Comparative Analysis for Boosting Classifiers in the Context of Higher Education

https://doi.org/10.3991/ijet.v15i10.13663

Eslam Abou Gamie, M. Samir Abou El-Seoud (✉), Mostafa A. Salama
British University in Egypt, Cairo, Egypt
selseoud@yahoo.com

Abstract—Machine learning techniques are applied on higher education data for analyzing the interaction between the students and electronic learning systems. This type of analysis serves in predicting students' scores, in alerting students-at-risk, and in managing the degree of student engagement to educational system. The approaches in this work implements the divide and conquer algorithm on feature set of an educational data set to enhance the analysis and prediction accuracy. It divides the feature set into a number of logical subgroups based on the problem domain. Each subgroup is analyzed separately and the final result is the combination of the results of the analysis of these subgroups. The classifier that shows the best prediction accuracy is dependent on the logical non-statistical nature of the features in each group. Both traditional and boosting classifiers are utilized on each dataset, from which a comparison is conducted to show the best classifiers along with the best dataset. This approach provides the possibility to apply a brute force algorithm in the selection of the best feature subgroups with a low computational complexity. The experimental work shows a high prediction accuracy of the students-at-risk relative to the current research, and provides a list of new important features in the field of electronic learning systems.

Keywords—Learning Analytics, Education Data Mining, AdaBoost, XgBoost, Random Forest, Support Vector Machine, OULAD, Virtual Learning Environment, Learning Management System.

1 Introduction

With the rapid expansion in technology; educational institutions generate tons of statistical and behavioral student's records; such data could be analyzed to reveal useful knowledge in order to facilitate the learning and teaching processes. Students may not perform well in some modules not because they are not diligent enough, but also because the educational settings or modules representation does not fit them. Accordingly educational data mining plays an important role to reveal hidden knowledge from which early interventions could be made to detect at risk students. Education data mining EDM is defined as "a growing discipline which is concerned with the development of methods for exploring the unique types of data that come

from an educational setting, and use this data in order to better understand students and the settings in which they learn in" [1]. EDM main concern is developing models to improve learners experience and to enhance academic performance. Academic progress can be monitored by predictive models; such models use various data mining techniques to analyze students' performance. Data collection, pre-processing, modeling and finally interpretation are the four main phases for any traditional data mining problem. FS algorithms are applied in the pre-processing step of data. Its aim to select the most appropriate set of features prior building the data mining model in order to enhance model accuracy and lower computational complexity [4]. Modeling phase concerns with developing techniques to categorize data based on similar characteristics. Clustering and classifications are the most popular DM techniques (learning methods) that could produce an effective predictive model. Major EDM problems can be categorized into non-standardization in educational settings and the associated generated data, leading to difficulty in selecting the best features and the optimum mining technique that could enhance the accuracy of the predictive model. Studies have shown that educational settings can have a great impact on students' performance and grades [5]. The main objective of this paper is to study and analyze benchmarked student's data, generated from educational settings, in order to propose a predictive model with enhanced accuracy rates compared to others on same data set. Moreover to find the most significant features that may affect students' performance by adding an extra layer of logically grouped features prior applying classification techniques. A various set of classifiers are applied on dataset with focus on traditional classifiers and newly enhanced ensemble techniques, finally a comparison is conducted between classifiers to reveal the best model from accuracy perspective. Facilitating teaching and learning process are the main motivations for this research; by early interventions of at risk students at early stages could aid decision makers to detect drawbacks in students' learning behavior. While data mining DM in other contexts is known for its effectiveness in other sciences like e-commerce, banking, digital marketing and other business industries, yet the applications of DM in the educational context is still limited. Academic progress can be monitored by predictive models; such models use various data mining techniques to analyze students' performance. Major EDM problems can be categorized into non-standardization in educational settings and the associated generated data, leading to difficulty in selecting the best features and the optimum mining technique that could enhance the accuracy of the predictive model. Accordingly the work in this paper tries to discuss the answers for the following questions: what are the most significant features that may affect students' performance? What is the best fit classification model with highest accuracy rates? Will a hybrid classification model produce better accuracy results than a single model? Will a specific combination of dimensions of features utilizing a certain classifier acts better than a full data set of features?

