A Conceptual Framework to Aid Attribute Selection in Machine Learning Student Performance Prediction Models

https://doi.org/10.3991/ijim.v15i15.20019

Ijaz Khan[1(✉)], Abdul Rahim Ahmad[1], Nafaa Jabeur[2], Mohammed Najah Mahdi[1]
[1]Universiti Tenaga Nasional, Kajang, Malaysia
[2]German University of Technology, Muscat, Oman
ijaz@buc.edu.om

Abstract—One of the key applications of Learning Analytics is offering an opportunity to the institutions to track the students' academic activities and provide them with real-time adaptive consultations regarding the students' academic progression. However, numerous barriers exist while developing and implementing such kind of learning analytics applications. Machine learning algorithms emerge as useful tools to endorse learning analytics by building models capable of forecasting the final outcome of students based on their available attributes. Machine learning algorithm's performance demotes with using the entire attributes and thus a vigilant selection of predicting attributes boosts the performance of the produced model. Though, several constructive techniques facilitate to identify the subset of significant attributes, however, the challenging task is to evaluate if the prediction attributes are meaningful, explicit, and controllable by the students. This paper reviews the existing literature to come up with an exhausting list of attributes used in developing student performance prediction models. We propose a conceptual framework which identifies the nature of attributes and classify them as either latent or dynamic. The latent attributes may appear significant, but the student is not able to control these attributes, on the other hand, the student has command to restrain the dynamic attributes. The framework presents an opportunity to the researchers to pick constructive attributes for model development. We apply artificial neural network, a supervised learner, over a dataset to compare the performance of prediction models with distinct classes of attributes. It confirms the significance of dynamic attributes for student performance prediction models.

Keywords—learning analytics, student performance prediction, academic analytics, machine learning

1 Introduction

Learning analytics (LA) examine student's academic activities by making use of the data collected from different sources and search for correlations linking student's

activities and learning outcomes. LA put forward a favorable approach to understand the learning atmosphere and support learners during the instructive process [1]. The raw data is collected from learners, normally through learning management systems, browsing and interaction behavior and then the stakeholders of the process act upon the data to arrange constructive procedures. However, the stakeholders may be expanded or substituted by other groups such as researchers, service providers, or governmental agencies [2]. LA helps to improve teaching methods, learning activities, and extract concealed knowledge about learners [3].

Monitoring student performance is one of the core applications of Learning Analytics. It monitors student's academic performance, while the course is still in progress, and intervenes when the students are leading towards a disappointing academic ending [4]. LA make use of technologies, for instance, educational data mining, machine learning, classical statistical analysis techniques, and social network analysis to accomplish the projected purposes [5]. Machine Learning classifiers are among the tools appearing productive in monitoring students' academic performance.

Some machine learning algorithms, particularly supervised, have been broadly used to build prediction models. The supervised algorithms construct a classification model with the training dataset. The model encompasses the taxonomy conventions revealed from the training dataset. The testing phase executes the model with a subset of the training dataset to evaluate the performance of the model. The training dataset consists of instances (records) and each instance has a number of attributes. For instance, gender, age, and grades are some of the students' attributes. Several authors preferred mining the attributes to come up with a set of significant attributes, although a number of authors preferred to make use of the entire set of attributes [6, 7]. The performance of the prediction model relies on the student attributes available in the dataset. Sudani et al. [8] categorized the prediction attributes as academic, psychological, demographic, and others. Papadogiannis et al. [9] concluded grades, demographics, and academic data besides grades are widely used prediction attributes. The review by Shahiri et al. [10] shows Cumulative Grade Points Average (CGPA) and internal assessments as the most widely used prediction attributes.

One of the essential phases of prediction model development is the identification of suitable prediction attributes [11]. The performance of machine learning classifiers might downgrade if the whole set of attributes is used [12]. It points towards the fact that a careful selection of predicting attributes tends to improve the performance of the prediction model [13, 14]. Several features selection algorithms facilitate the researcher in this phase. Though, a chosen attribute may appear significant for the classification model but may have rigid nature and thus the student will fail to boost its performance. The evident concern demands the nature of predicting attributes must be flexible so the student can get an opportunity to get inspiration and make improvements in each of the predicting attributes. For instance, gender may emerge as an appropriate predicting indicator however its fixed nature prevents students from any variation. The selected attributes must appear as proxies for learning [15]. The moment a student is forecasted with a probability of ending up with an unsatisfactory outcome, then, the student must have a command to rework and improve its performance in the prediction indicators.

The major contribution of this paper is to review the existing students' performance prediction models and identify the prediction attributes preferred by the authors under distinct educational settings. This paper proposes a conceptual framework to categorize and subcategorize the prediction attributes based on whether the attributes are

meaningful, precise and controllable by the student. The paper is organized as; section 2 provides a literature review of learning analytics, machine learning and student performance prediction models. Section 3 describes various classes and levels of prediction attributes and demonstrates the proposed framework. Section 4 provides experimental evaluation to validate the conceptual framework and section 5 concludes the paper.

2 Literature review

The institutions are craving to collect more and more data than the past to maximize their strategic planning [16]. Despite possessing a huge amount of data, the institutes require standard means of organizing the data and utilizing it for appropriate decision making. Various computer technologies offer numerous opportunities to transform complex educational material into a form easy to understand and remember [17]. The educational institutions adopt novel technologies such as Interactive Learning Environments (ILE), learning management systems (LMS), intelligent tutoring systems (ITS), and online learning platforms which results in an access to a huge quantity of data about the students and the underlying learning environment [18, 19]. The intention of transforming large amount of data into constructive information, leads towards the significance of learning analytics [20]. Learning Analytics acquires the data relevant to students and instructors at both individual learner or course level and applies analytic techniques to improve students' learning outcomes through better instructional, curricular well as supporting resources, interventions, and learning culture empowerment [21].

