Predicting Students' Performance Using Machine Learning Algorithms

Charalampos Dervenis
Department of Business Administration, University of Thessaly, Larisa, Greece

Vasileios Kyriatzis
Department of Digital Systems, University of Thessaly, Larisa, Greece

Spyros Stoufis
Computer Science Department, Hellenic Open University, Patras, Greece

Panos Fitsilis
Department of Business Administration, University of Thessaly, Larisa, GreeceComputer Science Department, Hellenic Open University, Patras, Greece

ABSTRACT

Learning analytics (LA), defined as "the measurement, collection, analysis and reporting of data about learners and their contexts for the purposes of understanding and optimizing learning and the environments in which it occurs", is a research topic that has gained significant importance and visibility among researchers during the past few decades. It is a research domain where the use of modern machine-learning (ML) algorithms and big data management provide timely and actionable information that can transform the overall learning experience for both students and educational institutions. In this paper we use ML algorithms in order to predict the performance of students, taking into account both past semester grades and socioeconomic factors. We run two models; a 2-class one predicting a "pass" or "fail" result and then we expanded this to a 5-class model, where we predict in which grading group the student will fall in the next semester. The results acquired indicate that it is possible to accurately predict the student's performance in both cases, with the 2-class model performing better than the 5-class one, which of course opts in providing more fine grain results.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Cross-validation**;

KEYWORDS

student performance, learning analytics, machine learning algorithms, KNN, SVM, Random Forest

ACM Reference Format:
Charalampos Dervenis, Vasileios Kyriatzis, Spyros Stoufis, and Panos Fitsilis. 2022. Predicting Students' Performance Using Machine Learning Algorithms. In *2022 The 6th International Conference on Algorithms, Computing and Systems (ICACS 2022), September 16–18, 2022, Larissa, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3564982.3564990

ICACS 2022, September 16–18, 2022, Larissa, Greece

ACM ISBN 978-1-4503-9740-7/22/09…$15.00
https://doi.org/10.1145/3564982.3564990

1 INTRODUCTION

1.1 Learning Analytics

Learning analytics (LA), defined as "the measurement, collection, analysis and reporting of data about learners and their contexts for the purposes of understanding and optimizing learning and the environments in which it occurs" [1], is a concrete embodiment of a larger shift in an algorithmically permeated society. Since first mentioned in the New Media Consortium (NMC) Horizon Report 2012 [2], learning analytics has gained increasing importance in the educational domain, providing valuable insight and actionable data to stakeholders. The LA definition by Siemens "the use of intelligent learner-generated data and analytics models to discover information and social connection and to predict and advise on learning" [3] and also that of Greller "a key concern of Learning Analytics is the collection and analysis of data as well as the determination of appropriate interventions to improve the learning experience of students" [4] clearly highlight the aspects and aspirations of this area of research. Indeed, educational researchers are a community interested in applying "big data" approaches in the form of learning analytics, yet the critical questions are how exactly theory could or should shape research in this new paradigm and, as Wise & Shaffer, debate, what counts as a substantial finding when the amount of data is so large that something will always be important [5].

Learning analysis researchers opting to study learning using such tools must be aware that they have adopted a particular set of ways of looking at 'learning' which reinforce and distort in particular ways, and which may inadvertently alter the system under observation. Data collection, analysis, interpretation, and even intervention (in the case of adaptive software) is no longer the preserve of the researcher but shifted to an embedded sociotechnical educational infrastructure. Thus, for teachers and students, the focus is on the ability to gain timely knowledge that could improve outcomes, such as learner performance. For a that matter, a variety of machine learning (ML)-related approaches have been proposed.

