Data Mining for Student Outcome Prediction on Moodle: a systematic mapping

Igor Moreira Félix[1], **Ana Paula Ambrósio**[1],
Priscila da Silva Neves Lima[1], **Jacques Duilio Brancher**[2]

[1]Instituto de Informática – Universidade Federal de Goiás
Goiânia – Goiás – Brasil

[2]Departamento de Computação – Universidade Estadual de Londrina
Londrina – Paraná – Brazil

{igormoreira, apaula, prisciladasilva}@inf.ufg.br, jacques@uel.br

Abstract. *Virtual learning environments facilitate online learning, generating and storing large amounts of data during the learning/teaching process. This stored data enables extraction of valuable information using data mining. In this article, we present a systematic mapping, containing 42 papers, where data mining techniques are applied to predict students performance using Moodle data. Results show that decision trees are the most used classification approach. Furthermore, students interactions in forums are the main Moodle attribute analyzed by researchers.*

1. Introduction

Virtual Learning Environment (VLE) is a software that promotes distribution of online courses available on the Internet. These environments provide a number of facilities for managing distance learning courses, from delivering pedagogical content to monitoring student progress. Their tools include: forums, chats, educational resources and questionnaires [Romero et al. 2013b].

Moodle (Modular Object-Oriented Dynamic Learning Environment) is today the most widely used open source virtual learning environment for distance education around the world [EDUCAUSE 2014]. All interactions that occur inside Moodle are registered and stored in databases and logs [Moodle.org 2018]. They detail what activities were visualized and/or answered, as well as students' performance, registering when each resource is published, accessed, updated or removed. This data can be used to analyze students behavior, making it possible to monitor their progress [Romero and Ventura 2010], evaluate course structure, pedagogical activities and teacher performance.

Although this data can be partially visualized through reports and summaries, the information is scattered, creating difficulties for analysis. Furthermore, as the number of students has grown, teachers find it harder to deal with all this data in an appropriate way. Aware of this problem, some VLE, mostly proprietary, offer tools that allow data analysis. However, Moodle does not have an embedded analysis tool. To analyze data, users have relied on external software or plug-ins, available through community contributions.

This systematic mapping aims to identify what type of analysis is being used to predict students' performance on Moodle, using data mining. It is organized as follows:

DOI: 10.5753/cbie.sbie.2018.1393

Section 2 describes the mapping procedure adopted in this paper. Section 3 presents the results obtained through the answers to the research questions. Section 4 presents the conclusion by summarizing this research.

2. Systematic Mapping

A systematic mapping study is conducted to provide an overview of published research reports and their findings, categorizing them, giving a visual summary [Petersen et al. 2008]. This is an important resource for all knowledge domains, because often in most scientific researches, work starts with a deep analysis of the studied subject, so researchers can immerse in the state of the art in that domain.

The aim of this review is to analyze papers that applied data mining techniques or methodology to predict students performance in Moodle environment. For that, we have done a selection and inclusion of primary studies that are individual investigations presenting original research. Our general research question to be answered is:

"What type of analysis is being undertaken to predict students' performance on Moodle, using data mining techniques or methodology?"

Six specific subquestions were defined:

Q01. Which data mining techniques and methods were used?
Q02. Has the research developed some tool or presented only analysis results?
Q03. Which attributes were used?
Q04. What is the volume of data analyzed?
Q05. What is the accuracy obtained in students' performance prediction?
Q06. Which tools were used?

The search for studies was conducted in ten scientific databases, that have in their repository papers related to computational and technological areas: ACM Digital Library (dl.acm.org), CiteSeerX Library (citeseerx.ist.psu.edu), Keele University's Electronic Library (opac.keele.ac.uk), IEEE Xplore Digital Library (ieeexplore.ieee.org), LearnTechLib (www.learntechlib.org), Microsoft Research (www.microsoft.com/en-us/research), Science Direct (www.sciencedirect.com), Scopus (www.scopus.com), Semantic Scholar (www.semanticscholar.org), and Springer Link (link.springer.com).

