz 6.
self_check_7	Self-assesment quiz 7.
self_check_8	Self-assesment quiz 8.
self_check_9	Self-assesment quiz 9.
self_check_10	Self-assesment quiz 10.
self_check_11	Self-assesment quiz 11.
self_check_12	Self-assesment quiz 12.
access_map	Organized set of files.
access_file	Files about specific topics in the course.
access_forum	Forum is used for communication between teacher and students in asynchronous manner.
access_student_report	Report about student's progress. Individual view.
access_lesson	Usage of lesson materials.
pick_group	Selection of a group for the laboratory exercises.
upload_file	File submission (i.e. homework).
access_links	Links to other resources which are relevant for the course.
access_review_report	Review report about student progress. Individual view.
access_page	Short texts on course page which includes: literature list, guides, etc.
access_system	Accessing course main page.
access_exam	Exam grade reports.
access_homework	Accessing

	homework submission activities.

3.2 Data mining process

In the research we used dross-industry process for data mining (CRISP DM) standard. CRISP DM [21] consists of six phases presented in Fig. 1. CRISP DM was also used in the similar data contexts [23]. First phase of CRISP DM focuses on the objectives of research and definition of data mining problem. Main objective of this paper is to investigate machine learning algorithms performance on LMS data. Data understanding phase explores data: type of variables, their characteristics and distributions. Variables included in this research are explained in Table 1. Data preparation step deprives 70% to 90% of data mining process [24]. Feature reduction is the most important part of data preparation. Factor analysis was used in this research to perform feature extraction. Modelling phase consists of development and assessment of the models. In this step modeling method is selected. Hereinafter, we have used five different methods, of which each belongs to a different machine learning approach: neural network and support vector machines (SVM) as an error based machine learning approach, decision tree as information based machine learning approach, k-nearest neighbors as similarity based machine learning approach and Naïve Bayes classifier as probability based machine learning approach. Comparison of machine algorithm approaches on this specific domain is applied. Evaluation phase explores how well model performs on the test data. In order to answer the main research question we have used two performance measures: root-mean-square error as an accuracy measure [25] and execution time of machine learning algorithms. Deployment phase explains how modelling results need to be utilized.