1.2 Machine Learning Algorithms and techniques

Liu predicted learner retention by combining SVM (Support Vector Machine) and a shallow neural network to improve classification accuracy [6]. Musso applied traditional artificial neural networks to predict general academic performance [7]. Kotsiantis used the

Table 1: Overview of stakeholders (Romero & Ventura 2013)

Stakeholder	Goals, benefits and prospects
Student	Support the learner with adaptive feedback, recommendations, responsiveness to his/her needs, to improve learning performance
Educational	Understanding of student learning process, reflection on teaching methods and performance, understanding of social, cognitive and behavioral aspects
Researcher	Using the right Educational Data Mining technique that fits the problem, evaluating the effectiveness of learning for different settings
Administrator	Evaluation of institutional resources and their educational offer

regression method to predict students' grades in a distance education system [8]. Wolff developed a prediction model using decision trees and SVM with data from several Open University courses to predict student performance pattern [9]. These methods are all based on shallow architectures that implement one- or two-level feature representation [10]. Predictive analytics is a group of techniques used to draw conclusions about uncertain future events. For example, in the field of education, one may be interested in predicting a measure of learning (e.g., academic success or learner skill acquisition), teaching (e.g., the impact of a particular teaching style or a particular teacher on an individual) or other proxy metrics of value to managers (e.g., retention forecasts or course enrollment). At this point, it would be beneficial to distinguish between two important lines of modelling; explanatory and predictive modeling. In explanatory modeling, forecasting is based on the assumption that a set of known data can be used to predict the value or class of new data based on observed variables. On the other hand, in predictive modeling, the goal is to create a model that will predict the values (or category if the prediction does not deal with numerical data) of new data based on various observations. Therefore, the main difference between explanatory and predictive modeling lies in the application of the model to future events, whereas in contrast to predictive modeling, explanatory modeling does not aim at future claims. Most often, this evaluation concerns the model's ability to correctly predict successes and failures in a set of learner response outcomes. Less commonly, models can be validated based on their ability to predict posttest outcomes [11] or pretest/posttest gains [12]. In the vast majority of educational data mining research, models are evaluated based on their predictive accuracy.

In recent years, many universities have been using/researching machine learning in order to gain findings about students' academic progress, predict future behaviors, identify potential problems at an early stage or even improve inter-institutional collaboration and develop an agenda for the larger community of students and teachers [13] [14]. Learning Analytics in the context of Higher Education (HE) is a suitable tool for reflecting the learning behavior of students and providing appropriate help from teachers. This individual or group support, as shown in Table 1, offers new ways of teaching and provides a way to reflect on the student's learning behavior [15] [16].

1.3 Research Questions

In our paper we used educational data about students in a course in order to perform classification with machine learning methods and to predict their performance. The research questions that guided this research are:

- Does the use of machine learning data enable more effective evaluation of training programs?
- Can we predict student performance based on personal data with machine learning?

In Section 2, we describe the methodology we followed to achieve our results. Then, in Section 3, relevant tables are used to present these results and information in a clear and concise manner. Finally, Section 4 closing remarks are provided.

2 METHODOLOGY

In this section we present and describe the stages of the process we followed to reach our results, as illustrated in the following block diagram (figure 1). Each stage is described in the following subsections.

2.1 Data Collection

We used student performance data to classify and predict student Grading. More specifically, we used the Student Performance Data Set, which is publicly available in the UCI repository (https://archive.ics.uci.edu/ml/datasets/student+performance), a popular repository of datasets for testing machine learning algorithms. These data relate to the performance of secondary school students in two Portuguese schools [17]. Data characteristics include student grades, demographic, social, and school characteristics and were collected using school reports and questionnaires. The subject we focused at was Mathematics (mat) and the variable to predict was the third quarter grade. The dataset has 649 samples (students) and 33 variables, presented in table 2.

2.2 Model Design

We created our model in Orange, as it is illustrated in Figure 2. The Orange platform [18] is an open-source data visualization, machine learning and data mining toolkit with a visual programming front-end that allows users to expedite data analysis and easily produce interactive data visualizations.

2.3 Testing Algorithms and Classification

We tested and compared different classification algorithms in order to evaluate which one achieves the highest accuracy. In particular, we tested the K-nearest neighbors (KNN), Radom Forest and Support Vector Machines (SVM) algorithms.