The keywords used to filter papers in databases are: "moodle", "data mining", and "predict". For each database, a search string was defined based on its requirements. Were included in this systematic mapping, studies with an emphasis on: **I1.** Moodle data analysis; **I2.** Use of data mining; **I3.** Student performance prediction.

Furthermore, a Quality Criteria was adopted, based on the H-index of the publisher, obtained in Scimago Journal & Country Rank (www.scimagojr.com/journalrank.php). Only publishers with H-index higher than 5 were considered.

Were removed from the review, the primary studies that: **E1.** Are duplicated papers; **E2.** Are not accessible; **E3.** Are not a book chapter or journal/conference article.

Application of the research protocol recovered 1.006 papers. Their title and abstract were analyzed using the selection criteria. When this was not enough, the conclusion was also analyzed. 122 papers were selected for complete analysis. 42 were

selected to compose this systematic mapping. Figure 1 presents them organized according to database and publication type: (i) journal article, (ii) conference article, and (iii) conference short paper.

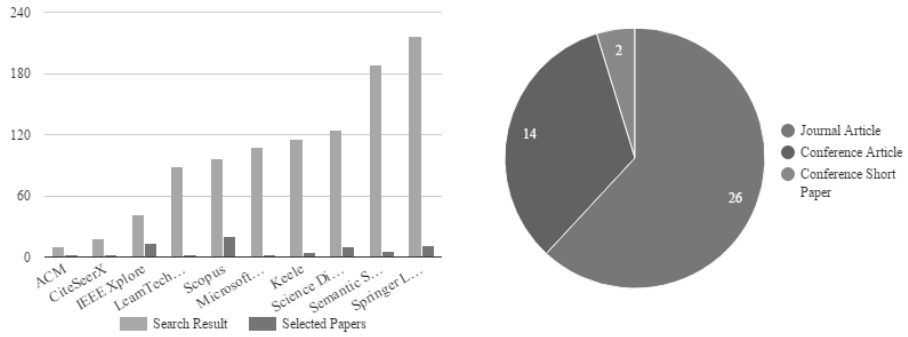

Figure 1. Search results X Selected papers and Papers' types

3. Results

Q01. *Which data mining techniques and methods were used?*

The selected papers are organized according to the method used (Table 1), grouped by general technique. Some appear more than once as they use more than one data mining technique or method. Methods for Classification are the most used for prediction (28 papers), followed by Clustering (11 papers), and Association (5 papers).

Some papers apply statistical techniques, like ***Linear Regression*** [You 2016], [Strang 2016], [Hu et al. 2014], [Černezel et al. 2014], [Kotsiantis 2012], [Kato and Ishikawa 2013], [Gasevic et al. 2016]; ***Stepwise Regression*** [Dascalu et al. 2016]; and ***Multilevel Mixed*** [Joksimović et al. 2015], [Joksimović et al. 2015].

The methods are implemented by several algorithms, and several were used in the selected papers. The most widely used Decision Tree algorithm is C4.5 [Zorrilla and Garcia-Saiz 2014], [Romero et al. 2013a], [Hu et al. 2014], followed by Simple Cart [Zorrilla and Garcia-Saiz 2014], [Romero et al. 2013a], [Hu et al. 2014], and Random Trees [Romero et al. 2008], [Romero et al. 2013a], [Hu et al. 2014], [Márquez-Vera et al. 2013].

Neural Networks algorithms include: ***Multilayer Perceptron*** [Romero et al. 2013a]; ***Radial Basis Function Network*** [Romero et al. 2013a]; and ***Fuzzy Learning*** [Shana and Abdulla 2015], [Romero et al. 2013a]. Other popular algorithms are: ***K-means clustering*** [Moradi et al. 2014], [Jovanovic et al. 2012], [Pardos et al. 2012], [Mogus et al. 2012], [Sorour et al. 2014]; ***K-nearest neighbor (KNN)*** [Zorrilla and Garcia-Saiz 2014], [Kotsiantis et al. 2010], [Minaei-Bidgoli et al. 2003], [Gamulin et al. 2016]; and ***JRip*** [Zorrilla and Garcia-Saiz 2014], [Márquez-Vera et al. 2013].