2 Background and Literature Review

Applications and methods for mining students' data can be categorized based on problem domain; accordingly many surveys in the last decade have listed possible applications of EDM. According to intensive survey provided by Behdad, Osmar and Samira they identified 13 categories of applications, forming a new taxonomy tailored specifically to educational domain [25]. Student modeling is a cognitive operation devoted to representing cognitive prospect of students' activities, such as analyzing the student's behavior in order to detect their performance, isolating underlying misconceptions, representing students' goals and plans, identifying prior and acquired knowledge, maintaining describing personality characteristic [26]. Under modeling category is performance prediction, which is the main aim of this paper. Detecting students' engagement in a web-based course contents by utilizing machine learning (ML) techniques to measure the effect of such interactions on student's performance. With engagement level as target variable; the level is classified as high and low. The input features of the model included highest education level, final results, assessment score, and number of clicks on the virtual learning environment [6]. The activities is considered also as a group of important features, it includes the data plus, the forums, the glossary, the resources, the subpages, the homepages and the URL during the first course assessment [10] .Absence of data from previous courses which are usually used in training the model has been tackled by data generated from early course assessments, the approach tries to find the correlation between the first assessment and the final grade; by applying ML techniques to extract students' behavior who submitted their assignments earlier than others, hence apply it to other students [7]. Rules extractions from eLearning systems to detect frequent patterns is not always enough, normal association roles like Apriori algorithm do not take infrequent associations into consideration, despite the fact that relatively infrequent associations could be of significant interest [8], that's why Rare Association Mining Technique could play an important role to detect infrequent student's behaviors. In [9] the author deals with variance of courses types and number of activities generated from eLearning systems; by detecting the relationship between activities and resources in a certain course along with students' final grades. He did so by applying different Multiple Instance learning techniques and results were compared. Although the research is well organized the main focus was on the techniques, and not the data attributes, without mentioning the reason behind choosing only three specific students' online activities. The author in [10] started with a question if it is possible to predict student's success enrolled in a course with a small dataset? And that datasets associated with students are considered small even if with a big number of students. Student attributes considered in this paper for prediction: gender, year of birth, Employment, status, registration, type of study, Exam condition and activities. Several perspectives affect student's behavior over the academic period, the first perspective is partitioning the factors that are affecting the student behavior according to the institutional and family support and degree of the student awareness ,the second perspective studies the students who perform an improvement during their study in the university, another perspective is the addition of the external factorials like the economic status, finally the interaction to

the electronic educational systems [11][12][13]. The current research trends in this area examine the different activities performed by the student on the Electronic learning systems. The work in [14] studies the frequency of online interaction of the students; by measuring the percentage of accessing the virtual classroom and discussion boards. The work in [15] provides an evaluation to the E-learning systems by categorizing the different factors that may affect the student performance. These factors are divided into six dimensions: system quality, service quality, content quality, learner perspective, instructor attitudes, and supportive issues. The purpose of the previous work is to gain the benefit of all factors that affect the student performance and build a machine learning model that enables the decision makers in altering the teaching methodology.

3 Model Implementation

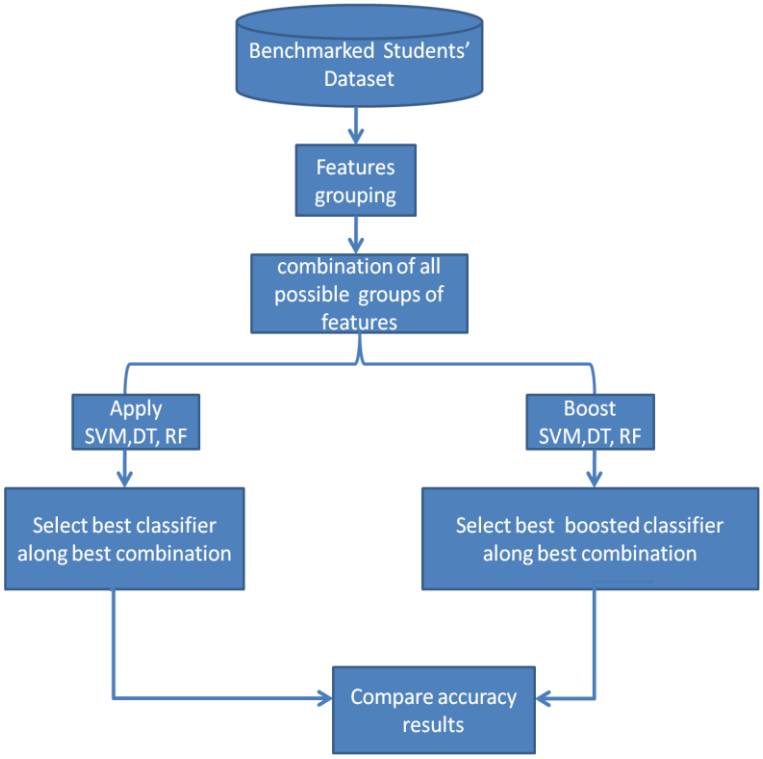

Fig. 1. Ensemble with Boosting model

The main model in this paper is ensemble model with boosting as shown on figure 1. The approach starts by categorizing the OULAD education dataset into four groups. As mentioned these groups are students' demographic data, module registration, assessments, and finally records from the virtual learning environment. A combination