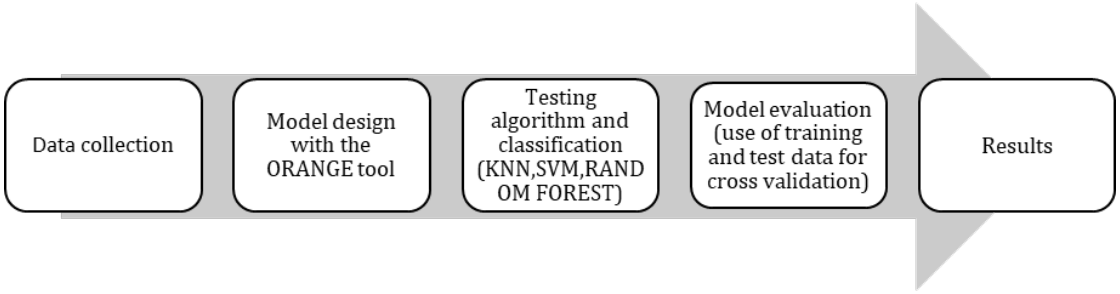

Figure 1: The procedure followed

Table 2: Model variables and description

Variable	Description
school	School student (binary: 'GP' - Gabriel Pereira ή 'MS' - Mousinho da Silveira)
sex	student gender (binary: 'F' - female ή 'M' - male)
age	student age (numeric: from 15 to 22)
address	type of residence (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' – less than or equal to 3 or 'GT3' – greater than 3)
Pstatus	parents' marital status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (arithmetic: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father's education (arithmetic: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or - higher education)
Mjob	mother's work (categorical: 'teacher', 'health' care related, civil 'services' (e.g. administrative ή police), 'at home', 'other')
Fjob	father's work (categorical: 'teacher', 'health' care related, civil 'services' (e.g. administrative ή police), 'at home', 'other')
reason	reason for school choice (categorical: close to 'home', school 'reputation', 'course' preference, 'other')
guardian	student guardian (categorical: 'mother', 'father' or 'other')
travel time	school-home travel time (arithmetic:1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, 4 - >1 hour)
study time	weekly study time (arithmetic:1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, 4 - >10 hours)
failures	number of previous failures in the course (arithmetic: n if 1<=n<3, else 4)
schools up	additional school support (binary: yes or no)
famsup	educational support from family (binary: yes, ή no)
paid	additional paid courses (Math or Portuguese) (binary: yes or no)
activities	extracurricular activities (binary: yes or no)
nursery	went to kindergarten (binary: yes or no)
higher	wants higher education (binary: yes or no)
internet	Internet at home (binary: yes or no)
romantic	with relation (binary: yes, or no)
famrel	level of family relationships (arithmetic: from 1 - very bad to 5 - excellent)
free time	free time after school (arithmetic: from 1 - very low to 5 - very high)
Go out	going out with friends (arithmetic: from 1 - very low to 5 - very high)
Dalc	alcohol consumption on weekdays (arithmetic: from 1 - very low to 5 - very high)
Walc	alcohol consumption SCs (arithmetic: from 1 - very low to 5- very high)
health	health status (arithmetic: from 1 - very bad to 5 - very good)
absences	number of absences (arithmetic: from 0 to 93)
G1	first term grade (arithmetic: from 0 to 20)
G2	second term grade (arithmetic: from 0 to 20)
G3	third term grade (arithmetic: from 0 to 20)