Q02. *Has the research developed some tool or presented only analysis results?*

As EDM is still a new research domain, the great majority of the papers present data mining analysis of specific situations, experimenting with different types of methods

Table 1. Papers according to data mining method

Classification	
Decision Trees	[Hung et al. 2016], [Zorrilla and Garcia-Saiz 2014], [Hu et al. 2014], [Márquez-Vera et al. 2013], [Romero et al. 2013a], [Jovanovic et al. 2012], [Kotsiantis 2012], [Minaei-Bidgoli et al. 2003], [Thai-Nghe et al. 2009], [Pardos et al. 2012], [Romero et al. 2008], [Sharma and Mavani 2011a]
Neural Network	[Gamulin et al. 2016], [Cambruzzi et al. 2015], [Shana and Abdulla 2015], [Sorour et al. 2014], [Romero et al. 2013a], [Kotsiantis 2012], [Sharma and Mavani 2011a], [Lykourentzou et al. 2009b], [Lykourentzou et al. 2009a]
Bayesian Classification	[Gamulin et al. 2016], [Zorrilla and Garcia-Saiz 2014], [Romero et al. 2013b], [Sharma and Mavani 2011b], [Sharma and Mavani 2011a], [Kotsiantis et al. 2010], [Thai-Nghe et al. 2009], [Minaei-Bidgoli et al. 2003]
Support Vector Machines	[Gamulin et al. 2016], [Kotsiantis 2012], [Lykourentzou et al. 2009b], [Thai-Nghe et al. 2009]
Genetic Algorithm	[Márquez-Vera et al. 2016], [Xing et al. 2015], [Romero et al. 2013a], [Zafra and Ventura 2012], [Zafra et al. 2011], [Zafra and Ventura 2009], [Minaei-Bidgoli et al. 2003]
Lazy Learning	[Gamulin et al. 2016], [Zorrilla and Garcia-Saiz 2014], [Kotsiantis et al. 2010], [Minaei-Bidgoli et al. 2003]
Rule-Based Classification	[Zorrilla and Garcia-Saiz 2014], [Márquez-Vera et al. 2013]
Clustering	
Partitioning Methods	[Gamulin et al. 2016], [Moradi et al. 2014], [Sorour et al. 2014], [Romero et al. 2013b], [Jovanovic et al. 2012], [Mogus et al. 2012], [López et al. 2012], [Pardos et al. 2012], [Kotsiantis et al. 2010], [Obadi et al. 2010], [Minaei-Bidgoli et al. 2003]
Association	
Apriori Algorithm	[Neto and Castro 2015], [Romero et al. 2013a], [Romero et al. 2009], [Carmona et al. 2010]
Regression	[Kotsiantis 2012]

and techniques (detailed in Q01). Total papers by analysis or system are presented in figure 2.

Few tools have been developed (Figure 2). These include [Cambruzzi et al. 2015] that implemented a system that allows to integrate different data sources, including Moodle database for dropout prediction applying Artificial Neural Networks. This tool was used in a distance education university, presenting satisfactory rates of correct predictions, enabling pedagogical actions to be taken. Another tool developed is presented by [Hu et al. 2014], that using decision trees and time-dependent variables, implemented an early warning web system, achieving high prediction accuracy.

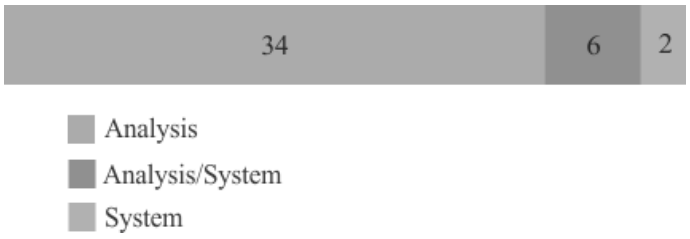

Figure 2. Total papers by analysis or system

Q03. *Which attributes were used?*

In the context of predicting students' performance using Moodle data, attributes are usually information related to interaction of students with other participants, e.g. forums; interaction with environment, e.g. access times; interaction with activities, e.g. submitted work or quiz answers. Other attributes often used include students' demographic data, like age and gender (Table 2).