of such groups is then conducted without redundancy, resulting in 10 combinations in addition to the 4 partitions of a singular dimension. Various machine learning techniques are applied on each partition and a combination is used to detect the best fitting classifier. A classifier could perform better on one of the partitions, while the same classifier may act less on another partition in the same data domain. In addition to the non-ensemble classifiers applied in model one, bagging and boosting techniques are utilized on each partition and combinations of features. Both bagging and boosting are forms of ensemble techniques that combine multiple learners to generate a more accurate model. By default, bagging utilizes bootstrap sampling to get the data subgroups for training the base learners. To assemble the outputs of base learners, bagging uses voting analysis for classification. On the other hand, boosting combines multiple weak classifiers to create a single strong classifier. A weak classifier is a learner whose prediction accuracy is slightly better than random guessing. Decision trees and decision stumps are examples for weak classifiers. However, theoretically boosting can be possible with any base classifier that accepts features weighting. In real life practices it seems that mostly used (boosted) base learners are tree-based classifiers. On the other hand, classifiers like SVM and RF are considered as strong classifiers, whose classification accuracy will outperform DT. As such, boosting SVM and boosting Random Forest are introduced, in order to detect the classification accuracy compared the default Boosting DT. The steps below summarize model implementation:

1. Initialize features groups based on data sources.
2. Generate all possible combination of groups with no repetition
3. Apply SVM, DT, NN as non-ensemble learners on each partition/combination of features
4. Apply RF as bagging technique on each partition/combination of features
5. Apply XgBoost and AdaBoost with default DT as the base learner each partition/combination of features.
6. Select best fit classifier along with the best combination of groups from 3, 4 and 5
7. Boost linear and non-linear SVM, boost RF and save results
8. Compare classification accuracy from 6 and 7

To simplify the output of model two, table 1 below summarizes the utilized classification techniques applied across various groups of features grouping, along with their combinations, with classification accuracy results for each:

Table 1. The first four rows represent features partitions, along with their classification accuracy across different classifiers. The rest of rows represent all possible combinations of partitions. The last row represents the full features dataset.

		Classifiers Accuracy Results					
		SVM	***DT***	***NN***	***AdaBoosts***	***XGBoosts***	***RF***
1	Student info - D	0.422	0.409	0.4207701	0.418776	0.4123332	0.4297
2	Registration - R	0.671	0.6699	0.6721966	0.67235	0.6725038	0.6736
3	Assessment - A	0.662	0.6597	0.6624031	0.664729	0.626938	0.661
4	VLE Interaction	0.669	0.6536	0.6684913	0.664044	0.6583989	0.6842
5	DR	0.655	0.6498	0.6021505	0.684793	0.6537634	0.68
6	DA	0.66	0.664	0.4804264	0.671512	0.6277132	0.6773
7	D_VLE	0.68	0.6618	0.6563462	0.671741	0.6713992	0.6861
8	RA	0.805	0.6581	0.7226207	0.806939	0.7615817	0.8197
9	R_VLE	0.789	0.6586	0.767796	0.798297	0.801848	0.8109
10	A_VLE	0.682	0.6575	0.6737864	0.692039	0.7174757	0.7353
11	DRA	0.807	0.6534	0.7414228	0.802675	0.7749564	0.8174
12	DR_VLE	0.794	0.6635	0.7979124	0.798426	0.7956879	0.8198
13	RA_VLE	0.803	0.6555	0.7253836	0.836862	0.851039	0.844
14	DRA_VLE (Full set)	0.797	0.6646	0.7945232	0.835308	0.8428821	0.8499

4 Experimental Work and Results Analysis

The experimental work is applied on a benchmark education data set; Open University Learning analytics database (OULAD). The Open University uses a technological platform that includes MOOC along with Moodle LMS. The dataset includes the various interaction activities of the students to the Virtual Learning Environments (VLE), along with the demographic data and the assessment results of the students. The analysis study applied here focuses on for seven selected courses (modules).