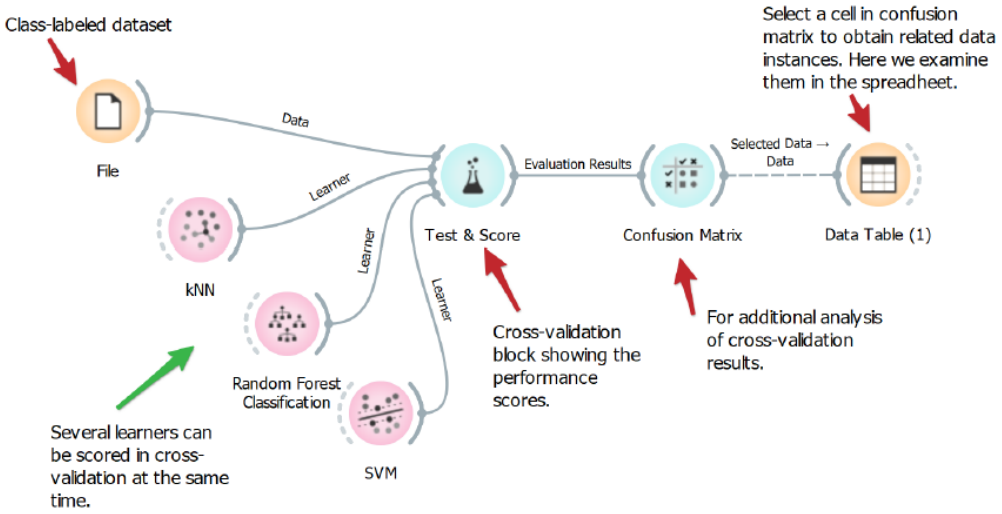

Figure 2: Model designed

2.3.1 K-Nearest-Neighbors (KNN).. KNN is a simple but efficient classification algorithm, hence it is quite widespread [19]. For each sample of the test data, it finds the K nearest neighbors from the training data. Finding the nearest neighbors is done with some distance metric, such as the Euclidean distance [20]. It then finds which class of the K neighbors has the majority and returns as output.

2.3.2 Support Vector Machine (SVM).. In the SVM algorithm the categorization of the data is based on finding an optimal line (for two-dimensional data) or an optimal hyperplane (for higher dimensions) that separates the data creating the maximum margin [21]. The ability to generalize the use of SVMs to non-linear data relies on the kernel trick. In the event that linear separation is not possible, appropriate visualizations are used that transfer the set of data to a larger dimension in order to finally achieve their separation [22] [23]. SVM is a binary classifier, i.e. it has the ability to categorize into two classes. A common kernel is the radial basis function:

$$f(x_1, x_2) = \exp\left(-\gamma |x_1 - x_2|^2\right) \quad (1)$$

Where x_1, x_2 two points and γ a parameter of the function.

2.3.3 Random Forest. Random Forest is a classification method that uses a large number of Classification and Regression Trees (CART) in order to provide higher accuracy than a single decision tree [24] [25]. Random Forest generates a large number of unpruned trees, which are quite different from each other due to their random construction. Thus, the trees are not correlated with each other, so Random Forests can avoid overfitting to the training data and can achieve higher accuracy than a single tree. Moreover, they can handle large data sets efficiently and can be used for both classification and regression.

2.4 Model Evaluation

For evaluation purposes, training and test data are needed. An algorithm builds a model based on training data, but its performance

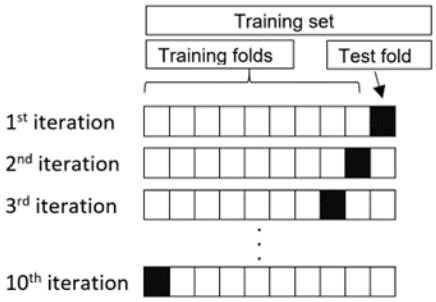

Figure 3: Diagram of k-fold cross-validation with k = 10

is measured against test data that has not been used to build the model. Since there is usually only one data set, the cross-validation technique [26] is widely used to evaluate an algorithm. Under this technique, we repeatedly divide the data set 10 times (figure 3) into training data (90%) and test data (10%) (10-fold cross-validation) so that all data are passed through the test subset, in which the outcome is predicted for evaluation.

To measure performance, we calculate the confusion matrix, as shown below in table 3, for two categories, which we call Positive and Negative:

That is, in the rows we have the actual class and in the columns the class predicted by the algorithm. Correct predictions are on the diagonal (starting from cell (1,1)).