Table 2. Papers according to attributes used to predict students performance

Attributes	Papers
Forums participation	[Dascalu et al. 2016], [Hung et al. 2016], [Gasevic et al. 2016], [Neto and Castro 2015], [Cambruzzi et al. 2015], [Hu et al. 2014], [Romero et al. 2013b], [Romero et al. 2013a], [Zafra and Ventura 2012], [López et al. 2012], [Jovanovic et al. 2012], [Mogus et al. 2012], [Zafra et al. 2011], [Obadi et al. 2010], [Carmona et al. 2010], [Zafra and Ventura 2009], [Romero et al. 2009]
Assessment data/grades	[Kostopoulos et al. 2015], [Romero et al. 2009], [Lykourentzou et al. 2009b], [Hu et al. 2014], [Gasevic et al. 2016], [You 2016], [Kotsiantis 2012], [Moradi et al. 2014], [Jovanovic et al. 2012], [Romero et al. 2013a], [Černezel et al. 2014], [Pardos et al. 2012], [Carmona et al. 2010], [Hung et al. 2016]
Interaction logs	[Joksimović et al. 2015], [Xing et al. 2015], [Kotsiantis et al. 2010], [Zacharis 2015], [You 2016], [Zorrilla and Garcia-Saiz 2014], [Cambruzzi et al. 2015], [Gamulin et al. 2016], [Sharma and Mavani 2011a], [Sorour et al. 2014], [Romero et al. 2008], [Sharma and Mavani 2011b]
Quizzes data	[Kato and Ishikawa 2013], [Zafra et al. 2011], [Lykourentzou et al. 2009a], [Zafra and Ventura 2012], [Gasevic et al. 2016], [Jovanovic et al. 2012], [Strang 2016], [Romero et al. 2013a], [Obadi et al. 2010], [Carmona et al. 2010], [Zafra and Ventura 2009]
Access logs	[Hu et al. 2014], [Gasevic et al. 2016], [Strang 2016], [Romero et al. 2013a], [Neto and Castro 2015], [Minaei-Bidgoli et al. 2003]
Resources logs	[Hu et al. 2014], [Gasevic et al. 2016], [Strang 2016], [Mogus et al. 2012], [Obadi et al. 2010], [Hung et al. 2016]
Tasks data	[Romero et al. 2009], [Zafra et al. 2011], [Zafra and Ventura 2012], [Romero et al. 2013a], [Černezel et al. 2014], [Neto and Castro 2015], [Minaei-Bidgoli et al. 2003], [Carmona et al. 2010], [Zafra and Ventura 2009]
Chats logs	[Romero et al. 2009], [Gasevic et al. 2016], [You 2016], [Neto and Castro 2015], [Carmona et al. 2010]
Academic registers	[Lykourentzou et al. 2009b], [Márquez-Vera et al. 2013], [Márquez-Vera et al. 2016], [Hung et al. 2016], [Shana and Abdulla 2015]
Demographic data	[Kostopoulos et al. 2015], [Lykourentzou et al. 2009b], [Gasevic et al. 2016], [You 2016], [Kotsiantis 2012], [Márquez-Vera et al. 2013], [Márquez-Vera et al. 2016], [Strang 2016], [Mogus et al. 2012], [Hung et al. 2016]

Q04. *What is the volume of data analyzed?*

The number of students used in the papers vary a lot, with [Thai-Nghe et al. 2009] using data for more than 10.000 students (Figure 3).

Figure 3 gives an overview about diversified volume of students applied in studies. Most papers collected students data from universities (38 researches), while in 3 papers the data analyzed comes from students in high school.