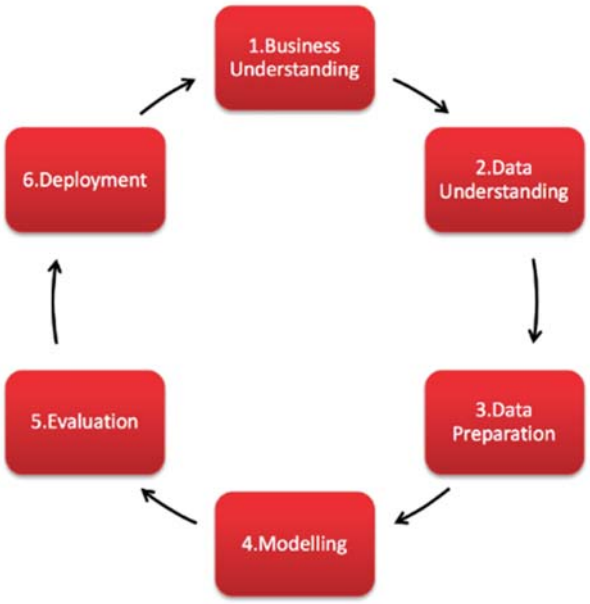
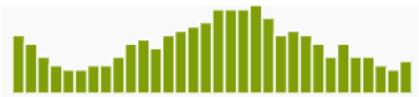
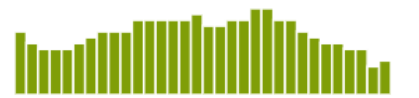
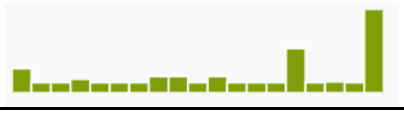
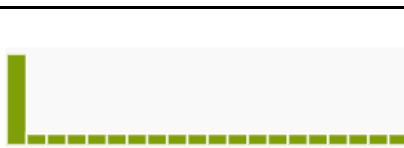
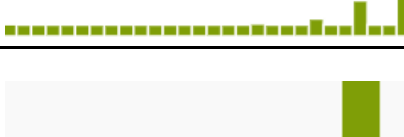
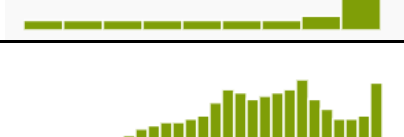
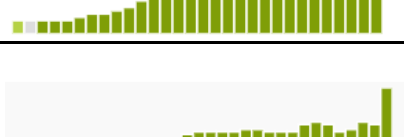
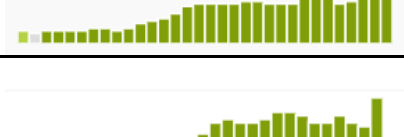
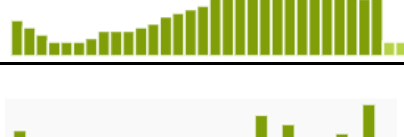
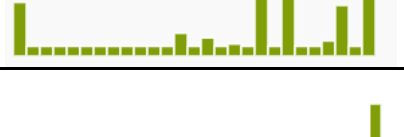
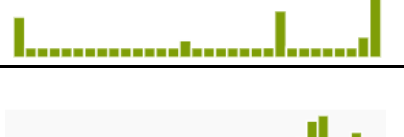
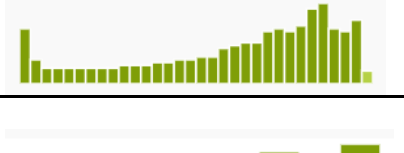
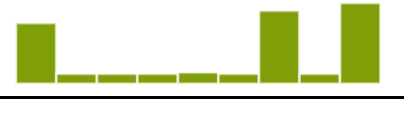
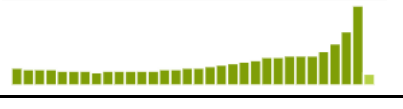
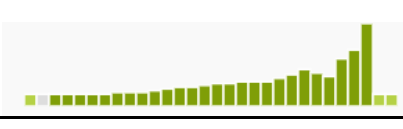
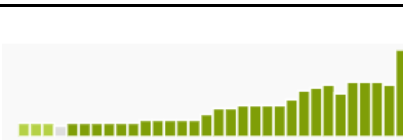
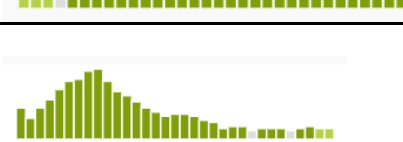
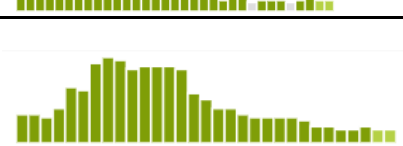
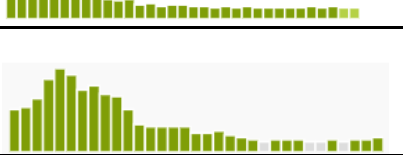
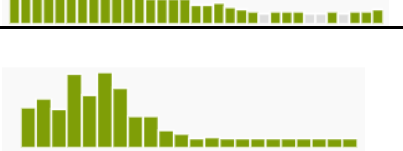
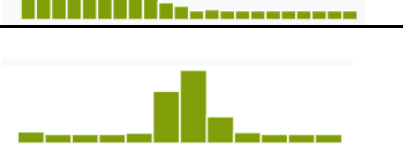
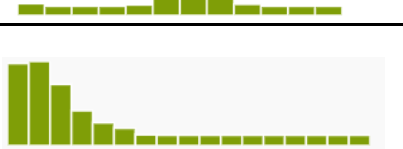
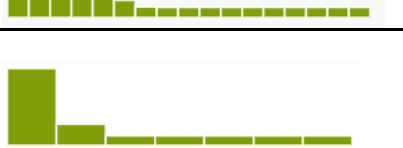
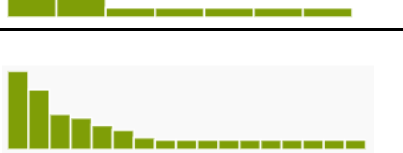