For the evaluation, the following metrics were used:

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})}. \quad (2)$$

$$Precision = \frac{TP}{TP + FP)} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Table 3: Confusion Matrix

		Category prediction	
		Negative	Positive
Real Category	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 4: Evaluation Results (2-class implementation)

Algorithm	AUC	CA	F1	Precision	Recall
KNN	0.938	0.881	0.881	0.880	0.881
SVM	0.904	0.833	0.829	0.830	0.833
Random Forest Learner	0.968	0.906	0.907	0.907	0.906

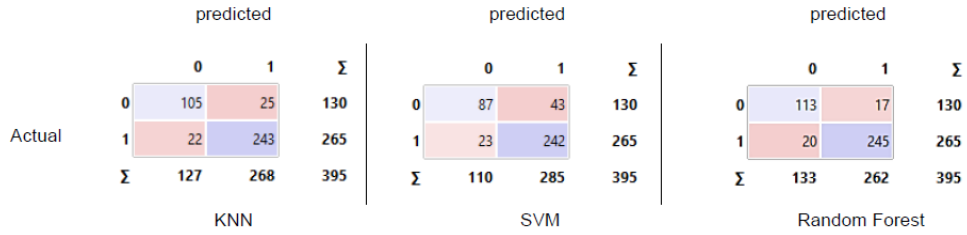

Figure 4: KNN, SVM and Random Forest confusion matrix (2-class implementation)

$$F1score = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (5)$$

Where,
Accuracy: The classification accuracy (CA) of the proposed system.
Precision: Out of all the positive predicted, what percentage is truly positive.
Recall: Out of the total positive, what percentage are predicted positive.
F1score (or F1): A weighted average of precision and recall.

3 RESULTS AND DISCUSSION

3.1 Class Implementation

In this run, we used 2 classification categories, pass or fail (i.e. predict if the third semester grade is going to be less than or greater than 10), and tested the three classification models described earlier. The values of the parameters K and N used were the following; for KNN we choose number of neighbors K=5, for Random Forest we increased the trees to N=50, while for SVM we choose RBF kernel. These values are widely used in the literature and in the algorithms' documentation, leading to good results. The evaluation results of the models are shown in Table 4.

As we can see, the metrics AUC (area under curve), CA (classification accuracy), F1, Precision and Recall are indicative of the performance of the algorithms (the higher the values, the higher the performance). Between all three cases we have the best result with the Random Forest algorithm, followed by KNN and finally SVM. The same holds if we compare the algorithms over CA, F1, precision and recall. It is noted that two-class random classification has an accuracy of 50%, so the level achieved (90.6%) for Random Forest is extremely satisfactory.

An important observation stemming from the confusion matrices of figure 4 is that all algorithms' performance is commensurate in predicting a "pass"/"fail" classification. For example, Random Forest predicts 113 failures (130 are true failures) and 245 successes (265 are true successes), while KNN predicts 105 failures and 243 successes (88.1% accuracy) and SVM predicts 87 failures and 242 successes (83.3% accuracy). Therefore, it is safe to claim that indeed, we are in a position to safely predict, taking into account both past semester grades and socioeconomic factors, whether a student will pass of fail the math course next semester.

3.2 Class Implementation

In order to further investigate the algorithms' capability in providing more fine-grain predictions of the students' performance, we increased the number of classes from 2 ("pass"/ "fail") to 5 (see table 5), and keeping the rest of the parameters unaltered, we rerun the models. The class distribution used is adopted from the original work of Cortez and Silva and shown in Table 5, along with the distribution of the dataset's grades per class [17].