Monitoring individual student performance, in a course, is one of the key areas of LA applications [22] and it appeared, as one of the eight categories of instructional applications, in a white paper entitled "Analytics for Achievement" published by IBM [22]. Monitoring represents the observation and inspection of the progress or quality of something over some time [23]. Student monitoring appears as an incredibly imperative factor in higher educational institutions since the institutions have been systematically monitored by governmental bureaus and accreditation agencies [24, 25]. Therefore, to remain competent in pursuing an admirable reputation in pedagogical society, the institutes ought to implement novel procedures to track students' academic progress. This monitoring emerges as a supportive tool for an instructor to identify the students with unsatisfactory academic progress. The instructor can accordingly provide instantaneous interventions to help students understand their endangered circumstances and rework to improve. In order to achieve such goals, the institutions need to put into action novel methodologies that foretell the outcome of students based on their ongoing academic activities. Course Signals implemented at Purdue University [26, 27] is an eminent application that extracts data from several sources and the statistical techniques are applied to forecast students at the risk of failing the courses. The instructors then intervene and organize counseling sessions with the weak students.

2.1 Machine learning

There are various tools used to develop models which are capable of predicting student's outcome. The models are used as effective means of identifying the students having a high probability to end up with disappointing final results. Machine Learning [28]

is considered a useful tool to develop prediction models. Machine Learning (a branch of Artificial Intelligence) refers to learning from the previous dataset to enhance future performance automatically without any explicit programming [29]. The two major categories of machine learning algorithms are supervised and unsupervised. Supervised learning consists of input and output and the aim is to estimate the mapping function sufficiently well so it can forecast the output for unseen data. Supervised learning can be either classification or regression. In classification, the output is a category, while in regression it is a real value. Naive Bayes, Decision Trees, Support Vector Machines (SVM), Linear Regression, are some of the popular supervised learning algorithms. Unsupervised learning has only input data and it has to model the underlying structure in the data to learn more about the data. Unsupervised learning can be clustering and association. The aim of clustering is to find out the inherent groupings in the data, while in association rule the fundamental rules that represent the large segments of data are discovered. K-means and Apriori are well-liked clustering and association algorithms respectively.

Particularly supervised algorithms have been implemented to develop models that accurately predict the characteristics of students which provoke their behavior and performance [30]. Figure 1 illustrates the working of supervised learning. It is mainly a two step process of model development and validation. In the first step, the supervised classifier constructs a classification model with the input dataset (called a training dataset). The training dataset consists of instances (records) and each instance has a number of attributes where an attribute denotes a single characteristic of the student which may influence its academic behavior. These attributes constitute the set of independent variables that forecast students' outcome by labeling them into the most probable category. The produced classification model constitutes classification rules as it is discovered from the training dataset. The subsequent step, called testing, runs the model with a subset of data from the training dataset to determine the effectiveness of the model. The model gains knowledge from the prearranged training dataset, henceforth, ready to classify the instances from the unseen dataset (validation dataset) [31]. The validation data consists of instances with unknown classes. This validation data is provided to the classification model as an input. The model evaluates each instance and tags it with an appropriate class label.

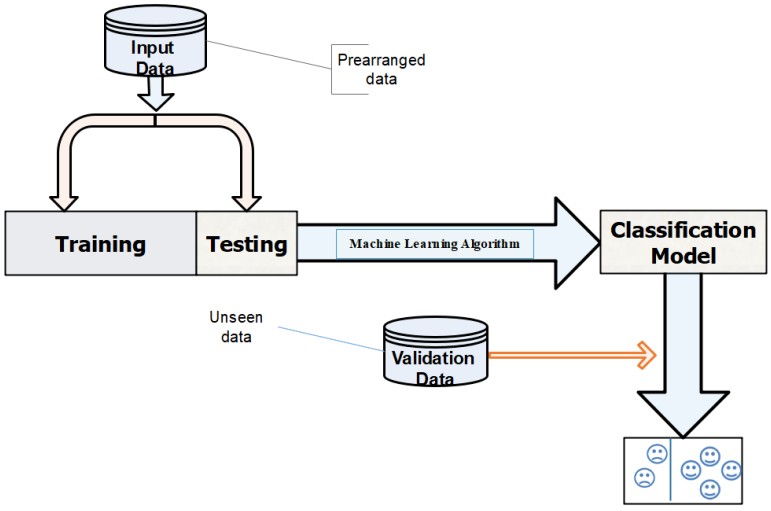

Fig. 1. An illustration of the supervised classification process

Numerous prediction models make use of students academic, demographic, and social attributes to develop prediction models which can forecast students' outcome based on several prediction features at a specific stage of the semester [32]. Asif et al. [33] applied supervised classifiers to forecast the final result of university students at an early phase. Al-Sudani et al. [8] implemented Artificial Neural Network for the discovery of low-performing students at an early stage of the semester so that the university can propose suitable interference procedures to decrease the attainment gap. Ghosh et al. [34] made use of lazy algorithms to identify the student vulnerable of failing mathematics courses and forward the information to the corresponding instructors. Kausar et al. [35] made use of ensemble techniques to examine the relationship between students' semester course and final results. Hussain et al. [36] concludes decision tree as a robust solution in successfully recognizing the students who truly exhibit low-engagement during assessment activities. Similarly, Jishan et al. [37] used Naïve Bayes, Decision Tree, and Artificial Neural Networks to forecast students' final result before the final exam. There are a number of models which are based on decision tree [38, 39], lazy classifiers [40], Artificial Neural Networks [6, 41, 42], Naïve Bayes [37, 43] and Support Vector Machine [44, 45].