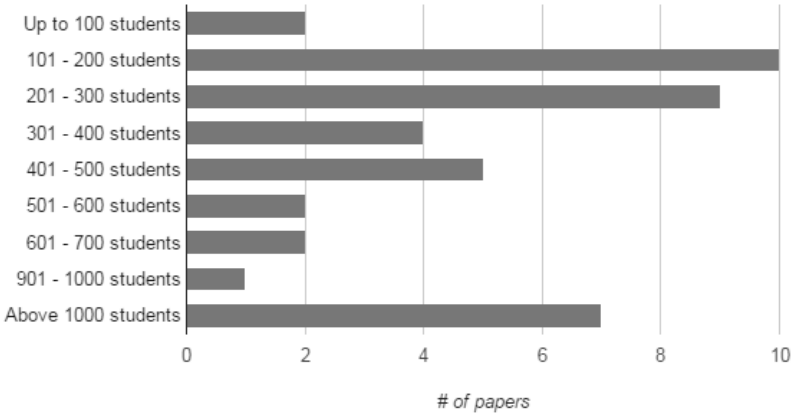

Figure 3. Total papers by number of students used in the analysis

Q05. *What is the accuracy obtained in students' performance prediction?*

The accuracy indicates the percentage of correctly classified instances using the data mining algorithm. In the case of students' prediction, this rate indicates the percentage of students whose outcome (failed or passed, completed or dropout) was successfully predicted.

Two of the selected papers did not provide this information. Most of the other 40 papers reported an accuracy between 75% and 100%, as seen in figure 4.

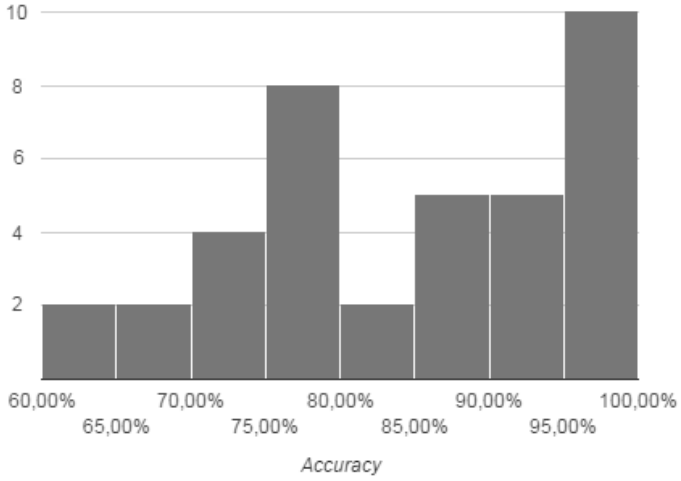

Figure 4. Total papers by accuracy

Q06. *Which tools were used?*

Several data mining support tools are available to the research community. Some are free and open source, while others are private and expensive.

WEKA (Waikato Environment for Knowledge Analysis - www.cs.waikato.ac.nz/ml/weka) is the most widely used tool by researchers to conduct performance prediction, being explicitly presented by 14 papers in their methodology. In second is KEEL (Knowledge Extraction based on Evolutionary Learning - www.keel.es), used by 5 researches. Other tools that have been employed are: KNIME (2 papers),

MatLab (3 papers), SPSS and R. Twelve papers did not inform what tool was used to conduct their research.

4. Conclusion

Students performance prediction allows teachers and institutions to monitor students progress, so early intervention can be made and to improve learning process. This is specially important in online courses that present high failure and dropout rates. The systematic mapping study presented in this paper offers to researchers a starting point that allows them to have an overview of this domain.

This mapping has analyzed papers focused on students' performance prediction using Moodle data, including 42 publications. Results show that relevant research in this domain has steadily grown in high impact vehicles, which highlights the importance of the topic and it has attracted attention. Analysis is presented according to the protocol questions that guided the mapping.

As a new research domain, EDM still doesn't have established approaches and protocols. This is verified by the high number of papers focused on analysis of specific situations, that serve to define the feasibility of the area and best practices which may lead to the definition of appropriated approaches to students outcome prediction. Only eight papers described an implemented system for prediction.