As we can see in table 6, AUC, the measure of the ability of the classifier to distinguish between classes, is 90.5% for Random Forest, 87.1% and 82.2% for KNN and SVM respectively, indicating a good performance for our model. However, we notice that the percentages for CA are lower than those obtained with the 2-class implementation (67.6% vs 90.6% for Random Forest). Nevertheless,

Table 5: Categories of students grades (for Math)

Country	0 (fail)	I (sufficient)	II (satisfactory)	III (good)	IV (excellent/very good)
Portugal/France	0-9	10-11	12-13	14-15	16-20
Student distribution per grade category	130	103	62	60	40

Table 6: Evaluation Results (5-class implementation)

Algorithm	AUC	CA	F1	Precision	Recall
KNN	0.871	0.600	0.597	0.602	0.600
SVM	0.822	0.491	0.466	0.488	0.491
Random Forest Learner	0.905	0.676	0.666	0.672	0.676

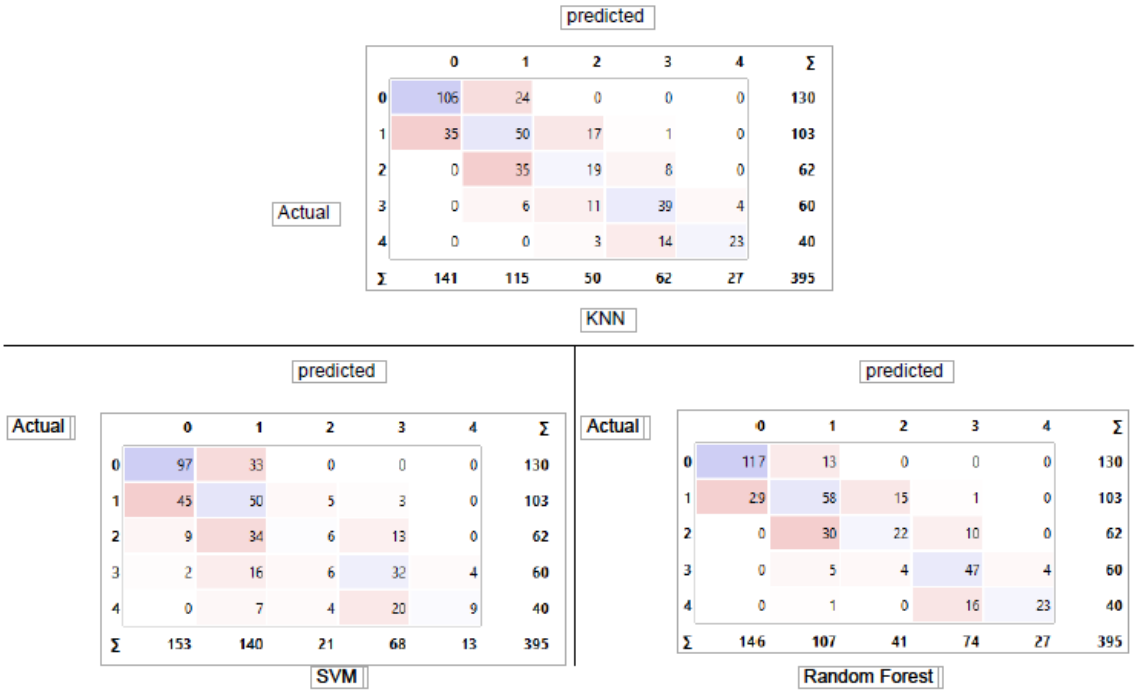

Figure 5: KNN, SVM and Random Forest confusion matrix (5-class implementation)

since we are dealing with imbalanced data and non-binary classification, AUC is deemed as a more fit metric to use in order to assess the model's performance.

This can be also deducted from the confusion matrices of figure 5, where for the classification to be correct, we must have the larger values on the diagonal of the confusion matrix. Indeed, in the case of the Random Forest implementation, we can see that the errors are mostly evident in class I and class II, while for the rest of the classes the error is quite negligible, and in all cases, it mostly concerns classifying a sample in the immediately neighboring classes, with almost 81.3% of data points getting classified in the correct class. Overall, in this 5-class implementation, the performance of the KNN and especially of the SVM algorithms deteriorate compared to the 2-class implementation, suggesting that either they are not fit for the purpose used (especially SVM) or their parameters need to be adjusted in order to perform better.