3 Prediction attributes conceptual framework

In educational institutions, success is measured by the students' academic performance or how well the students achieved the standards set out by the instructor and the institution [46, 47]. In the dataset, an instance possesses several attributes. An attribute or feature demonstrates the unique characteristic of a student. Not all the attributes emerge vital in designing prediction models. The machine learning algorithm's performance relegates if all the attributes are used. Therefore, a selection of predicting attributes is required to enhance the performance of the prediction model. The selected attributes must exhibit the dominance to measure student learning rather than boosting student retention [48]. Therefore, it is an obligatory task to appropriately deal with the raw data and recognize the attributes reliable for decision making [49]. This turns towards the need for appropriate and accurate indicators that are meaningful for predicting student's performance. We provide a review of the attributes utilized for predicting students' final outcome through machine learning classifiers.

In the light of this review, we believe that based on the significance, the attributes can be either latent or dynamic. Latent attributes stay along with the student but a student does not have a control to modify these attributes and improve academic status. On the other hand, dynamic attributes not only compute student current status, but the student has command and tends to modify these attributes. As an example, the age is an attribute where a student has no control to alter and thus is a latent attribute. On the other hand, grades in assessment tool (such as assignment) are dynamic attributes as the student can rework to improve their attribute and enhance academic standings. Further, there exist several levels of both latent and dynamic attributes. Here we explain, the levels of attributes and justify whether it is a latent or dynamic.

3.1 Latent attributes

3.1.1. Presage. This set of attributes is associated with the past academic record of the student. Several authors used attributes that highlight the type of previous institution

or school, the medium of instruction in the school, and the type of program of study in the school. A set of attributes stores the academic results of the students in the previous school or institution. Student's grades at various milestones in school life such as grade-10 and grade-12 are considered essentials by several authors [50–52]. These attributes may appear handy in the model developed for forecasting students' dropout at the early stages of the higher institution and for admission eligibility. Several authors [38, 53, 54] and [55] made use of types and location of the school. Hamsa et al. [56, 57] considered the student's admission score and gap years [41, 58] as an essential attribute in the prediction models. However, due to the rigid nature, the student is never able to control these attributes.

3.1.2. Demographic. Statistically, demographic attributes are the quantifiable attributes of the population [59]. Student gender, age, disability, ethnicity, and place of birth are some of the widely addressed demographic attributes. Student age (year of birth) and gender are the broadly used demographic attributes in prediction models in pedagogical environment. Several studies [8, 60, 61] made use of nationality/place of birth, ethnicity, and disability among the prime prediction attributes.

3.1.3. Academic non-reactive. The academic non-reactive attributes may perhaps inspire a student to improve, but it is hard to grasp them and modify them. For instance, student major of study, year of study, type of degree, and scholarship go along with the student academic period, but a change in such attributes appear very rare. Several authors [6, 62] made use of student ID, course ID, and section ID in their models. Several models [36, 54, 63] considers scholarship as an essential attribute in students' performance prediction.

3.1.4. Social behavior. These attributes interrelate with the students' academic and social life. This can be subdivided into family background and social behavior. The social behavior of the student may include attributes such as the employment status of the student, the time spent with friends and family, social relationships, and interest in extracurricular activities. Several others attributes such as student health issues, commuting, and stress management capability may affect a student's academic performance.

3.1.5. Family background. The family background includes attributes such as parents' education [53, 63, 64] and occupation [63, 65], family size [38], parents influence, living area and other related attributes.

3.2 Dynamic attributes

3.2.1. Academic reactive. These attributes are related to the students' academic activities and thus can appear useful to compute the current academic status of the student. The academic reactive attributes stick to the student throughout studies and calculate the academic standings of the student. The most broadly applied attributes consist of; Cumulative Grade Point Average (CGPA), student attendance in the course, grades in assessment tools (such as quizzes, assignments, and lab work), and grades in the midterm exam [6, 62, 66]. Indeed, the students have command to control these attributes to recover and improve their academic standings. Several other attributes we suggest include; students' GPA in the previous semester, number of subjects registered in the current semester, and grades in the prerequisite subject.

3.2.2. Psychometric attributes. The psychometric attributes underscore the students' behavior, interest, and barriers towards their studies. Several authors [58, 67] used psychometric attributes in designing performance prediction models. These attributes may

be reactive such as private tuition, internet access, motivation for higher education, or maybe non-reactive, for instance, homesickness, self-motivation, extra abilities in the student. Thoroughly organizing these attributes might improve the reactive attributes and thus enhance student's academic performance. Table 1 provides a summary of the various prediction models and the number of attributes they used from each of the categories.