Data used in the analysis, includes mainly Moodle attributes that can be used raw or processed by text analysis and statistical tools. Commonly used attributes include: interaction logs, forum participation, grades and evaluation data, quizzes, and others. The most used attributes by 42 papers included in this mapping (Q02) involved forums analysis, examples: number of posts, total of posts reads, number of enjoyed discussions and others. While the highest accuracy levels were based on interaction logs attributes [Sharma and Mavani 2011a, Xing et al. 2015, Joksimović et al. 2015, Gamulin et al. 2016] and demographic data [Márquez-Vera et al. 2016, Lykourentzou et al. 2009b, Márquez-Vera et al. 2013].

One of the main questions in this domain is "what is the best algorithm to be used?" and its answer depends on data analyzed and the goals of analysis. The 42 papers included in this systematic mapping, have presented a variety of applied algorithms, but the most part related the use of classification methods to their educational data mining process, such as decision trees (C4.5, random forests, and others). And these methods show to be appropriated to students' prediction on Moodle data, because between the 10 highest accuracy (95% - 100%) presented by publications, 9 are reached through classification methods [Márquez-Vera et al. 2016, Sharma and Mavani 2011a, Lykourentzou et al. 2009b, Shana and Abdulla 2015, Márquez-Vera et al. 2013, Xing et al. 2015, Gamulin et al. 2016, Lykourentzou et al. 2009a, Hu et al. 2014].

However these results must be further analyzed. Often, the number of failing students is much smaller than successful students, leading to unbalanced datasets. In these cases, the use of only accuracy measures can be misleading. Other measures such as confusion matrix and Area Under ROC curve must also be considered. Also, balancing techniques can be used to correct dataset imbalance.

All educational institutions that use Moodle to support learning process, have in

their databases valuable data for useful and interesting analysis, with accessible tools. In fact, Moodle and the most widely used data mining software, including WEKA, KEEL and R, are free and open source, facilitating implementations. It is now a question of testing different approaches that will lead to an implemented system, allowing non experts to benefit of this information and help at risk students through early intervention.

References

Cambruzzi, W., Rigo, S., and Barbosa, J. (2015). Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach. *Journal of Universal Computer Science*, 21(1):23–47.

Carmona, C. J., González, P., del Jesus, M. J., Romero, C., and Ventura, S. (2010). Evolutionary algorithms for subgroup discovery applied to e-learning data. In *IEEE EDUCON 2010 Conference*, pages 983–990.

Černezel, A., Karakatič, S., Brumen, B., and Podgorelec, V. (2014). *Predicting Grades Based on Students' Online Course Activities*, pages 108–117. Cham.

Dascalu, M., Popescu, E., Becheru, A., Crossley, S., and Trausan-Matu, S. (2016). *Predicting Academic Performance Based on Students' Blog and Microblog Posts*, pages 370–376. Cham.

EDUCAUSE (2014). The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives. Accessed 27/07/2016.

Gamulin, J., Gamulin, O., and Kermek, D. (2016). Using fourier coefficients in time series analysis for student performance prediction in blended learning environments. *Expert Systems*, 33(2):189–200. EXSY-Aug-14-172.R2.

Gasevic, D., Dawson, S., Rogers, T., and Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28:68 – 84.

Hu, Y.-H., Lo, C.-L., and Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 36:469 – 478.

Hung, J. L., Wang, M., Wang, S., Abdelrasoul, M., y. li, and He, W. (2016). Identifying at-risk students for early interventions? a time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, PP(99):1–1.

Joksimović, S., Gašević, D., Loughin, T. M., Kovanović, V., and Hatala, M. (2015). Learning at distance: Effects of interaction traces on academic achievement. *Computers & Education*, 87:204 – 217.

Jovanovic, M., Vukicevic, M., Milovanovic, M., and Minovic, M. (2012). Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems*, 5(3).

Kato, T. and Ishikawa, T. (2013). *Detection and Presentation of Failure of Learning from Quiz Responses in Course Management Systems*, pages 64–73. Cham.

Kostopoulos, G., Kotsiantis, S., and Pintelas, P. (2015). *Predicting Student Performance in Distance Higher Education Using Semi-supervised Techniques*, pages 259–270. Cham.