The number of features extracted from the Open University database is 41 features over 30k students. Different classical Machine Learning algorithms are applied on this extracted dataset, including SVM, DT and NN. In addition, AdaBoost and XGBoost classifier algorithms are used as boosting algorithms, and RF algorithm is used as a bagging algorithm. The target variables of the models fall in one of three classes: pass, fail or distinguish as label for scores categories. A comparison among these algorithms shows that the RF algorithm has the best classification/prediction accuracy with 84.99%. Fig 2 demonstrates the utilized ML techniques along their accuracy results.

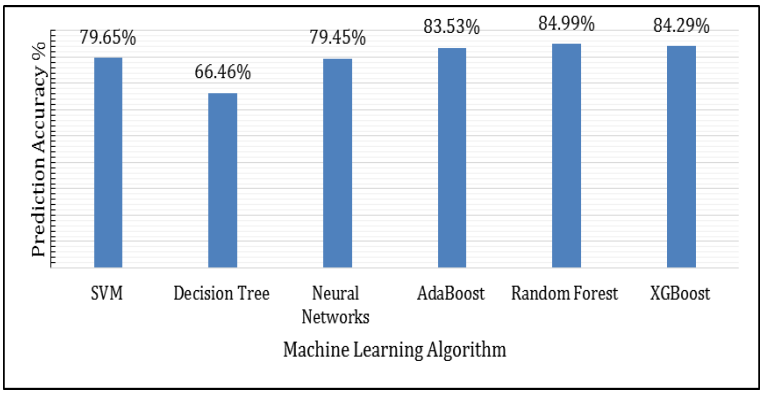

Fig. 2. Classification accuracy of machine learning algorithms applied on full set of 41 features

According to the proposed model, the features within this extracted dataset are categorized into four logical groups: Student demographic Information (D), Assessment (A), VLE interactions (VLE) and Registration (R). Each group contains a different number of features. When applying the different machine learning algorithms on each of these subgroups, the prediction accuracy of the independent datasets shows that the features of student demographic (S) performed the least, with around 40% classification accuracy. On the other hand, features generated from the Virtual Learning Environment and R data sets shows the best prediction accuracy percentage as shown in figure 3. This demonstrates that the student demographic information has the least discrimination power between students of various performance results in the final score of the module. While the information about the interaction of the students to the VLE system and the registration of the student to module has significance in predicting the final performance of this student.

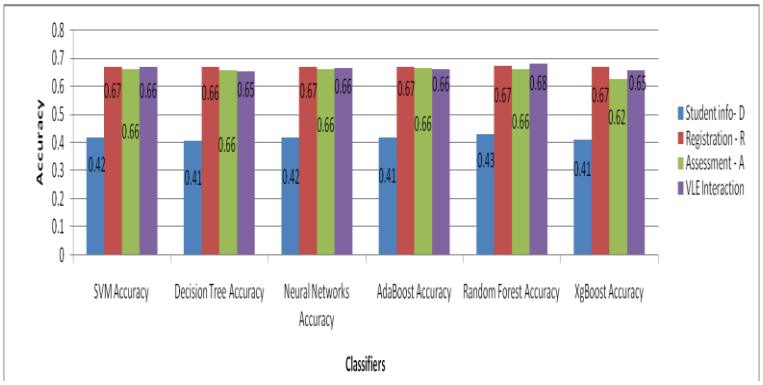

Fig. 3. Classification accuracy of various machine learning algorithms applied on the four subgroup datasets (Student demographic Information (S), Assessment (A), VLE interactions (V), and Registration (R) independently) with lowest accuracy to demographic data represented in blue bar

Accordingly, instead of applying a feature selection and feature reduction techniques, the features in the subgroup of student info-D can be simply excluded from the modelling process. When excluding the subgroup of features that reflects that student demographic information from the full dataset; the prediction accuracy is enhanced in SVM, AdaBoost and XgBoost, which prove that this sub group of features (demographic student info-D) is of least importance in the prediction problem, thus can be neglected. Although the prediction accuracy across other techniques deteriorates by a small percentage, the value of this deterioration can be ignored against the advantage of reducing the computation time by reducing number of features input to the model. Figure 4 shows the difference in classification accuracy between the full data set, and the subset of combining registration, assessment and VLE interactions (RA_VLE), followed by Figure 7 that shows all classification results with respect to accuracy across different combinations of features, with maximum accuracy of almost 85% for both RF on the full dataset, and XGBoost on combinations of registration, assessment and VLE interactions.