Table 1. Prediction models and the number of attributes used from each category

Reference	Presage	Demographic	Academic Non-reactive	Social Behavior	Family Background	Academic Reactive	Psychometric
[6] [60]	0	3	5	2	0	1	4
[34]	0	0	2	0	0	0	0
[65] [7]	1	3	1	7	6	4	5
[68]	1	5	0	1	0	2	0
[54]	2	2	1	3	0	1	2
[62]	1	0	7	0	0	4	0
[69]	0	0	1	0	0	1	1
[70]	0	0	1	3	5	0	1
[56]	1	0	0	0	0	3	0
[38]	1	1	2	5	2	3	1
[64]	0	0	1	0	2	1	3
[71]	0	1	6	1	1	0	0
[37]	0	0	0	0	0	5	0
[72]	0	0	0	0	0	6	0
[41]	5	2	1	1	1	0	0
[73]	3	0	0	0	1	0	0
[63]	0	1	1	1	3	2	0
[74]	0	0	0	0	0	8	0
[66]	0	0	1	0	0	3	0
[51]	2	1	0	0	1	0	1
[75]	0	2	5	1	0	3	0
[57]	1	1	4	0	1	1	0
[76]	4	0	1	1	3	0	1
[52]	2	2	2	0	0	8	4
[77]	0	0	0	0	0	1	0
[58]	5	1	2	0	5	4	8
[61]	2	3	0	1	0	1	0
[67]	1	1	0	0	0	2	4
[55]	3	0	4	1	6	2	1
[78]	0	0	2	0	0	1	0

Table 2 provides a brief summary to illustrate the class of attribute, the frequency they appear in models and broadly used attributes in each category.

Table 2. The category and the potential attributes included

Category	Attributes
Presage	**Grade history:** (high school grades, grades in specific subjects at school, admission grades) **Schooling:** (location of school, type of school, prestige of school) **Others:** (gap years)
Demographic	Gender, age, nationality, marital status, disability
Academic Non-Reactive	**Student:** (ID, major, semester) **Period:** (section, academic year) **Lecturer:** (qualification, gender) **Others:** (past failures, study type, dorm)
Social Behavior	Social relationships, free time spending, extra-curricular activities, health issues, commuting, stress management capability
Family Background	**Financial support:** (parents occupation , family income), Family literacy level (Parents education), family setup (family size, siblings) family support, residential area
Academic Reactive	CGPA, GPA in previous semester, attendance, marks (quizzes, internal exams, assignments, lab work)
Psychometric	Available resources (internet access, library visits), score for additional activities, time management (study hours), self-motivation, homesickness

3.3 Conceptual framework

We propose a conceptual framework to visualize the sets of prediction attributes and their impact in the prediction models. The framework, in Figure 2, illustrates the two major classes of prediction attributes, which are further subdivided into classes. Some of the most widely used attributes from each category are listed. The conceptual framework proposed does not attempt to explain or describe every possible attribute and relationship among attributes in a machine learning prediction model. It rather provides a set of concepts that can be used to think about developing prediction models. It can be used as a heuristic tool to examine relationships among concepts, prediction models, develop additional lines of research, and raise new questions.

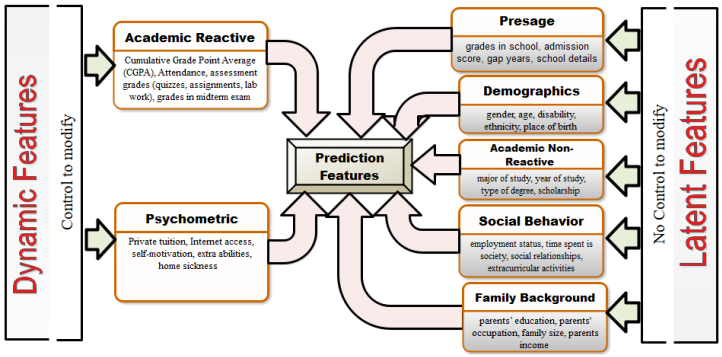

Fig. 2. The proposed conceptual framework to classify the available student's attributes

Attribute selection is among the key stages of developing machine learning prediction models. Besides a number of techniques exist to rank the attributes in a dataset however, our conceptual framework emphases over the nature of attributes as well. Different levels of attributes tend to appear beneficial in distinct situations. The dynamic attributes requires additional weight whilst designing prediction models to forecast the final outcome of the students. This prediction can be more beneficial if the prediction model is accompanied with adapted recommendation module. The recommendation model informs the students of their current academic standings. The students can then have a chance to work hard and improve their academic position. This is more meaningful if the model is designed based on dynamic attributes.

However, the latent attributes emerge constructive in the models intending the classification of students and with no recommendation is envisioned. For instance, forecasting the dropout rate [79, 80] at an institute may consider different levels of latent attributes. Similarly, several models, for instance [81], finds latent attributes meaningful in models for judging the admission eligibility of students in the higher education institute. The latent attributes can support administration in decision making after exploring useful patterns [53].

4 Experimental evaluation

In this section we apply supervised learner, Artificial Neural Networks [82], to observe the delineation between latent and dynamic attributes. The dataset consists of the student academic records for a course taught at Al-Buraimi University College (BUC), Sultanate of Oman. There are total of 151 instances in the training dataset with 12 attributes and one prediction class. The experiments are performed in Waikato Environment for Knowledge Analysis (WEKA) with 10-fold cross-validation.

Initially, we produced a model with only latent attributes in the training dataset. As Figure 3 shows, the model is able to achieve an accuracy of nearly 52%. The following experiment eradicated the latent attributes and the model is built with merely the entire set of dynamic attributes. An increase in the accuracy evidences the significance of dynamic attributes for better prediction performance.