4 CONCLUSION

In this case study, we used a dataset of student scores in one module. We initially investigated a two-category classification, i.e., whether the student passed the course or not. The results were very satisfactory, with our model predicting the third semester "pass" or "fail" with a very high level of precision. In order to investigate the algorithms' performance in giving more fine grain result, we then

classified student grades into 5 categories. The results in this case were also satisfactory, with the best algorithm based on the AUC metric being Random Forest (as compared to SVM and kNN). Analysis of the results, using confusion matrices, revealed that although some of the performance indicators were reduced when compared to the two class implementation, the results are commensurately high taking in consideration that with the five class implementation we opt for more fine grain classification results. It is well established that a huge amount of educational data is generated every day and remains untapped. Educational institutions must, by all means exploit this data in order to get insight and support accurate and timely interventions towards improving various aspects of educational services provided. Our approach revealed that techniques and methods using machine learning algorithms can contribute in harnessing this vast amount of data with multifaceted benefits for the entire educational community.

REFERENCES

[1] Lias, T. E., & Elias, T. (2011). Learning analytics: The definitions, the processes, and the potential.

[2] Johnson, L., Brown, S., Cummins, M., & Estrada, V. (2012). The technology outlook for STEM+ education 2012-2017: an NMC horizon report sector analysis (pp. 1-23). The New Media Consortium.

[3] Siemens, G. (2010). What are learning analytics? eLearnspace. Recuperado de http://www. elearnspace. org/blog/2010/08/25/what-are-learning-analytics/. Consultado el, 1(07), 2012.

[4] Greller, W., Ebner, M., & Schön, M. (2014, June). Learning analytics: From theory to practice–data support for learning and teaching. In International Computer Assisted Assessment Conference (pp. 79-87). Springer, Cham.

[5] Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2(2), 5-13.

[6] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8759-8768).

[7] Musso, M. F., Kyndt, E., Cascallar, E. C., & Dochy, F. (2013). Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks. Frontline Learning Research, 1(1), 42-71.

[8] Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artificial Intelligence Review, 37(4), 331-344.

[9] Wolff, A., Zdrahal, Z., Herrmannova, D., & Knoth, P. (2014). Predicting student performance from combined data sources. In Educational data mining (pp. 175-202). Springer, Cham.

[10] Bengio, Y. (2009). Learning deep architectures for AI. Foundations and trends® in Machine Learning, 2(1), 1-127.

[11] Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), 253-278.

[12] Liu, R., & Koedinger, K. R. (2015). Variations in Learning Rate: Student Classification Based on Systematic Residual Error Patterns across Practice Opportunities. International educational data mining society.

[13] Atif, A., Richards, D., Bilgin, A., & Marrone, M. (2013). Learning analytics in higher education: a summary of tools and approaches. In ASCILITE-Australian Society for Computers in Learning in Tertiary Education Annual Conference (pp. 68-72). Australasian Society for Computers in Learning in Tertiary Education.

[14] Iatrellis, O., Savvas, I. K., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. Education and Information Technologies, 26(1), 69-88.

[15] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

[16] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.

[17] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

[18] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research* 14(Aug): 2349−2353

[19] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 1-19.

[20] Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. ACM Computing Surveys (CSUR), 54(6), 1-25.

[21] Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. In Data mining techniques for the life sciences (pp. 223-239). Humana Press.

[22] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

[23] Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. Artificial Intelligence Review, 52(2), 857-900.

[24] Resende, P. A. A., & Drummond, A. C. (2018). A survey of random forest based methods for intrusion detection systems. ACM Computing Surveys (CSUR), 51(3), 1-36.

[25] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[26] Ashfaque, J. M., & Iqbal, A. (2019). Introduction to support vector machines and kernel methods. publication at https://www. researchgate. net/publication/332370436.