Educational Data Driven Decision Making: Early Identification of Students at Risk by Means of Machine Learning

Romano Kovač
Poslovna inteligencija d.o.o.
Krste Pavletića 1, 10000 Zagreb
romano.kovac@outlook.com

Dijana Oreški
University of Zagreb, Faculty of Organization and Informatics
Department of Information Systems Development
Pavlinska 2, 42000 Varaždin
dijana.oreski@foi.hr

Abstract. *In the last few years there has been a notable increase in the data mining usage for educational purposes. Educational data mining is emerging field of research which has the aim of analysing data about students` activities. Prediction of student achievements is among the fastest growing research in this domain. Main goal of this paper is to provide useful knowledge to faculties and their management using data about students` activity at the LMS Moodle and comparing different machine learning techniques in order to analyse this data. In this paper we have evaluated four machine learning algorithms: neural networks, decision tree, support vector machines and logistic regression. Decision tree shown to be most accurate predictive model. Results indicated lecture and seminar attendance as significant predictors of academic success.*

Keywords. Data-driven educational decision making, decision support system, machine learning, academic performance.

1 Introduction

As education increasingly relies on technology, huge amount of data about students` activity is available. Records with student activities, ratings, interactions with teachers and other students are now collected through learning management systems (LMS) such as Edmodo or Moodle. Within all levels of education there is a need to use this data in order to develop systems that help to increase the likelihood of students' success. Data mining is one of the most popular tools for student performance analysis and it has been widely applied in the educational area leading to the development of special subfield called educational data mining. Purpose of educational data mining is to transform raw data into useful knowledge. Educational data mining enables data-driven educational decision making at all levels (Van Barneveld et al., 2012). Main direction of the research in educational data mining is analysing data from online courses. This research aims to analyze data from students` activity at the one Moodle course which was carried out at the University of Zagreb, Faculty of Organization and Informatics.

Paper is organized as follows. Second section gives brief overview of previous research in this domain. Framework for research is set up in the third section, following by modelling description and research results in fourth section. At the end, conslucion is given and directions for further research are presented.

2 Related work

This chapter will highlight some related scientific research on the issue of predictive models development for intervention system and student performance measurements by the means of data mining methods.
Al-Barrak and Al-Razgan (2015) collected data from 170 students enrolled ate the *Data Structure* course. *Data Structures* course has high failure among IT students and goal of the research was to identify how to predict students´ failure? Data mining methods are applied in the study of student success in this course. There were 158 students included in the research. Each student record has the following features: student ID, student name, test 1 grade, test 2 grade, test 3 grade, inter-exam 1, intermediate level 2, project, tutorials, final exam, and total number of points earned. The score for the course was 60 points per year and 40 points for the final exam. The student must have at least 60 out of 100 points to complete the course. Result of the research was predictive model with the precision of 91 %.
Daud, Aljohani, Abbas, Lytras, Abbas and Alowibdi (2017) performed research with the purpose to predict students' performance: whether they would successfully finish their study or they would fail. Data were collected by graduate and undergraduate students at various universities in Pakistan during the period of 2004 to 2011. Number of 776 student records were

collected, of which 690 students who completed their study and 86 students who dropout. Authors applied Support Vector Machines (SVM), C4.5, Classification and Regression trees (CART), Bayesian nets (BN), and Naive Bayes (NB) in their comparative analysis. Precision, recall and F1 were used as performance measures. SVM works the best on data set with a F1 result of 0.867, which is 13% better compared to the second best method. BN and NB classifiers gave better results compared to C4.5 and CART. Student academic achievements are based on various factors such as the environment, personality, social, psychological and other variables.

Bhardwaj and Pal (2011) use Naive Bayes classification algorithm. They found that student performance was highly dependent on their grade obtained from high school examinations. Furthermore, other important variable for the student success prediction was student's location of residence. In, general, their study shows that the academic outcomes are not always dependent on their own efforts but also other factors have a significant impact on student performance.

Romero, Ventura and Garcia (2007) investigated Moodle as source of a large amount of data that is very useful for student behavioral analysis and could create a real gold mine for educational data mining. Moodle records all the student activities involved, such as reading, writing, testing, performing different tasks, and even communicating with peers. Moodle also provides personal user information (profile), academic results, and user interaction data.