In order to further validate, the attribute selection is performed with Correlation Attribute Evaluator Filter with Ranker search to reduce the number of attributes. Correlation-based Feature Selector (CFS) measures the Pearson's correlation between each attribute and the prediction class and the unrelated features turn up with low correlation value. Table 3 provides the list of attributes, their description, class and the correlation values. The final model produced with the attribute selection achieves accuracy slightly better.

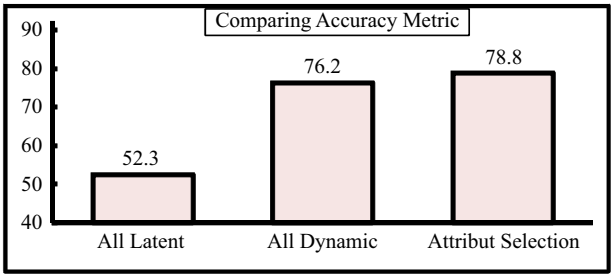

Fig. 3. Comparing the accuracy metric of the models

Table 3. Description of attributes and their details

Attribute	Description	Class	Pearson' Value
CGPA	Cumulative Grade Points Average of student	Dynamic	0.5438
Prev_Sem_GPA	GPA in last semester	Dynamic	0.5324
Test_1	Grades in Midterm Exam	Dynamic	0.4481
Cont_Assess	Grades in assignments, quizzes etc	Dynamic	0.3768
Register_Courses	Number of courses registered in current semester	Dynamic	0.1324
Sponsorship	Whether sponsored for study by government	Latent	0.2277
Dorms	Whether resides in hostel	Latent	0.1347
Nationality	Nationality	Latent	0.1198
Gender	Gender of the student	Latent	0.105
Session	Section of subject	Latent	0.0895
Major	Major of the student	Latent	0.0646
Year	Year of study	Latent	0.0237

These experiments confirm the significance of dynamic attributes over latent attributes for student prediction modeling. It is observed that most of the latent attributes have correlation values less than 0.15. On the other hand, the dynamic attributes possess a higher value. Non-reactive attributes have extremely low correlation comparing to other latent attributes. Dynamic attributes have highest correlation values, which confirm their significance in the prediction models.

5 Conclusion

Machine learning algorithms are constructive tools to support Learning Analytics by building prediction models capable to forecast the final outcome of students based on their key attributes. The dataset usually suffers from high dimensionality and not all the attributes play vital role in the prediction process. A cautious selection of predicting attributes can boost the performance of the produced model. However, it is necessary to consider the nature of the selected attributes, especially, if the prediction model is accompanied with recommendation practices.

This paper lists the attributes used in student performance prediction models and proposes a conceptual framework which demonstrates the attributes as either latent or dynamic. The conceptual framework provides a set of concepts for researchers while developing prediction models. Latent attributes may emerge essential prediction indicators, but the students are unable to modify and improve. On the other hand, the dynamic attributes are well in the students' control. An experimental evaluation confirms the importance of dynamic attribute in the student performance prediction modeling. We apply artificial neural network over a dataset to produce models with all latent attributes, with all dynamic attributes and in the third experiment Correlation-based Feature Selector algorithm was chosen to select the attributes with highest Pearson correlation value. The experiments illustrate the significance of the dynamic attributes.

6 References

[1] Schumacher C. and Ifenthaler D. (2018). Features students really expect from learning analytics. *Computers in Human Behavior*. 78: p. 397–407, 2018. https://doi.org/10.1016/j.chb.2017.06.030

[2] Greller W. and Drachsler H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*. 15(3): p. 42–57, 2012.

[3] 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): p. 601–618, 2010. https://doi.org/10.1109/TSMCC.2010.2053532

[4] Wong B.T.M. (2017). Learning analytics in higher education: an analysis of case studies. *Asian Association of Open Universities Journal*, 2017. https://doi.org/10.1108/AAOUJ-01-2017-0009

[5] Shum S.B. and Ferguson R. (2012). Social learning analytics. *Journal of Educational Technology & Society*. 15(3): p. 3–26, 2012.

[6] Mondal A. and Mukherjee J. (2018). An Approach to predict a student's academic performance using Recurrent Neural Network (RNN). *Int. J. Comput. Appl*. 181(6): p. 1–5, 2018. https://doi.org/10.5120/ijca2018917352

[7] Sekeroglu B.; Dimililer K. and Tuncal K. (2019). Student performance prediction and classification using machine learning algorithms. *in Proceedings of the 2019 8th International Conference on Educational and Information Technology.* 2019. https://doi.org/10.1145/3318396.3318419

[8] Al-Sudani S. and Palaniappan R. (2019). Predicting students' final degree classification using an extended profile. *Education and Information Technologies*. 24(4): p. 2357–2369, 2019. https://doi.org/10.1007/s10639-019-09873-8

[9] Papadogiannis I.; Poulopoulos V. and Wallace M. (2020). A Critical Review of Data Mining for Education: What has been done, what has been learnt and what remains to be seen. *International Journal of Educational Research Review*. 5(4): p. 353–372, https://doi.org/10.24331/ijere.755047

[10] Shahiri A.M. and Husain W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*. 72: p. 414–422, 2015. https://doi.org/10.1016/j.procs.2015.12.157

[11] Xue B.; Zhang M.; Browne W.N. and Yao X. (2015). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*. 20(4): p. 606–626, 2015. https://doi.org/10.1109/TEVC.2015.2504420

[12] Cai J.; Luo J.; Wang S. and Yang S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*. 300: p. 70–79, 2018. https://doi.org/10.1016/j.neucom.2017.11.077

[13] Khan I.; Al Sadiri A.; Ahmad A.R. and Jabeur N. (2019). Tracking student performance in introductory programming by means of machine learning. *in 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*. IEEE 2019. https://doi.org/10.1109/ICBDSC.2019.8645608

[14] Miao J. and Niu L. (2016). A survey on feature selection. *Procedia Computer Science*. 91: p. 919–926, 2016. https://doi.org/10.1016/j.procs.2016.07.111

[15] Watters A. (2012). Learning Analytics: Lots of Education Data… Now What. *Hack Education*, 2012.