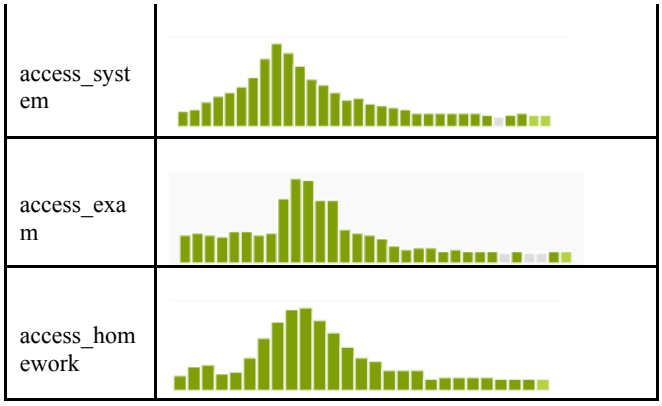
Fig. 1. CRISP DM methodology [22]

First and second step of CRISP DM standard, business and data understanding, are presented through distributions of variables (see table 2).

TABLE II. DATA DISTRIBUTION

Variable	Distribution
Pass	
additional_grades	
blitz_1	
blitz_2	
blitz_3	
blitz_4	

exam_1_excel	
exam_2_accessdb	
retry_full_exam	
attend_lab	
attend_lectures	
self_check_1	
self_check_2	
self_check_3	
self_check_4	
self_check_5	
self_check_6	
self_check_7	
self_check_8	
self_check_9	
self_check_10	
self_check_11	
self_check_12	
access_map	
access_file	
access_forum	
access_student_report	
access_lesson	
pick_group	
upload_file	
access_links	
access_review_report	
access_page	



4 Results and discussion

How does different kinds of machine learning algorithms affect students` performance prediction was investigated on the LMS dataset. In this section we present results of the application of five machine learning algorithms and one feature extraction algorithm. Firstly, factor analysis was performed. Factor analysis transforms the input features into new features (called factors) which are not correlated. First four factors extracted explain 61,08% of the total variance.Many of the previous papers explained possibilities of factor analysis in reduction of the number of input features in the data mining models which led to reduction in modelling time e.g. [26, 27, 28, 29]. When applying the factor analysis we aim at optimizing the performance of a machine learning algorithms by removing a number of features which presence hinders the predictive models accuracy. Results of the factor analysis served as the input for all five machine learning algorithms. The modeling was conducted to find the most accurate and reliable predictive model with a high level of accuracy on one hand, and minimal number of predictors on the other hand. Out of 35 input variables, four variables were excluded from factors analysis: additional_grades, exam_1_excel, exam_2_accessdb, retry_full_exam since they are part of output variable, pass. Since factor analysis explained 61% of variance with 4 factors, we have extracted 4 factors. Four extracted factors are named: Self-check_first, Access materials, Attendance and blitz tests, Self-check_second. Results of the factor analysis served as input into predictive modelling by five different machine learning approaches.

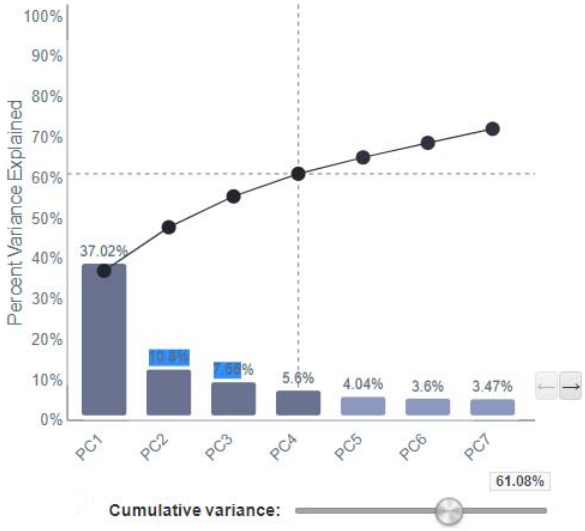


Fig 2. Scree plot of the eigenvalues

4.1 Comparative analysis of predictive accuracy

Table 3. depicts accuracy achieved by each of the hybrid data mining models. Neural networks prediction with factor analysis in data reduction performs with the highest accuracy.

