Predicting University Students' Grades Based on Previous Academic Achievements

Maria Tsiakmaki
Department of Mathematics
University of Patras
Patras, Greece
m.tsiakmaki@gmail.com

Georgios Kostopoulos
Department of Mathematics
University of Patras
Patras, Greece
kostg@sch.gr

Giannis Koutsonikos
Department of Business Administration
T.E.I. of Western Greece
Patras, Greece
gkoutson@gmail.com

Christos Pierrakeas
Department of Business Administration
T.E.I. of Western Greece
Patras, Greece
pierrakeas@teipat.gr

Sotiris Kotsiantis
Department of Mathematics
University of Patras
Patras, Greece
sotos@math.upatras.gr

Omiros Ragos
Department of Mathematics
University of Patras
Patras, Greece
ragos@math.upatras.gr

***Abstract*—Several data mining techniques are currently applied in many areas with great success. Various educational systems collect and use huge amount of data on students, staff and faculty. Researchers have the potential to employ such datasets to examine students' performance over their learning time—from one semester to another or from one academic year to another. Student's key demographic characteristics, number of examination attempts per course and final grade of each course can provide the training data and infer a regression function that estimates the performance of upcoming courses. We have carried out several experiments using eight courses modeled for regression tasks and familiar data mining models (i.e. Linear Regression, Support Vector Machines, Decision Trees, M5 Rules, and k-Nearest Neighbors). The results reported that satisfactory accuracy is achieved, provided that the first semester grades are available, indicating that an early identification of students at risk of underperforming in a course can be obtained.**

Keywords—educational data mining; learning analytics; regression; grades prediction

I. INTRODUCTION

Large amounts of data within educational settings are becoming more and more available [1]. Most of these data emanate from institutional student information systems, virtual learning environments, attendance monitoring systems (e.g. "card swipes") and library systems [2]. These systems can record any student activity that is supported, such as reading, writing, exams taking, tasks performed and peer communications, information on staff, content and the institution and so forth. The analysis of these datasets attracted much attention among researchers for its potential to identify patterns, trends and predictions that can be used to optimize the learning process and its outcomes.

Data collected from learning systems tend to be large and may contain many features that data mining algorithms explore for model building [3]. Preprocessing and data mining techniques (such as statistics, visualization, classification, clustering and association rule mining) can be applied, providing useful information and interesting automated tools that aid the education domain [4, 5, 6]. The development of methods for exploring these types of data is the main purpose of the Educational Data Mining (EDM) [7, 8].

As a discipline that has emerged in recent years, EDM applies specific knowledge discovery process, adapted to treat the special characteristics and objectives of this mining problem. According to Romero and Ventura, EDM is an interdisciplinary area, drawn as the intersection of computer science, education and statistics areas [9]. It is generally looking for new patterns in data and developing new algorithms and/or models. An area that is most closely to EDM is Learning Analytics (LA).

LA is focused on data-driven decision-making. It applies known predictive models to turn the unstructured, raw data into actionable, high-level information. LA uses high-level knowledge to improve education. Sclater identifies four main areas of learning analytics applications that have emerged driven by this belief [10]. The early prediction of future performance and identification of students at risk are indicated as the most common applications and describe the first area. The second area is the course recommendation, dealing with the evaluation of available learning material in order to suggest particular courses and sequence for learners. The third one is the adaptive learning systems, where the predictions are used to adjust and personalize the future learning content, while the last one concerns the curriculum design and just-in-time pedagogic strategies. The scope of the present study concerns the first area of the applications.

The ability to predict students' learning difficulties is a significant task and has different interest groups: students, tutors, administrators and educational leaders [4, 11]. It is beneficial for students that are in danger of falling behind [11], as appropriate actions could be provided leading to the improvement of their performance and learning outcomes [12]. Even effective decisions about implementing beneficial change improvements at the institutional level could be made [13].

978-1-5386-8161-9/18/$31.00 ©2018 European Union

In the present study, we analyzed recent data provided by the Department of Business Administration of the Technological Educational Institute (TEI) of Western Greece. Several state of the art regression techniques were applied for predicting undergraduate students' marks in the final examinations of a second semester course, given the grades from the first semester courses, the number of unsuccessful attempts to pass the corresponding courses in previous semester examinations and selected demographic variables. The aim of the study is to predict student grades and, if possible, to identify the key variables that improve the performance of the results.