[16] Leitner P.; Khalil M. and Ebner M., *Learning analytics in higher education—a literature review*, in *Learning analytics: Fundaments, applications, and trends*. 2017, Springer. p. 1–23. https://doi.org/10.1007/978-3-319-52977-6_1

[17] Romanenko V.; Tropin Y.; Boychenko N. and Goloha V. (2019). Monitoring student performance using computer technology. *Slobozhanskyi herald of science and sport*. 7(2 (70)): p. 36–39, 2019. https://doi.org/10.15391/snsv.2019-2.013

[18] Gašević D.; Dawson S. and Siemens G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*. 59(1): p. 64–71, 2015. https://doi.org/10.1007/s11528-014-0822-x

[19] Rabiman R.; Nurtanto M. and Kholifah N. (2020). Design and Development E-Learning System by Learning Management System (LMS) in Vocational Education. *Online Submission*. 9(1): p. 1059–1063, 2020.

[20] Knight S. and Shum S.B. (2017). Theory and learning analytics. *Handbook of learning analytics*: p. 17–22, 2017.

[21] Huda M.; Sabani N.; Shahrill M.; Jasmi K.A.; Basiron B., and Mustari M.I., *Empowering learning culture as student identity construction in higher education*, in *Student Culture and Identity in Higher Education*. 2017, IGI Global. p. 160–179. https://doi.org/10.4018/978-1-5225-2551-6.ch010

[22] Picciano A.G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of asynchronous learning networks*. 16(3): p. 9–20, 2012. https://doi.org/10.24059/olj.v16i3.267

[23] Finata R. and Andrawina L. (2019). A Systematic Literature Review: Framework Design of Student Performance Monitoring System in Higher Education. *in IOP Conference Series: Materials Science and Engineering*. IOP Publishing 2019. https://doi.org/10.1088/1757-899X/598/1/012024

[24] Marks A.; Maytha A.-A. and Rietsema K. (2016). Learning systems' learning analytics. *in 2016 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE 2016. https://doi.org/10.1109/PICMET.2016.7806510

[25] Wong J.; Baars M.; de Koning B.B.; van der Zee T.; Davis D.; Khalil M.; Houben G.-J., and Paas F., *Educational theories and learning analytics: From data to knowledge*, in *Utilizing learning analytics to support study success*. 2019, Springer. p. 3–25. https://doi.org/10.1007/978-3-319-64792-0_1

[26] Arnold K.E. (2010). Signals: Applying academic analytics. *Educause Quarterly*. 33(1): p. n1, 2010.

[27] Arnold K.E. and Pistilli M.D. (2012). Course signals at Purdue: Using learning analytics to increase student success. *in Proceedings of the 2nd international conference on learning analytics and knowledge*. ACM 2012. https://doi.org/10.1145/2330601.2330666

[28] Marsland S. (2014). *Machine learning: an algorithmic perspective*. 2014: Chapman and Hall/CRC. https://doi.org/10.1201/b17476

[29] Alghamdi M.I. (2020). Survey on Applications of Deep Learning and Machine Learning Techniques for Cyber Security. *International Journal of Interactive Mobile Technologies*. 14(16): 2020. https://doi.org/10.3991/ijim.v14i16.16953

[30] Livieris I.E.; Drakopoulou K.; Tampakas V.T.; Mikropoulos T.A., and Pintelas P. (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of educational computing research*. 57(2): p. 448–470, 2019. https://doi.org/10.1177/0735633117752614

[31] Duda R.O.; Hart P.E. and Stork D.G. (2012). *Pattern classification*. 2012: John Wiley & Sons.

[32] Sunday K.; Ocheja P.; Hussain S.; Oyelere S.; Samson B., and Agbo F. (2020). Analyzing Student Performance in Programming Education Using Classification Techniques. *International Journal of Emerging Technologies in Learning (iJET)*. 15(2): p. 127–144, 2020. https://doi.org/10.3991/ijet.v15i02.11527

[33] Asif R.; Merceron A.; Ali S.A. and Haider N.G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*. 113: p. 177–194, 2017. https://doi.org/10.1016/j.compedu.2017.05.007

[34] Ghosh C.; Saha S.; Saha S.; Ghosh N.; Singha K.; Banerjee A., and Majumder S. (2020). Machine Learning Based Supplementary Prediction System Using K Nearest Neighbour Algorithm. *Available at SSRN 3517197*, 2020. https://doi.org/10.2139/ssrn.3517197

[35] Kausar S.; Oyelere S.; Salal Y.; Hussain S.; Cifci M.; Hilcenko S.; Iqbal M.; Wenhao Z., and Huahu X. (2020). Mining Smart Learning Analytics Data Using Ensemble Classifiers. *International Journal of Emerging Technologies in Learning (iJET)*. 15(12): p. 81–102, 2020. https://doi.org/10.3991/ijet.v15i12.13455