TABLE III. COMPARISON OF ACCURACY

Model	Accuracy
PCA + NN	74,15 %
PCA + DT	71,23 %
PCA+SVM	69.89 %
PCA + KNN	67,61 %
PCA + BC	64,36 %

The performance metrics from Table 3 allow us to make conclusions about different machine

learning approaches on our dataset. Neural networks modeling resulted with the most accurate prediction. Our conclusions imposes the following question: could the results be generalized? Statistical testing was performed in order to answer this question. The purpose of statistical significance testing is to find the level in which accuracy represents behavior of machine learning algorithms. We have tested two approaches on one domain and used two matched sampled t-test. In this part of the research our aim is to test if the differences in means are significant. The assumption is that difference between the means is zero (the null hypothesis). Assumptions for performing t-test were satisfied.

TABLE IV. TESTING STATISTICAL SIGNIFICANCE OF DIFFERENCES BETWEEN ACCURACIES

Hypothesis	Model	t-test
H0: PCA+NN = PCA+DT	PCA+NN PCA+DT	*P=0,04*
H0: PCA+NN = PCA+SVM	PCA+NN PCA+SVM	*P=0,02*
H0: PCA+NN = PCA+KNN	PCA+NN PCA+KNN	*P=0,01*
H0: PCA+NN = PCA+BC	PCA+NN PCA+BC	*P=0,01*
H0: PCA+DT = PCA+SVM	PCA+DT PCA+SVM	*P=0,05*
H0: PCA+DT = PCA+KNN	PCA+DT PCA+KNN	*P=0,02*
H0: PCA+DT = PCA+BC	PCA+DT PCA+BC	*P=0,02*
H0: PCA+ SVM = PCA+KNN	PCA+ SVM PCA+KNN	*P=0,03*
H0: PCA+ SVM = PCA+BC	PCA+ SVM PCA+BC	*P=0,05*
H0: PCA+KNN = PCA+BC	PCA+KNN PCA+BC	*P=0,04*

Table 4 show results of t-test. Results indicate statistically significant differences in performances of NN with DT, k-NN and BC. According to the results error based approach was the best predictive modelling approach on the LMS data. Furthermore, both DT and k-NN perform significantly better then Naïve Bayes classifier. Naïve Bayes classifier gave poor results. Since most of the variables in LMS dataset are numerical continuous, neural networks gave the best results since transformations are not needed. k-NN, an algorithm which also requires numerical variables, was ranged third. Decision tree algorithms work with both categorical and numerical variables, whereas Bayesian classifier performs well with categorical inputs. Type of variables shown to be crucial in LMS dataset case.After testing of statistical significance the results confirm superiority of the approach which consists of factor analysis in data reduction and neural networks in classification. The results showed that error based machine learning approach outperforms principal information based, similarity based and probability based approaches in predictive modelling.

4.2.Comparative Analysis of Modelling Time

In addition, we have performed performance analysis regarding execution speed. In this research, execution speed is overall time needed to build predictive models consisting of feature reduction time and machine learning based predictive models development. Results of t-test are presented in Table 5. Here we tested whether the difference in execution speed is significant.The results showed that information based machine learning algorithm (decision tree) outperforms other approaches in speed. The second fastest approach is similarity based

machine learning approach, k-nearest neighbors. Those two machine learning approaches are statistically significantly faster than other approaches. Neural networks shown to be the slowest predictive modelling approach. Neural network algorithm is computationally extensive and time consuming.