The rest of this paper is organized as follows: The background of learning analytics is briefly described in Section II. Section III presents recent studies regarding the specified area. Section IV presents our data collection and preprocessing tasks. In Section V our experimental evaluation and results are described. Finally, Section VI concludes the study considering some future research directions.

II. LEARNING ANALYTICS

Educational data mining, academic analytics and learning analytics form a set of related and divergent fields that have been widely used in the research area of Technology-Enhanced Learning. Educational data mining is a discipline based on data mining and machine learning techniques exploring data from educational settings to address important educational issues [9]. Even more, it concerns the development of methods to better understand students' learning behavior and the settings which they learn in [14]. At the same time, it adds new dimensions to be taken into consideration by examining patterns and predictions that describe students' accomplishments, content, evaluations, learning processes and applications.

Meanwhile, the term analytics combines big data sets, statistics and predictive modelling [15]. It is mainly focused on quantitative analysis. Additionally, the field of business intelligence covers the data-driven decision making and the achievement of desired business results. These terms are also related to educational challenges, through academic and learning analytics.

Academic analytics is most often applied in education for improving administrative decisions and performance measuring [15, 12]. They can even be used to benchmark one's learning processes and performance metrics from other institutions [12].

While academic analytics reflects the role of data analysis at an institutional level, learning analytics focuses on the learning process, which provides better knowledge and understanding of the relationship between learner, content, institution, and educator [3]. According to the 1st International Conference on Learning Analytics and Knowledge (LAK 2011), Learning Analytics "*is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs*" [16].

A. A growing field

Learning analytics gather multi-disciplinary and complementary approaches from different fields. It incorporates concepts and techniques from information science, sociology, psychology, statistics, machine learning and data mining [17]. Ferguson argues that it has roots in many fields, such as business intelligence, web analytics, educational data mining and recommender systems [11], whereas Johnson et al. [18] simply connect it with web analytics. Even more, the "Research Evidence on the Use of Learning Analytics: Implications for Education Policy" report, published by the European Commission's Science and Knowledge Service in 2016 [19], acknowledges its connections with assessment, personal learning and social learning. Subsets of the field such as predictive analytics, social learning analytics and discourse analytics are also identified [11].

The Higher Education Edition of the NMC Horizon Report in 2016 presented an estimate of one year or less for the adoption of Learning Analytics [18]. The same edition in 2017 [20] goes even further by reporting measuring learning as a new trend for the next three to five years, where learning analytics is the fundamental assessment tool for learning outcomes. Sclater et al. [2] refer to learning analytics as a new field that appears to be growing rapidly. Campbell et al. [15] state that analytics will likely grow in various significant areas, Ferguson [11] depicts it as a field in its own right.

The last author also gives credit to three principal factors for the emergence of learning analytics as a field: (a) big data (i.e. the increasingly large amounts of data from learning management systems, and other related systems), (b) online learning (i.e. the worldwide increased use of online and blended teaching and learning), and (c) national concerns (i.e. the growing interest of countries in measuring and improving education).

Also the JRC-led study [19], which gathered evidence of implementation of learning analytics from five case studies and the LAEP inventory of 60 tools, practices and policies in the field[1], agrees with the above. The study also concludes that analytics enables a new type of support for teachers, school leaders and other educational staff.

Various goals are achieved using learning analytics. Generally, it is a tool for improving learning and teaching [2]. Such a tool may be able to build better pedagogies, promote active learning, identify at-risk students, and assess factors that influence student success and retention [18]. Ifenthaler et al. [1] state that analytics can be used for real-time modeling, prediction and optimization of learning processes. The study [19] notably illustrates a variety of implemented tools from different continents and educational levels, lists also objectives such as the identification of isolated students, group malfunction and student inequalities, the reduction of the time-to-degree, the improvement of retention, academic performance, persistence and graduation rates. Jayaprakash et al. [21] also indicate the gain in decision making in the admission process, financial and operational efficiency.

B. Overlaps and diverges between the fields

While overlaps between educational data mining, academic analytics, and learning analytics remain [11], the variety of aims and practices separate these three groupings. Ferguson

[1] LAEP Inventory - http://cloudworks.ac.uk/cloudscape/view/2959

[11] highlights the different challenges that each field addresses. According to her research, educational data mining is more focused on the technical challenge of extracting value from learning-related datasets. Learning analytics focuses on the educational challenge of optimizing learning opportunities, and academic analytics on the political/economic challenge of improving educational results at national or international levels.