[36] Hussain M.; Zhu W.; Zhang W. and Abidi S.M.R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience*. 20182018. https://doi.org/10.1155/2018/6347186

[37] Jishan S.T.; Rashu R.I.; Haque N. and Rahman R.M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*. 2(1): p. 1, 2015. https://doi.org/10.1186/s40165-014-0010-2

[38] Kiu C.-C. (2018). Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities. *in 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE 2018. https://doi.org/10.1109/ICACCAF.2018.8776809

[39] Pandey M. and Taruna S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*. 8: p. 364–366, 2016. https://doi.org/10.1016/j.pisc.2016.04.076

[40] Alfere S.S. and Maghari A.Y. (2018). Prediction of Student's Performance Using Modified KNN Classifiers. *Prediction of Student's Performance Using Modified KNN Classifiers*, 2018.

[41] Oladokun V.; Adebanjo A. and Charles-Owaba O. (2008). Predicting students academic performance using artificial neural network: A case study of an engineering course. 2008.

[42] Asogwa O. and Oladugba A. (2015). Of students academic performance rates using artificial neural networks (ANNs). *American Journal of Applied Mathematics and Statistics*. 3(4): p. 151–155, 2015.

[43] Lagman A.C.; Calleja J.Q.; Fernando C.G.; Gonzales J.G.; Legaspi J.B.; Ortega J.H.J.C.; Ramos R.F.; Solomo M.V.S., and Santos R.C. (2019). Embedding naïve Bayes algorithm data model in predicting student graduation. *in Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering*. 2019. https://doi.org/10.1145/3369555.3369570

[44] Liao S.N.; Zingaro D.; Thai K.; Alvarado C.; Griswold W.G., and Porter L. (2019). A robust machine learning technique to predict low-performing students. *ACM Transactions on Computing Education (TOCE)*. 19(3): p. 1–19, 2019. https://doi.org/10.1145/3277569

[45] Ma X.; Yang Y. and Zhou Z. (2018). Using Machine Learning Algorithm to Predict Student Pass Rates In Online Education. *in Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing*. 2018. https://doi.org/10.1145/3220162.3220188

[46] Amazona M.V. and Hernandez A.A. (2019). User Acceptance of Predictive Analytics for Student Academic Performance Monitoring: Insights from a Higher Education Institution in the Philippines. *in 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*. IEEE 2019. https://doi.org/10.1109/TSSA48701.2019.8985457

[47] Asiah M.; Zulkarnaen K.N.; Safaai D.; Hafzan M.Y.N.N.; Saberi M.M., and Syuhaida S.S. (2019). A review on predictive modeling technique for student academic performance monitoring. *in MATEC Web of Conferences*. EDP Sciences 2019. https://doi.org/10.1051/matecconf/201925503004

[48] Watters A. (2012). Learning Analytics: Lots of Education Data... Now What. *LAK12*, 2012.

[49] Chatti M.A.; Dyckhoff A.L.; Schroeder U. and Thüs H. (2013). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*. 4(5–6): p. 318–331, 2013. https://doi.org/10.1504/IJTEL.2012.051815

[50] Ramesh V.; Parkavi P. and Ramar K. (2013). Predicting student performance: a statistical and data mining approach. *International journal of computer applications*. 63(8): p. 35–39, 2013. https://doi.org/10.5120/10489-5242

[51] Abu Tair M.M. and El-Halees A.M. (2012). Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study*. 2(2): 2012.

[52] Mayilvaganan M. and Kalpanadevi D. (2014). Comparison of classification techniques for predicting the performance of students academic environment. *in 2014 International Conference on Communication and Network Technologies*. IEEE 2014. https://doi.org/10.1109/CNT.2014.7062736

[53] Lesinski G.; Corns S. and Dagli C. (2016). Application of an artificial neural network to predict graduation success at the United States Military Academy. *Procedia Computer Science*. 95: p. 375–382, 2016. https://doi.org/10.1016/j.procs.2016.09.348

[54] Osmanbegovic E. and Suljic M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*. 10(1): p. 3–12, 2012.

[55] Hussain S.; Dahan N.A.; Ba-Alwib F.M. and Ribata N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*. 9(2): p. 447–459, 2018. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

[56] Hamsa H.; Indiradevi S. and Kizhakkethottam J.J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*. 25: p. 326–332, 2016. https://doi.org/10.1016/j.protcy.2016.08.114

[57] Christian T.M. and Ayub M. (2014). Exploration of classification using NBTree for predicting students' performance. *in 2014 International Conference on Data and Software Engineering (ICODSE)*. IEEE 2014. https://doi.org/10.1109/ICODSE.2014.7062654

[58] Mishra T.; Kumar D. and Gupta S. (2014). Mining students' data for prediction performance. *in 2014 Fourth International Conference on Advanced Computing & Communication Technologies*. IEEE 2014. https://doi.org/10.1109/ACCT.2014.105

[59] Rizvi S.; Rienties B. and Khoja S.A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*. 137: p. 32–47, 2019. https://doi.org/10.1016/j.compedu.2019.04.001

[60] Wafi M.; Faruq U. and Supianto A.A. (2019). Automatic Feature Selection for Modified K-Nearest Neighbor to Predict Student's Academic Performance. *in 2019 International Conference on Sustainable Information Engineering and Technology (SIET)*. IEEE 2019. https://doi.org/10.1109/SIET48054.2019.8986074

[61] Budiman E.; Kridalaksana A.H. and Wati M. (2017). Performance of Decision Tree C4. 5 Algorithm in Student Academic Evaluation. *in International Conference on Computational Science and Technology*. Springer 2017. https://doi.org/10.1007/978-981-10-8276-4_36

[62] Manhães L.M.B.; da Cruz S.M.S. and Zimbrão G. (2014). WAVE: an architecture for predicting dropout in undergraduate courses using EDM. *in Proceedings of the 29th annual acm symposium on applied computing*. ACM 2014. https://doi.org/10.1145/2554850.2555135

[63] Quadri M.M. and Kalyankar N. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 2010.