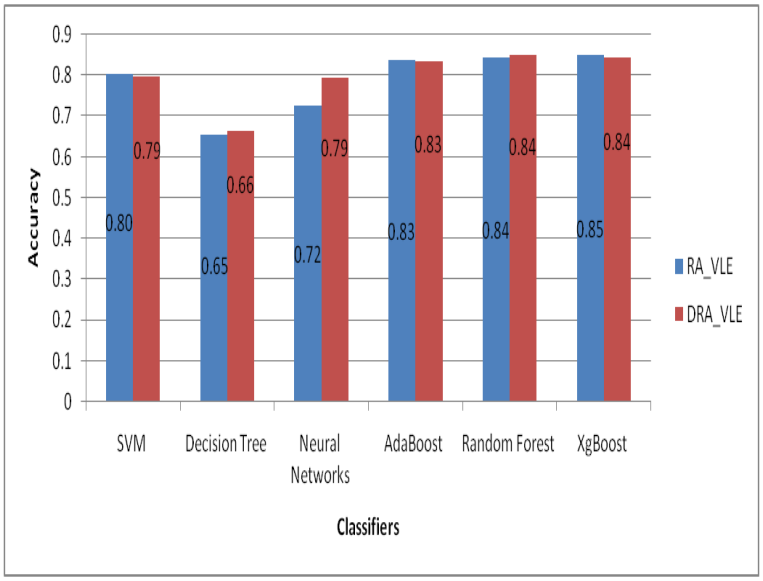

Fig. 4. Classification accuracy of machine learning algorithms applied on the dataset that includes three subgroup features Assessment (A), VLE interactions (V), and Registration (R) in comparison to results when applied on the whole dataset (DRA_VLE).

Furthermore, a brute force analysis is applied to detect the best combination of feature subgroups. The computation complexity of testing all the combinations of the four subgroups is lowered when compared to that of all the combinations of the 41 features; this allows the detection of the combination of feature-subgroups that are important to the prediction process. Figure 5 demonstrates prediction accuracy when different machine learning techniques are applied on all possible combination of features. Results analysis clarifies th t while utilizing support vector machine, the combi-

nation of the subgroups AV (Student Assessment and VLE interactions) provide the same and highest prediction accuracy compared to the whole dataset including all features of the subgroups.

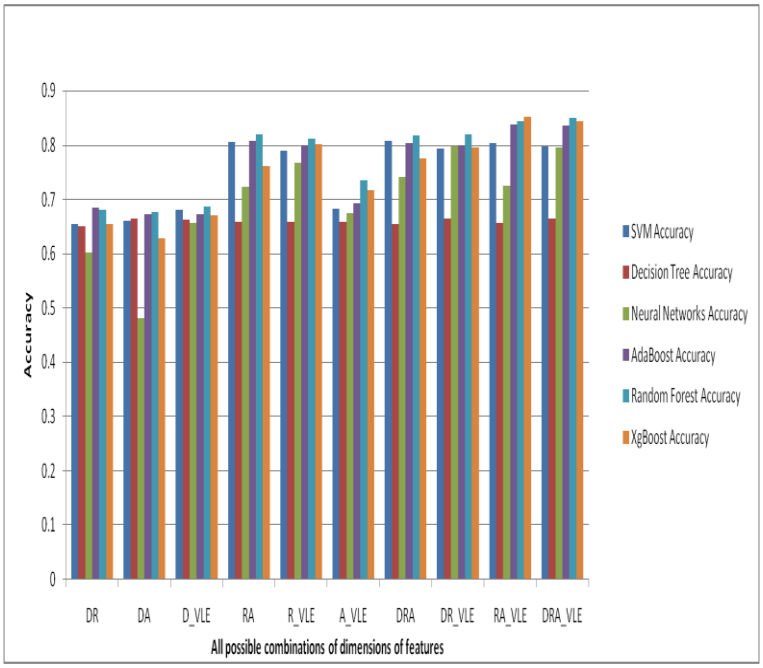

Fig. 5. The classification accuracy of machine learning algorithms is applied on all possible combinations of feature-subgroup.

5 Conclusion

A brute force analysis is applied to detect the best combination of feature subgroups. Results analysis illustrates the extent to which classification results measure up with respect to accuracy across different combinations of features, with maximum accuracy of almost 85% for both Random Forest on the full dataset including demographic data, and XGBoost on combinations of registration, assessment and VLE interactions. This leads to the conclusion that demographic data does not affect the accuracy of final results when excluded from the full dataset, and thus can be neglected in favour of reducing computational time and complexity. The final step in model two introduced an adjustment for boosting techniques; with a trial to increase the classification accuracy by more than 85%.