TABLE V. TESTING STATISTICAL SIGNIFICANCE OF DIFFERENCES BETWEEN EXECUTION SPEED

Hypothesis	Model	t-test
H0: PCA+DT = PCA+KNN	PCA+DT PCA+KNN	*P=0,01*
H0: PCA+DT = PCA+SVM	PCA+DT PCA+SVM	*P=0,03*
H0: PCA+DT = PCA+BC	PCA+DT PCA+BC	*P=0,02*
H0: PCA+DT = PCA+NN	PCA+DT PCA+NN	*P=0,04*
H0: PCA+KNN = PCA+SVM	PCA+KNN PCA+SVM	*P=0,02*
H0: PCA+KNN = PCA+BC	PCA+KNN PCA+BC	*P=0,02*
H0: PCA+ KNN = PCA+NN	PCA+ KNN PCA+NN	*P=0,01*
H0: PCA+SVM = PCA + BC	PCA+SVM PCA + BC	*P=0,03*
H0: PCA+SVM = PCA + NN	PCA+SVM PCA + NN	*P=0,05*
H0: PCA+BC= PCA+NN	PCA+BC PCA+NN	*P=0,02*

4.3.Interpretation of Predictive Models

Results of the sensitivity analysis are presented in the Fig. 3. As stated in the chapter IV., four factors which were extracted are: Self-check_first, Access materials, Attendance and blitz tests, Self-check_second.Out of 36 variables presented in the Table 1 factor Self-check first includes 7 out of 12 self-assessment quizzes, while the factor Self-check second includes the remaining 5 out of 12 self-assessment quizzes. Interesting to note is that the factors have clear grouping of the self-assessment quizzes. We need to further analyze why the results are grouped this way (ie. have the concept of self-assessment quizzes changed or what is different, did students get familiar with the quizzes and they gain better results, etc.) so we can provide valid interpretation of the factors.

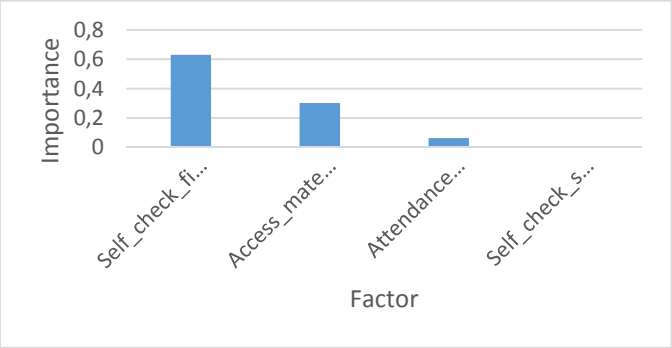


Fig 3. Sensitivity analysis

Factor Access materials consists of 8 out of 11 variables which are focused of different student's accessing actions in the LMS. Factor Attendance and blitz tests consists of both attendance variables and three out of four unannounced exams.With such factors we see clear connections between variables, where the main rule of the grouping is the similarity. The only deviation is two factors which are both consisted of self-assessment quizzes, but here again we have a clear separation among the factors where one has first five self-assessment quizzes, and the other has the later seven quizzes.For deeper understanding of the factors we need to conduct further research which will be part of our next scope. In the paper we will focus on predictive model results which will be explained within a pedagogical context, in order

to be used as part of a student support mechanism.

4 Conclusion

In this paper we have focused on the modelling phase of data mining and investigated which of the five machine learning approaches is more effective for classification. The experiment has demonstrated that machine learning modelling performed by error based approach (neural network modelling) can better classify students than other machine learning approaches. Neural network modelling of students` performance with previously conducted data reduction by factor analysis yields numerous benefits when analyzing categorical data. Type of variables included in the dataset is very important in data analysis and represents most important dataset characteristic for machine learning algorithm selection.

This study presents a contribution to knowledge in early prediction of students at-risk of low performance, determining students likely to withdraw from modules and ascertaining significant features that enable a student to outperform others.

This paper demonstrated how machine learning algorithm can be used to identify the most important attributes in a LMS data. Students achievement could be improved by using data mining techniques. Results of predictive models bring the benefits to management of academic institutions, teachers and students. In the future research, we will investigate implications of our research and explore predictive model results within pedagogical context.

References:

[1] [1] Aldowah, H., Al-Samarraie, H., & Fauzy, M. W. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics. 13-49.

[2] Chung, J. Y & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review. 346-353.

[3] Gray, C. C. & Perkins, D. (2019) Utilizing early engagement and machine learning to predict student outcomes. Computers & Education Journal, 22-32.