Corsate and Walker (2015) performed comparison of machine learning algorithms application to the Moodle data set. The results of the modelling were analyzed with the aim of transforming available information into structured intelligent system. This study used unsupervised learning method, K-means clustering, method that divides data into clusters. Students with the highest level of similarity are grouped together in the same cluster.

Keshtkar, Cowart and Crutcher (2018) used data set that contains metadata for interactivity of students and professors on LMS Moodle through 11 programming courses over two semesters at the State University South East Missouri. The final grade was output variable and consisted of five values: A, B, C, D and F, with A to C being a passing grade. They analyzed the success and failure of the students. Number of included students was 195, with 157 passes and 37 drops. Data set contained the following features: average and total number of interactions per session, whether the interaction is performed within a campus or outside the campus, the result of the first exam and the final grade of the course. The results show that Logistics Tree Model can serve as tool for prediction of dropout. In this way, it is possible to find out which students are at risk of failure.

Mayilvaganan and Kalpanadevi (2014) collected data about 197 students for the purpose of data analysis by means of data mining techniques. Student data includes the following attributes: specialty, previous grades, additional knowledge or skills, resources, class attendance, time spent learning, grade exam, seminar achievement, laboratory work, tests results and online assignments results. In this study, the discussion focuses on three classification techniques: decision tree, Naive Bayes, and k-nearest neighbor. The result concluded that the k-nearest neighbor had the best accuracy in the classification compared to other techniques due to the significance of the test results.

Saa (2016) also performed a study with objective to detect the relationship of personal and social factors of students on their educational performance by using data mining methodology. The data set used in this study was collected by a survey distributed to students in their daily classes and also as an online survey. The initial set of data set consisted of 270 records. Four decision tree algorithms have been implemented: Naive Bayes algorithm, ID3, C4.5 and CART. Interesting results from the classification models were extracted. It has been discovered that the performance of students is not entirely dependent on their academic efforts and there are many other factors that are equally influential.

Shahiria, Husain and Rashida (2015) claim that final grades are based on the structure of the course, assessment methods, results of the final exam and extracurricular activities. They applied several techniques for evaluating students' success. The results indicated neural networks as the method with the highest accuracy of predictions (98%), followed by decision trees (91%). Support vector machines and KNN gave the same accuracy of 83%, and Naive Bayes shown to be the method with the lower predictive accuracy (76%).

Based on the literature review, we have identified mostly used machine learning algorithms which will be used in our research.

3 Research framework

Graduation rates are important parameters and management of educational institutions are looking for new ways of predicting success and failure early enough in student education to achieve effective interventions as well as identifying the effectiveness of different interventions. The need to analyze the large amounts of data generated from the educational ecosystems urged development of educational data mining. Al-Barrak and Al-Razgan (2015) define educational data mining as a process of applying tools and techniques for data analysis in educational setting. The application of data mining techniques to educational data will help the education sector to improve its learning process. This is main motive of this paper.

There are two main elements in educational data mining: data set consisting of features which are

describing student activity and methods for analysis of data set.

Several methods are developed for predictive modelling. Depending on the task employed, methods are categorized into classification or regression methods. The most popular task of student performance prediction is mostly categorized as classification task. There are several algorithms in the classification task that are applied to predict student success. Literature review revealed decision tree, artificial neural network, logistic regression, and Support Vector Machine as mostly used methods.
An example of data stored in databases of educational institutions includes information on students enrollment on courses, their scores, activities, teacher notes, socio demographic characteristics. Once this data is properly analyzed, it will help improve knowledge in the education sector.
EDM can help universities to better plan the foreseen number of students enrolling in specific programs, predicting a dropout ratio, identification of students at risk of dropout, and better exploiting the available resources.
This will help the educational institutions to evaluate, plan and decide on their educational programs. It is expected that this new knowledge will reveal hidden forms that will help academic programs to use their resources more efficiently.

The system design in this research was carried out through four steps:
(i) data collection: students use LMS system and their interaction data is stored in the database. In this paper we have used data from students enrolled at the Moodle course.
(ii) data preprocessing: includes data cleaning and data transformation to the appropriate mining format. To pre-process Moodle data, we have used pre-processing possibilities of Azure tool.
(iii) modelilng: Four data mining algorithms were applied in order to develop a model that find patterns and summarizes knowledge to the teachers, administartors and students about students` behaviour.
(iv) system evaluation and interpretation: evaluate and implement results. The results obtained by model are interpreted and used by instructors for further procedures. The instructor can use disclosed information to make decisions about students' activities and the Moodle program to enhance student learning.
Next section explains implementations of this steps as well as model design and evaluation.