6 References

[1] Romero, c., & ventura, s. (2010). Educational data mining: a review of the state of theart.ieee transactions on systems, man, and cybernetics, part c (applications and reviews), 40(6), 601-618. https://doi.org/10.1109/tsmcc.2010.2053532

[2] Ranjan, J., & Malik, K. (2007). Effective educational process: a data-mining approach. VINE, 37(4), 502-515. https://doi.org/10.1108/03055720710838551

[3] Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. International Journal of Modern Education and Computer Science (IJMECS), 8(11), 36-42. https://doi.org/10.5815/ijmecs.2016.11.05

[4] Ramaswami, M., &Bhaskaran, R. (2009). A Study on Feature Selection Techniques in Educational Data Mining. Journal of Computing, 1(1), 7-11. doi: arXiv:0912.3924

[5] Zhang, M., Zhu, J., Zou, Y., Yan, H., Hao, D., & Liu, C. (2015). Educational Evaluation in the PKU SPOC Course "Data Structures and Algorithms". In Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15), (pp. 237-240). New York, NY, USA:ACM. https://doi.org/10.1145/2724660.2728666

[6] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computational Intelligence and Neuroscience, 2018(6347186), 21 pages. https://doi.org/10.1155/2018/6347186

[7] Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data. In proceedings ofthe Seventh International Learning Analytics & Knowledge Conference (LAK '17) (pp. 6-15). New York, NY, USA: ACM. https://doi.org/10.1145/3027385.3027449

[8] Romero, C., Romero, J. R., Luna, J. M., & Ventura, S. (2010). Mining Rare Association Rules from e-Learning Data. In R. Baker, A. Merceron, & P. Pavlik Jr (Eds.), proceedings of the 3rd International Conference on Educational Data Mining (pp. 171-180). Pittsburgh: International Educational Data Mining Society. https://doi.org/10.1201/b10274

[9] Zafra, A., & Ventura, S. (2009). Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming. In proceedings of the 2nd International Conference on Educational Data Mining (EDM) (pp. 307-314). Cordoba, Spain: International Working Group on Educational Data Mining. https://doi.org/10.1201/b10274-16

[10] Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications, 41(14), 6400-6407. https://doi.org/10.1016/j.eswa.2014.04.024

[11] Hijazi, S. T., &Naqvi, S. M. (2006). Factors affecting students'performance: A case of private colleges. Bangladesh E-Journal of Sociology, 3(1), 65-99.

[12] Farooq, M. S., Chaudhry, A. H., Shafiq, M., &Berhanu, G. (2011). Factors affecting students' quality of academic performance: a case of secondary school level. Journal of Quality and Technology Management, 7(2), 1-14.

[13] Davies, J., & Graff, M. (2005). Performance in e-learning: online participation and student grades. British Journal of Educational Technology, 36(4), 657-663. https://doi.org/10.1111/j.1467-8535.2005.00542.x

[14] Whalen, S., & Pandey, G. (2013). A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. In proceedings of the 13th IEEE International Conference on Data Mining (pp. 807-816). IEEE. https://doi.org/10.1109/icdm.2013.21

[15] Topaloglu, M., &Ekmekci, S. (2017). Gender detection and identifying one's handwriting with handwriting analysis. Expert Systems with Applications, 79, 236-243. https://doi.org/10.1016/j.eswa.2017.03.001

[16] Rao, Z., Zeng, C., Wu, M., Wang, Z., Zhao, N., and Wan, M. L. X. (2018). Research on a handwritten character recognition algorithm based on an extended nonlinear kernel residual network. KSII Transactions on Internet and Information Systems, 12(1), 413-435. https://doi.org/10.3837/tiis.2018.01.020

[17] Bresfelean, V. P., Bresfelean, M., Ghisoiu, N., & Comes, C.-A. (2008). Determining students' academic failure profile founded on data mining methods. In proceedings of ITI 2008 - 30th International Conference on Information Technology Interfaces (pp. 317-322). Dubrovnik, Croatia: IEEE. https://doi.org/10.1109/iti.2008.4588429

7 Authors

Eslam Abou Gamie, **M. Samir Abou El-Seoud** and **Mostafa A. Salama** work at the British University in Egypt, Cairo, Egypt.

Article submitted 2020-02-09. Resubmitted 2020-03-07. Final acceptance 2020-03-09. Final version published as submitted by the authors.