Apart from the above different challenges that these sets face, the issue brief [17] also concludes that educational data mining produces new methods and models, while learning analytics uses existing. Mining promotes new discoveries based on data (e.g. better understanding of educational systems and learning, pattern discovery, solve educational problems), while analytics offer more actionable information (e.g. provide feedback to students and tutors, proactive intervention techniques, customize learning paths).

C. *Datasets and applications of learning analytics*

As learning analytics is concerned with the analysis of the relationships between learner, content, institution, and educator, for ultimately improving learner success, behavior-specific data, that come from the educational context, are gathered and explored [13]. Similarly, [12] argues that learning analytics targets data related to learners' interactions with course content, other students, and instructors. Some tools also employ third party data such as social media and statistical services [19].

Various types of data are collected during educational processes and analyzed. There are three main types of data: (a) data related to student's usage and interaction with educational systems, (b) data about the courses and the curriculum, and (c) academic information [11, 17, 9]. The collected data contain many variables, such as student's entrie, course assessments, discussion board, blog and wiki entries, number of emails sent, number of quizzes completed.

Learning analytics tools and techniques are designed and maintained for multiple purposes. For example, they indicate students that need support and issue warnings to students and tutors about inadequate performance [19]. The same report continues that such type of applications can make predictions about the future behavior of learners and resources recommendations. A majority of tools incorporate visual representations, summaries and dashboards to generate actionable analyses [17]. These analyses can be used to tailor content to student's needs, intervene with at-risk learners, provide individual feedback and influence educational practice.

III. RELATED WORK

Various studies provide evidence of the use and impact of learning analytics on higher education.

The existence of gendered performance differences in Science, Math, Engineering, and Technology (SMET) disciplines is observed in [22]. To introduce the conducted analysis, they compared student grades in one chosen course to their GPAO, i.e. a credit hour weighted average of grades a student has received in all other classes. Their dataset contained (627,998) students from the University of Michigan in 116 courses from a wide range of subjects. Apart from exhibiting performance differences between the two genders, they also report patterns of grade anomaly from differences in grading practice among the disciplines. On average, students that choose STEM courses face significant grade penalties as they enter the field, due to their known demanding philosophy. For female students these penalties were reported to be larger than those of men. Finally, the authors suggest actions that could reduce these performance gaps.

The importance of understanding the course and disciplinary context when developing and interpreting predictive models for identifying student's academic performance is demonstrated at [23]. The data were collected from nine courses with representative diversity of disciplines, and differences in the use of the learning management system. The dataset included various student characteristics (such as age, gender, nationality, spoken language, if previous enrollment, the course start access) and trace data from the LSM (such as the usage of forums, course logins, resources, file submission, assignments, book, quizzes, feedback, map, virtual classroom, lessons, and chat). Multiple linear and logistic regression models were performed for each course to explore the associations between students' characteristics or trace data-based variables and student performance across courses. The results demonstrated significant differences in the association between student characteristics and trace data variables and student percent marks among courses.

Personalized regression based methods and matrix factorization approaches based on recommender systems where used in [24] to forecast students' grades in future courses and in-class assessments. Briefly, the first method that they investigated was the course-specific regression (CSpR), which predicted the grade that a student will achieve in a specific course as a sparse linear combination of the grades that the student obtained in past courses. The second method was the personalized linear multi-regression (PLMR), used a linear combination of k regression models, which were weighted on a per-student basis. The third method was a standard matrix factorization (MF) approach that approximated the observed entries of the student–course grade matrix. The fourth method was a course-specific matrix factorization (CSpMF) approach that relied only on the subset of the data used by CSpR in order to estimate an MF model that is specific to each course. The evaluations showed that the FM and PLMR methods produce low error rates for the next-term grade prediction.