4 Model design and evaluation

This section explains empirical research based on the methodology described previously. Microsoft Azure Machine Learning platform for machine leraning is applied with the aim to develop predictive models. Four machine learning algorithms were employed to do so: logistic regression, neural networks, support vector machines and decision tree. First we will explain data set used for training and testing the model.

4.1 Data description

Data set used in this research was collected from the one course taught at the University of Zagreb, Faculty of Organization and Informatics. Data set consists of 235 records about students activities. Data were collected from LMS system Moodle. Range of transformations was made to the data set including feature selection and transformation of categorical features into numerical. Dimensionality of data set was 13, but only 11 variables was included in the data analysis since some of them were derived one from another. Description of the features is given in the Table A1. Furthermore, descriptive statistics for output variable is presented at Fig 1.

Statistics

Mean	1.4979
Median	1
Min	1
Max	5
Standard Deviation	0.993
Unique Values	5
Missing Values	0
Feature Type	Numeric Feature

Figure 1. Descriptive statistics

Output variable is binary and indicates pass or fail: fail = 0 and 1 = pass. Thus, we are dealing here with the classification since the mai aim is the prediction of binary target. Task of predicting a continuous target is referred to as a regression task (Kelleher, Mac Namee, D'Arcy, 2015).
Output variable is constructed based on the grades. Distribution of grades is given at Fig 2.

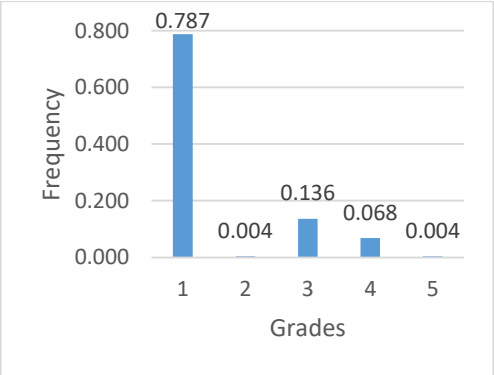


Figure 2. Distribution of grades

Following data preprocessing and data exploration phase, modelling was performed and implemented in Microsoft Azure ML cloud platform. Data set was split into trening (165 instances) and test data (70 instances). Results of the data modelling are presented in the next section.

4.2 Model description

Microsoft Azure Machine Learning served as platform for data model flow (Figure 3). First, data set was introduced. Then features were selected and input and output features were defined.

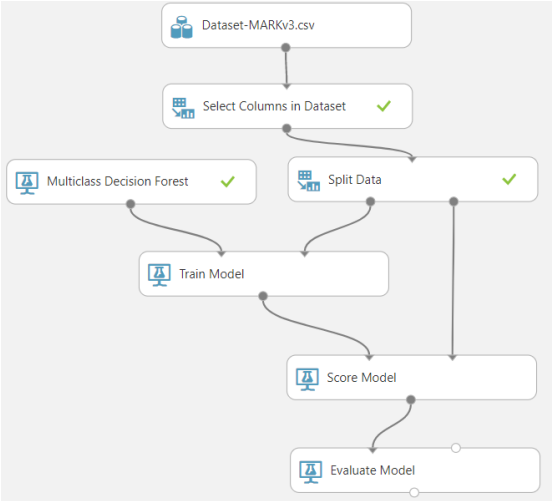


Figure 3. Data modeling flow

Data set is divided into the training set and the test set in a ratio of 70 % : 30 %. Training set is used to train the model while the test set is used to identify how precise a model is in generalizing over a new data set. Evaluation of the model completes the whole process by calculating accuracy, precision and recall measure. The results of the model training by using logistic regression are shown at figure 4. Logistic regression algorithm achieved a score of 92.85% accurate predictions.