[64] Kaunang F.J. and Rotikan R. (2018). Students' Academic Performance Prediction using Data Mining. *in 2018 Third International Conference on Informatics and Computing (ICIC)*. IEEE 2018. https://doi.org/10.1109/IAC.2018.8780547

[65] Chaudhury P.; Mishra S.; Tripathy H.K. and Kishore B. (2016). Enhancing the capabilities of student result prediction system. *in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. ACM 2016. https://doi.org/10.1145/2905055.2905150

[66] bin Mat U.; Buniyamin N.; Arsad P.M. and Kassim R. (2013). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. *in 2013 IEEE 5th Conference on Engineering Education (ICEED)*. IEEE 2013. https://doi.org/10.1109/ICEED.2013.6908316

[67] Kaur P.; Singh M. and Josan G.S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*. 57: p. 500–508, 2015. https://doi.org/10.1016/j.procs.2015.07.372

[68] Kotsiantis S.; Pierrakeas C. and Pintelas P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*. 18(5): p. 411–426, 2004. https://doi.org/10.1080/08839510490442058

[69] Mihăescu M.C. (2012). Classification of users by using support vector machines. *in Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*. 2012. https://doi.org/10.1145/2254129.2254211

[70] Orong M.Y.; Caroro R.A.; Durias G.D.; Cabrera J.A.; Lonzon H., and Ricalde G.T. (2020). A predictive analytics approach in determining the predictors of student attrition in the higher education institutions in the Philippines. *in Proceedings of the 3rd International Conference on Software Engineering and Information Management*. 2020. https://doi.org/10.1145/3378936.3378956

[71] Al-Radaideh Q.A.; Al-Shawakfa E.M. and Al-Najjar M.I. (2006). Mining student data using decision trees. *in International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*. 2006.

[72] Yadav S.K.; Bharadwaj B. and Pal S. (2012). Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815*, 2012.

[73] Ibrahim Z. and Rusli D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. *in 21st Annual SAS Malaysia Forum, 5th September*. 2007.

[74] Hämäläinen W. and Vinni M. (2006). Comparison of machine learning methods for intelligent tutoring systems. *in International Conference on Intelligent Tutoring Systems*. Springer 2006. https://doi.org/10.1007/11774303_52

[75] Natek S. and Zwilling M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications*. 41(14): p. 6400–6407, 2014. https://doi.org/10.1016/j.eswa.2014.04.024

[76] Ramesh V.; Parkavi P. and Ramar K. (2013). Predicting student performance: a statistical and data mining approach. *International journal of computer applications*. 63(8):2013. https://doi.org/10.5120/10489-5242

[77] Parack S.; Zahid Z. and Merchant F. (2012). Application of data mining in educational databases for predicting academic trends and patterns. *in 2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*. IEEE 2012. https://doi.org/10.1109/ICTEE.2012.6208617

[78] Ahadi A.; Lister R.; Haapala H. and Vihavainen A. (2015). Exploring machine learning methods to automatically identify students in need of assistance. *in Proceedings of the eleventh annual International Conference on International Computing Education Research*. ACM 2015. https://doi.org/10.1145/2787622.2787717

[79] Chung J.Y. and Lee S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*. 96: p. 346–353, 2019. https://doi.org/10.1016/j.childyouth.2018.11.030

[80] Solís M.; Moreira T.; Gonzalez R.; Fernandez T., and Hernandez M. (2018). Perspectives to predict dropout in university students with machine learning. *in 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. IEEE 2018. https://doi.org/10.1109/IWOBI.2018.8464191

[81] Aluko R.O.; Adenuga O.A.; Kukoyi P.O.; Soyingbe A.A., and Oyedeji J.O. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building*. 16(4): p. 86, 2016. https://doi.org/10.5130/AJCEB.v16i4.5184

[82] Mitchell R.; Michalski J. and Carbonell T. (2013). *An artificial intelligence approach*. 2013: Springer.

7 Authors

Ijaz Khan is with College of Graduate Studies Universiti Tenaga Nasional Kajang, Malaysia. Email: ijaz@buc.edu.om

Abdul Rahim Ahmad is with College of Computing and Informatics, Universiti Tenaga Nasional, Kajang, Malaysia. Email: abdrahim@uniten.edu.my

Nafaa Jabeur is with Computer Science Dept. German University of Technology, Muscat, Oman. Email: nafaa.jabeur@gutech.edu.om

Mohammed Najah Mahdi is with Institute of Informatics and Computing in Energy Universiti Tenaga Nasional, Kajang, Malaysia. Email: najah.mahdi@uniten.edu.my

Article submitted 2020-11-24. Resubmitted 2021-02-15 and 2021-04-10. Final acceptance 2021-04-12. Final version published as submitted by the authors.