The authors in [25] introduce a students' performance model based on data mining techniques with a new category of features, which they call behavioral features. These types of features are related to the learner's interaction with the e-learning system (such as discussion groups, visited resources, raised hand in class, viewing announcements) and parents participation in the learning process (parent answering survey, parent school satisfaction degree). The performance of students' predictive model was evaluated by using traditional classifiers: neural networks, naive Bayes and decision trees. They also applied ensemble methods to improve the results: Bagging, Boosting, and Random Forest. Among the results, it was proved that the classification results were significantly improved when the classifier used the behavioral features. Also, ANN model outperformed the other mining techniques. It

was also proved that all ensemble methods improved the predictions for almost every model, with Boosting to be the outperforming method.

Predicting student performance is also the main focus of [26] study. At first the researchers highlight the key challenges for an effective predictor, i.e. (a) the heterogeneous student background and chosen specializations, (b) the different degree of influence of each course, and (c) the continuous need for updating the model each time a new degree is available. Their goal was to predict the final cumulative GPA of a student, given his/her background and performance states of the known grades and the predictions for the courses that has not been taken. For enabling such progressive predictions, the authors proposed a two-layer architecture. The first layer implements the base predictors for each course given the performance state of graduate students on courses relevant to the targeted course. For discovering the relevant courses, a course clustering method was developed. In the second layer, ensemble-based predictors are developed, able to keep improving themselves by accumulating new student data over time. Four classic machine learning algorithms were implemented: linear regression, logistic regression, random forest and kNN, among which random forest performed the best.

IV. DATA DESCRIPTION

The dataset used in our study has been provided by the Business Administration department of the Technological Educational Institution of Western Greece. For a time period of four years (2013-2017), data of 592 anonymized students have been collected concerning mainly their academic performance during the first two semesters. Each of these semesters includes six courses and two laboratory courses, while students are obliged to attend the final examinations of the course at the end of the semester. The final exam is marked out of 10, and the successful completion of a course requires a grade of at least five in the course examinations.

TABLE I. INPUT ATTRIBUTES

Attribute	Type	Values	Description
Gender	nominal	m, f	Gender of Student
WayImport	integer	[1, 38]	Way that students entered the Department
A1	real	[0, 10]	Introduction to Information Technology
A2	real	[0, 10]	Introduction to Information Technology (Lab)
A3	real	[0, 10]	Business Administration-Management
A4	real	[0, 10]	Accounting
A5	real	[0, 10]	Accounting (Lab)
A6	real	[0, 10]	Introduction to Law
A7	real	[0, 10]	Mathematics
A8	real	[0, 10]	Introduction to Marketing
A1-fails	integer	0,1,2,3	Unsuccessful attempts to pass the corresponding course in previous semesters' examinations
A2-fails	integer	0,1,2,3	
A3-fails	integer	0,1,2,3	
A4-fails	integer	0,1,2,3	
A5-fails	integer	0,1,2,3	
A6-fails	integer	0,1,2,3	
A7-fails	integer	0,1,2,3	
A8-fails	integer	0,1,2,3	

Each instance in the dataset is characterized by the values of 18 attributes (Table I). Regarding students' characteristics, two attributes were used: "Gender" and "WayImport". The dataset distribution of male and female students was 55% and 45% respectively, while we also took into account the system of import in the department. Attributes Ai, i=1,2,…,8 correspond to first semester course grades, while attributes Ai-fails, i=1,2,…,8 correspond to the number of unsuccessful attempts to pass the corresponding courses in previous semester examinations.

The purpose of our study is to predict students' grades in the final examinations of a second semester' course, given the values of the aforementioned input attributes. Since the second semester includes six courses, we carried out a plethora of experiments applying familiar regression methods to predict the grade of each course Bi, i=1,2,…,6 (Table II).

TABLE II. OUTPUT ATTRIBUTES

Attribute	Type	Values	Description
B1	real	[0, 10]	Business Accounting
B2	real	[0, 10]	Fundamental Algorithms-Structured Programming
B3	real	[0, 10]	Introduction to Business Statistics
B4	real	[0, 10]	Commercial Law
B5	real	[0, 10]	Microeconomics
B6	real	[0, 10]	Financial Mathematics

V. EXPERIMENTAL SETUP AND RESULTS

A number of experiments were conducted in six distinct phases depending on the output attribute Bi, i=1,2,…,6. In each phase we evaluate the performance of familiar regression methods for predicting the grade to Bi course (i=1,2,…,6) based on the input attributes referred in the previous section IV. In particular, the following regression methods were applied using the Weka Machine Learning environment [27]:

- Linear Regression (LR), a widely used statistical approach for describing a continuous output variable y associated with a number of independent variables x_i, i=1,2,…,n [28].
- Random Forests (RF), a collection of tree-structured classifiers that have been implemented quite effectively for both classification and regression problems [29].
- Instance-based Regression (5NN) [30].
- M5 Rules, a rule learning algorithm for inducing rule sets from model trees [31].
- M5 algorithm, a popular method for inducing model trees [32].
- The Sequential Minimal Optimization algorithm for regression problems using Support Vector Machines (SMOreg) [33].
- Gaussian processes (GP), a stochastic process consisting of random variables [34].
- Bootstrap Aggregating (Bagging), a method using aggregation averages of multiple versions of a predictor for predicting a numerical outcome [35].

At first, the dataset was partitioned into 10 folds using the 10-cross validation procedure so that each fold had the same

distribution as the whole dataset. Nine folds were used for the training process, while one fold was kept for evaluating the predictive efficiency of each method. The evaluation metric used in our study for determining the efficiency of each regression method is the Mean Absolute Error (MAE).

The MAE values corresponding to the abovementioned regression methods in each one of the experiments phases are presented in Table III. At first glance it is observed that RF, Bagging and SMOreg take precedence over the rest methods with a MAE value ranging from 1.217 to 1.943 depending on the output attribute.

TABLE III. MAE RESULTS

Regression Method	Course of second semester to predict grade					
	B1	**B2**	**B3**	**B4**	**B5**	**B6**
LR	1.921	1.858	1.363	1.250	1.659	1.342
RF	1.910	1.779	1.424	1.198	1.594	1.286
SMOreg	1.907	1.801	1.385	1.207	1.623	1.266
5NN (Manhattan)	2.046	1.878	1.515	1.285	1.700	1.359
M5' Rules	1.924	1.939	1.607	1.384	1.760	1.438
M5	1.991	1.924	1.618	1.371	1.733	1.411
GP	1.923	1.902	1.487	1.237	1.690	1.350
Bagging	1.943	1.850	1.440	1.217	1.537	1.288

To compare the performance of the regression methods, the Friedman Aligned Ranks nonparametric test was applied, one of the most common tools for multiple statistical tests when comparing more than two methods. The previous observation is also statistically confirmed as illustrated from the test results (Table IV). According to the test results, the null hypothesis that the means of MAE values of two or more algorithms are the same is rejected (significance level of 0.05), while the algorithms are ordered from the best performer (lowest ranking value) to the worst one (highest ranking value). RF prevails, since it gives statistically better results, followed by the SMOreg (using an RBF kernel) and LR.

TABLE IV. FRIEDMAN ALLIGNED RANKS TEST RESULTS

Algorithm	Rank
RF	8.66667
SMOreg	9.16667
Bagging	14.83333
LR	21.00000
GP	26.83333
5NN	34.50000
M5' Rules	39.83333
M5	41.16667

VI. CONCLUSIONS

Predicting student performance in high education is beneficial to learning [36]. The early identification of learning difficulties triggers proactive actions that could improve the final outcome. Our study utilized eight familiar supervised learning algorithms to train the base regression models, given students' past achievements and selected demographic variables. A number of experiments were conducted, reporting that a fair accuracy is achieved. Future work includes extending the dataset with more semesters and curriculum data, and using the prediction results to recommend courses and specializations to students.

REFERENCES

[1] D. Ifenthaler and C. Widanapathirana, "Development and validation of a learning analytics framework: Two case studies using support vector machines," *Technology, Knowledge and Learning,* vol. 19, no. 1-2, pp. 221-240, 2014.

[2] N. Sclater, A. Peasgood and J. Mullan, "Learning analytics in higher education," *London: Jisc. Accessed February,* vol. 8, p. 2017, 2016.

[3] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education.," *EDUCAUSE review,* vol. 46, no. 5, p. 30, 2011.

[4] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[5] S. Kotsiantis, C. Pierrakeas and P. Pintelas, "Predicting students'performance in distance learning using machine learning techniques," *Applied Artificial Intelligence,* vol. 18, no. 5, pp. 411-426, 2004.

[6] J. M. Luna, C. Castro and C. Romero, "MDM tool: A data mining framework integrated into Moodle," *Computer Applications in Engineering Education,* vol. 25, no. 1, pp. 90-102, 2017.