Table 1. Performance scores of classification algorithms

Metrics/ Algorithm	*LR*	*NN*	*DT*	*SVM*
Overall accuracy	0,9286	0,9429	0,9714	0,9571
Average accuracy	0,9286	0,9429	0,9714	0,9571
Micro-averaged precision	0,9286	0,9429	0,9714	0,9571
Macro-averaged precision	0,8837	0,8839	0,9833	1
Micro-averaged recall	0,9286	0,9429	0,9714	0,8622
Macro-averaged recall	0,8577	0,9325	0,9167	0,7501

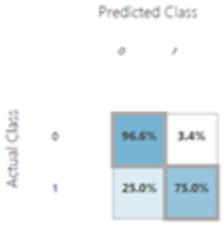


Figure 4. Confusion matrix of logistic regression model

Performance of the model developed by neural networks algorithm is shown at Figure 5. The neural network algorithm achieved a score of 94.28% accurate predictions.

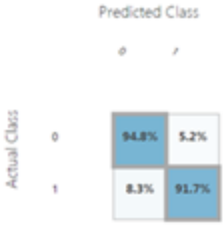


Figure 5. Confusion matrix of neural network model

Decision tree prediction model shown to be the most accurate model achieving 97.14% accurate predictions.

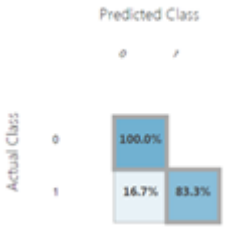


Figure 6. Confusion matrix of decision tree model

The results of the modelling by using support vector machine algorithm, are shown in Figure A1, attached to this paper. The support vector machine algorithm proved to be a very accurate model since achieved a score of 95.7% of accurate predictions.

4.3 Model interpretation

Sensitivity analysis is also performed in the research in order to identify predictors of academic performance. Seminar attendance and lecture attendance shown to be features with the highest impact on the passing or

failing the course in all machine learning models. Student attendance is a consistent source of discussion for researchers. Positive effect of attendance commonly reported in literature is noteable here. Our results are in line with previous work of Paisey and Paisey (2004). They identified strong positive relationship between attendance at classes and academic performance. However, Chen and Lin (2015) are sceptical and they consider that positive effects of an attendance considered in prior literature must be reassessed. Andrietti (2014) did find a positive and significant effect of attendance on academic performance. Furthermore, she used proxy variables regressions to capture the effect of unobservable student traits possibly correlated with attendance. As suggested in the research of Mearman et.al. (2014) wide range of factors affect attendance, such as the quality of teaching sessions or students aspirations. This must be investigated in future research. Our study highlights various interesting findings, but also opens questions that require further investigation.

Results regarding machine learning algorithms comparison emphasized decision trees as most accurate model. Decision tree has valuable characteristic of considering only those attributes that are helpful to the classification.

5 Conclusion

These paper deals with machine learning algorithms applications in educational domain with the aim to develop decision support system for identification of students at risk.
Through this paper, the basic concepts of four learning algorithms were presented: logistic regression, neural networks, support vector machines and decision tree are applied on real data set. Machine learning was applied in educational domain. Special emphasis was on model evaluation in order to compare different approaches to machine learning. Main research question was: in which extent the total data on students, generated by LMS Moodle, can be a good basis for predictive modelling. Such predictive models of student success based on Moodle activity serve as input into decision support system used in detecting behavior of future generations of students. Decision support system provides valuable tool for educational institutions in achievieng one of their main goals: increasing quality of the study. Management of educational institutions are potential users of such systems. At the beginning of the academic year, groups where students of similar profiles could be formed. Also, the system would briefly inform the relevant body about the student's performance during the key moments of the semester and highlight those students with whom a personalized program was needed for the purpose of studying. Results of this research yielded highly accurate models and provided basis for educational data driven decision making. However, there are several limitations of this research. First, models are developed using students` data from only course. Second, specific group of students was included, just informatics students. Thus, it is hard to generalize results. In the future research we will increase our sample and include students from different fields of study. Furthermore, other approaches to machine learning will be applied.

References

Andrietti, V. (2014). Does lecture attendance affect academic performance? Panel data evidence for introductory macroeconomics. International Review of Economics Education, 15, 1-16.

Al-Barrak, M. A., Al-Razgan, M. S. (2015) *Predicting Students' Performance Through Classification: A Case Study, Journal of Theoretical and Applied Information Technology*

Bhardwaj, B.K., Pal, S. (2011) Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4

Chen, J., & Lin, T. F. (2015). Effect of peer attendance on college students' learning outcomes in a microeconomics course. The Journal of Economic Education, 46(4), 350-359.