Kotsiantis, S., Patriarcheas, K., and Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, 23(6):529 – 535.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*.

López, M. I., Romero, C., Ventura, S., and Luna, J. (2012). Classification via clustering for predicting final marks starting from the student participation in forums. In *EDM*.

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., and Loumos, V. (2009a). Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology*.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., and Loumos, V. (2009b). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3):950 – 965.

Márquez-Vera, C., Cano, A., Romero, C., and Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, 38(3):315–330.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., and Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003.*, volume 1.

Mogus, A. M., Djurdjevic, I., and Suvak, N. (2012). The impact of student activity in a virtual learning environment on their final mark. *Active Learning in Higher Education*.

Moodle.org (2018). Moodle Philosophy. `https://docs.moodle.org/24/en/Philosophy`. Accessed: 2018-03-06.

Moradi, H., Moradi, S. A., and Kashani, L. (2014). *Students' Performance Prediction Using Multi-Channel Decision Fusion*. Cham.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124. EXSY-Dec-13-227.R3.

Neto, F. A. A. and Castro, A. (2015). Elicited and mined rules for dropout prevention in online courses. In *2015 IEEE Frontiers in Education Conference (FIE)*, pages 1–7.

Obadi, G., Dráždilová, P., Martinovic, J., Slaninová, K., and Snášel, V. (2010). Finding patterns of students' behavior in synthetic social networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 411–413.

Pardos, Z. A., Wang, Q. Y., and Trivedi, S. (2012). The real world significance of performance prediction. *ICEDM Proceedings*, 1(5):192–195.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swindon, UK. BCS Learning & Development Ltd.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., and Ventura, S. (2013a). Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146.

Romero, C., González, P., Ventura, S., del Jesús, M. J., and Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using moodle data. *Expert Syst. Appl.*, 36:1632–1644.

Romero, C., López, M.-I., Luna, J.-M., and Ventura, S. (2013b). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*.

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Romero, C., Ventura, S., Espejo, P. G., and Hervás, C. (2008). Data mining algorithms to classify students. In *In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187191, 2008. 49 Data Mining 2009*.

Shana, Z. and Abdulla, S. (2015). Educational data mining: an intelligent system to predict student graduation agpa. *International Review on Computers and Software (IRECOS)*, 10(6):593–601.

Sharma, M. and Mavani, M. (2011a). Accuracy comparison of predictive algorithms of data mining: Application in education sector. *Communications in Computer and Information Science*, 125 CCIS:189–194. cited By 0.

Sharma, M. and Mavani, M. (2011b). Development of predictive model in education system: Using naÏve bayes classifier. In *Proceedings of the International Conference - Workshop on Emerging Trends in Technology*, ICWET '11, pages 185–186, New York, NY, USA. ACM.

Sorour, S. E., Mine, T., Goda, K., and Hirokawa, S. (2014). Predicting students' grades based on free style comments data by artificial neural network. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pages 1–9.

Strang, K. D. (2016). Beyond engagement analytics: which online mixed-data factors predict student learning outcomes? *Education and Information Technologies*.

Thai-Nghe, N., Busche, A., and Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 878–883.

Xing, W., Guo, R., Petakovic, E., and Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, 47:168 – 181.

You, J. W. (2016). Identifying significant indicators using lms data to predict course achievement in online learning. *The Internet and Higher Education*, 29:23 – 30.

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, 27:44 – 53.

Zafra, A., Romero, C., and Ventura, S. (2011). Multiple instance learning for classifying students in learning management systems. *Expert Systems with Applications*, 38(12):15020 – 15031.

Zafra, A. and Ventura, S. (2009). Predicting student grades in learning management systems with multiple instance genetic programming. In *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*, pages 307–314.

Zafra, A. and Ventura, S. (2012). Multi-instance genetic programming for predicting student performance in web based educational environments. *Applied Soft Computing*, 12(8):2693 – 2706.

Zorrilla, M. and Garcia-Saiz, D. (2014). Meta-learning: Can it be suitable to automatise the kdd process for the educational domain? *Lecture Notes in Computer Science*, pages 285–292.