[4] Ud Din, I., Guizani, M., Rodrigues, J. J.P.C., Hassan, S. & Korotaev, V. (2019). Machine learning in the Internet of Things: Designed techniques for smart cities. Future Generation Computer Systems.https://doi.org/10.1016/j.future.2019.04.017

[5] Kininenko, I.,&Kukar,M.(2007). Machine learning and data mining:Introduction to principles and algorithms . Horwood Publishing Limited

[6] Ho, T.K. (2008).Data complexity analysis:Linkage between context and solution in classification. Lecture Notes in Computer Science,5342 , 986–995.

[7] Lim, T. S., Loh, W. Y., & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine learning, 40(3), 203-228.

[8] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural computation, 8(7), 1341-1390.

[9] Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98, 166–173.

[10] Zhang, X., Meng, Y., de Pablos, P. O., & Sun, Y. (2017). Learning analytics in collaborative learning supported by Slack: From the perspective of engagement. Computers in Human Behavior.

[11] Filva, D. A., Forment, M. A., Garcia-Penalvo, F. J., Escudero, D. F. & Casan, M. J. (2019). Clickstream for learning analytics to assess students' behavior with Scratch Future Generation Computer Systems, 673-686.

[12] Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: University student dropout case study. International Journal of Data Mining & Knowledge Management Process, 5(1), 15.

[13] Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57,500–508.

[14] Hardman, Julie, Alberto Paucar□Caceres, and Alan Fielding. "Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm." Systems Research and Behavioral Science 30, no. 2 (2013): 194-203.

[15] Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. International Journal of Innovative Technology & Creative Engineering, 1(12), 13–19. ArXiv Preprint ArXiv:1202.4815.

[16] Waheed, H., Hassan, S. U., Aljohani, N. R., & Wasif, M. (2018). A bibliometric perspective of learning analytics research landscape. Behaviour & Information Technology, 37(10-11), 941-957.

[17] Saad, D. (Ed.). (2009). On-line learning in neural networks (Vol. 17). Cambridge University Press.

[18] Kohavi, R., & Quinlan, J. R. (2002, January). Data mining tasks and methods: Classification: decision-tree discovery. In Handbook of data mining and knowledge discovery (pp. 267-276). Oxford University Press, Inc..

[19] Cunningham, P., & Delany, S. J. (2007). k-Nearest neighbour classifiers. Multiple Classifier Systems, 34(8), 1-17.

[20] Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In 2007 International Conference on Convergence Information Technology (ICCIT 2007) (pp. 1541-1546). IEEE.

[21] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000, pp. 29–39.

[22] A. I. R. L. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," in Proceedings of the IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, 2008, pp. 182.-185.

[23] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94, 335-343.

[24] G. J. Myatt, Making sense of data: a practical guide to exploratory data analysis and data mining. Hoboken, N.J: Wiley-Interscience, 2007.

[25] N. Japkowicz and M. Shah, Evaluating Learning Algorithms: a classification perspective. Cambridge ; New York: Cambridge University Press, 2011.

[26] Zekić-Sušac, M., Šarlija, N., & Pfeifer, S. (2013). Combining PCA analysis and artificial neural networks in modelling entrepreneurial intentions of students. Croatian Operational Research Review, 4(1), 306-317.

[27] Bucinski, A., Baczek, T., Wasniewski, T., and Stefanowicz, M. (2005), "Clinical data analysis with the use of artificial neural networks (ANN) and principal component analysis (PCA) of patients with endometrial carcinoma", Reports on Practical Oncology and Radiotherapy, Vol. 10, pp. 239-248.

[28] O'Farrella, M., Lewisa, E., Flanagana, C., Lyonsa, W.B., Jackman, N. (2005), "Combining principal component analysis with an artificial neural network to perform online quality assessment of food as it cooks in a large-scale industrial oven", Sensors and Actuators B, Vol. 107, pp. 104–112.

[29] Sousa, S.I.V. , Martins, F.G. , Alvim-Ferraz, M.C.M., Pereira, M.C. (2007), "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations", Environmental Modelling & Software, Vol. 22, pp. 97-103.