Corsatea, B., Walker, S. (2015) Opportunities for Moodle data and learning intelligence in virtual environments, School of Computer Science and Electronic Engineering, University of Essex, UK

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., Alowibdi, J. S. (2017) Predicting Student Performance using Advanced Learning Analytics, International World Wide Web Conference Committee (IW3C2)

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

Keshtkar, F., Cowart, J., Crutcher, A. (2018) Predicting Risk of Failure in Online Learning Platforms Using Machine Learning Algorithms for Modeling Students' Academic Performance, available at: http://medianetlab.ee.ucla.edu/papers/ICMLWS1.pdf

Mayilvaganan, M., Kalpanadevi D. (2014) Comparison of Classification Techniques for predicting the performance of Students Academic Environment, International Conference on

Communication and Network Technologies (ICCNT)

Mearman, A., Pacheco, G., Webber, D., Ivlevs, A., & Rahman, T. (2014). Understanding student attendance in business schools: An exploratory study. International Review of Economics Education, 17, 120-136.

Paisey, C., & Paisey, N. J. (2004). Student attendance in an accounting module–reasons for non-attendance and the effect on academic performance at a Scottish University. Accounting education, 13(sup1), 39-53.

Romero, C., Ventura, S., Garcia E. (2007) Data mining in course management systems: Moodle case study and tutorial, Department of Computer Sciences and Numerical Analisys, University of Cordoba, 14071 Cordoba, Spain

Saa, A. A. (2016) Educational Data Mining & Students' Performance Prediction, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5

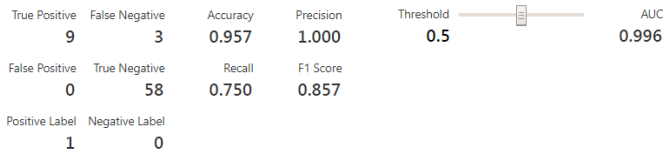
Shahiria, A. M., Husaina, W., Rashida, N. A. (2015) A Review on Predicting Student's Performance using Data Mining Techniques, School of Computer Sciences Universiti Sains Malayisa 11800 USM, Penang, Malaysia

Van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. Educause Learning Initiative, 1, 1–11. ELI Paper.

APPENDIX

Table A1. Description of features

Feature	Description	Values
Gender	Students gender	1 – female 2 – male
Status	Student status	1 – part time student 2 – full time student
Lecture_attendance	Frequency of lecture attendance	Numeric; scale: 0 – 6
Seminar_attendance	Frequency of seminar attendance	Numeric; scale: 0 – 4
Activity	Students activity at the seminars and lectures	Numeric; scale: 0 – 6
Mark_plan	Marketing plan as students assignment in the project	Numeric; scale: 0 – 25
Presentation	Presentation of the marketing plan	Numeric; scale: 0 – 10
Exam_1	First written exam	Numeric; scale: 0 – 25
Exam_2	Second written exam	Numeric; scale: 0 – 25
Assignemnt effort	Assignment for adittional points	Numeric; scale: 0 – 3
In class activity	Additional points for a course assignment	Numeric; scale: 0 – 3
Pass/Fail	Confirmation that the student has passed or the course	0 – Fail 1 – Pass



Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.829	0.000	1.000	0.000	0.829	1.000	0.000
(0.800,0.900]	1	0	0.014	0.843	0.154	1.000	0.083	0.841	1.000	0.000
(0.700,0.800]	1	0	0.029	0.857	0.286	1.000	0.167	0.853	1.000	0.000
(0.600,0.700]	6	0	0.114	0.943	0.800	1.000	0.667	0.935	1.000	0.000
(0.500,0.600]	1	0	0.129	0.957	0.857	1.000	0.750	0.951	1.000	0.000
(0.400,0.500]	2	3	0.200	0.943	0.846	0.786	0.917	0.982	0.948	0.047
(0.300,0.400]	1	3	0.257	0.914	0.800	0.667	1.000	1.000	0.897	0.099
(0.200,0.300]	0	6	0.343	0.829	0.667	0.500	1.000	1.000	0.793	0.203
(0.100,0.200]	0	17	0.586	0.586	0.453	0.293	1.000	1.000	0.500	0.496
(0.000,0.100]	0	29	1.000	0.171	0.293	0.171	1.000	1.000	0.000	0.996

Figure A1. Evaluation of support vector machines model