[7] C. Romero, S. Ventura and E. Garc\'\ia, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education,* vol. 51, no. 1, pp. 368-384, 2008.

[8] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications,* vol. 33, no. 1, pp. 135-146, 2007.

[9] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 3, no. 1, pp. 12-27, 2013.

[10] N. Sclater, Learning analytics explained, Taylor & Francis, 2017.

[11] R. Ferguson, "Learning analytics: drivers, developments and challenges," *International Journal of Technology Enhanced Learning,* vol. 4, no. 5-6, pp. 304-317, 2012.

[12] J. T. Avella, M. Kebritchi, S. G. Nunn and T. Kanai, "Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review.," *Online Learning,* vol. 20, no. 2, pp. 13-29, 2016.

[13] B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges," *British journal of educational technology,* vol. 46, no. 5, pp. 904-920, 2015.

[14] N. Bousbia and I. Belamri, "Which Contribution Does EDM Provide to Computer-Based Learning Environments?," in *Educational data mining*, Springer, 2014, pp. 3-28.

[15] J. P. Campbell, P. B. DeBlois and D. G. Oblinger, "Academic analytics: A new tool for a new era," *EDUCAUSE review,* vol. 42, no. 4, p. 40, 2007.

[16] " 1st International Conference on Learning Analytics and Knowledge, February 27–March 1, 2011, https://tekri.athabascau.ca/analytics/," Banff, Alberta,.

[17] M. Bienkowski, M. Feng, B. Means and others, "Enhancing teaching and learning through educational data mining and

learning analytics: An issue brief," *US Department of Education, Office of Educational Technology,* vol. 1, pp. 1-57, 2012.

[18] L. Johnson, S. Adams Becker, M. Cummins, V. Estrada, A. Freeman and C. Hall, "NMC Horizon Report: 2016 Higher Education Edition. Austin, Texas: The New Media Consortium.," 2016.

[19] R. Ferguson, A. Brasher, D. Clow, A. Cooper, G. Hillaire, J. Mittelmeier, B. Rienties, T. Ullmann and R. Vuorikari, "Research evidence on the use of learning analytics: Implications for education policy," 2016.

[20] S. A. Becker, M. Cummins, A. Davis, A. Freeman, C. G. Hall and V. Ananthanarayanan, "NMC horizon report: 2017 higher education edition," 2017.

[21] S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative.," *Journal of Learning Analytics},* vol. 1, no. 1, pp. 6-47, 2014.

[22] B. P. Koester, G. Grom and T. A. McKay, "Patterns of gendered performance difference in introductory STEM courses," *arXiv preprint arXiv:1608.07565,* 2016.

[23] D. Gašević, S. Dawson, T. Rogers and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *The Internet and Higher Education,* vol. 28, pp. 68-84, 2016.

[24] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis and H. Rangwala, "Predicting student performance using personalized analytics," *Computer,* vol. 49, no. 4, pp. 61-69, 2016.

[25] E. A. Amrieh, T. Hamtini and I. Aljarah, "Mining educational data to predict Student's academic performance using ensemble methods," *International Journal of Database Theory and Application,* vol. 9, no. 8, pp. 119-136, 2016.

[26] J. Xu, K. H. Moon and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 5, pp. 742-753, 2017.

[27] G. Holmes, A. Donkin and I. H. Witten, "Weka: A machine learning workbench," in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, 1994.

[28] R. A. Bottenberg and J. H. Ward, "Applied multiple linear regression," 1963.

[29] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[30] D. W. a. K. D. a. A. M. K. Aha, "Instance-based learning algorithms," *Machine learning,* vol. 6, no. 1, pp. 37--66, 1991.

[31] G. Holmes, M. Hall and E. Prank, "Generating rule sets from model trees," in *Australasian Joint Conference on Artificial Intelligence*, 1999.

[32] Y. Wang and I. H. Witten, "Inducing model trees for continuous classes," in *Proceedings of the Ninth European Conference on Machine Learning*, 1997.

[33] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing,* vol. 14, no. 3, pp. 199-222, 2004.

[34] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Advances in neural information processing systems*, 1996.

[35] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[36] M. Virvou, E. Alepis, G. Tsihrintzis, L. Jain and (eds.), Advances in Learning Analytics, Intelligent Systems Reference Library series, Springer 2018.