EXPLORING PREDICTING PERFORMANCE OF ENGINEERING STUDENTS USING DEEP LEARNING

Imran Zualkernan
Computer Science and Engineering
American University of Sharjah, UAE

ABSTRACT

A significant amount of research has gone into predicting student performance and many studies have been conducted to predict why students drop out. A variety of data including digital footprints, socio-economic data, financial data, and psychological aspects have been used to predict student performance at the test, course, or program level. Fairly good prediction results have been achieved using both traditional machine learning and more recently deep learning methods. While using diverse sets of data has achieved good results, this data is often difficult and expensive to collect, and may have privacy-related issues. This paper explores the extent to which only prior performance data readily available with registrars in most Universities can be used to predict student performance in future terms. Twenty term data from 789 students enrolled an engineering program at an American University were used to train long term short term (LSTM), Bi-directional LSTM and Gated Reference Units (GRU) models to predict student performance in future terms. The results are that all three types of models were able to reasonably predict the next term's performance (F1-score of about 0.70) regardless of the number of terms a student had spent the University. The models generally did not overfit. The prediction was reasonable until about trying to predict performance on seventh term in the future, but the performance dropped beyond this point primarily due to lack of sufficient data (F1-score of about 0.2).

KEYWORDS

Performance Prediction, Higher Education, Engineering, Deep Learning, LSTM, GRU

1. INTRODUCTION

Predicting student performance is a classical problem in educational data mining. A recent survey showed that student success could be predicted based on a number of variables including information from digital platforms, student demographics, socio-economic status, prior academic performance, course load, relationship to the educational institutions, access to counselling, and psychological factors, etc. (Ifenthaler and Yau, 2020). For example, engagement data from a learning analytics platform was a good predictor of student success for first year students (Foster and Siddle, 2020). Family commitments, financial strain, time management, expected study load, and work commitments were also found to be related to early dropout (Nieuwoudt and Pedler, 2021). A student's relationship to the educational institution also played a role as students who were strongly 'fused' with their university were more likely to not dropout (Talaifar et al., 2021). Institutional fit, high school performance, and financial aid were also significant predictors of dropout (Elder, 2021). Prior academic performance is important in predicting success. For example, first year grades were found to be the best predictor of graduation at the University level (Willoughby et al., 2021). A host of psychological factors like academic exhaustion, satisfaction with education and dropout intention have also been explored as well (e.g., (Casanova et al., 2021)).

Data required to predict performance can be divided into three broad categories. First, some data like digital platform data, or prior performance is readily available with the academic information technology (IT) systems. Second, other data like financial aid data can be requested from external information providers. Thirdly, data related to socio-economic factors and psychological factors may require conducting surveys or deploying similar instruments to collect and create data. In most higher education environments, the performance data including students' GPA on every course taken must be collected and retained by the registrar and hence this

data is always available in most higher education contexts and does not require any additional effort or cost for collection.

In higher education, students' performance can be predicted within a course, across semesters or across programs. For example, based on performance on a few quizzes, one can predict how a student will perform on final examination or overall, in the course. Similarly, one could also use performance on multiple courses within some number of terms (e.g., first four terms) to predict how well they will perform in the next term or subsequent terms. Finally, one could also compare students' performance to others to determine how well they will perform in the program overall. For example, would they be able to graduate or not. This paper explores if prior performance data alone can be used to predict future performance of students based on course GPA's from a set of terms. The primary contribution of this paper is building time-based models to predict student performance in future terms based on course GPA data alone.

Rest of the paper is organizes as follows. Related work is discussed next. This is followed by a decription of the methodology used. Results are presented next, and the paper ends with a discussion and a conclusion.

2. RELATED WORK

Prior research has shown that most work in machine learning (38%) is about predicting the final grade within a course, this is followed by work (14.7%) to predict an exam grade. The other predictive variables include program retention (13.4%), predicting the GPA (12.2%) and performance on a specific assignment (11.4%) (Hellas et al., 2018). Machine learning has achieved significant performance (e.g., 90-95% accuracy) in some cases but also has also failed to perform in many others (48-76% accuracy) (Namoun and Alshanqiti, 2021). A wide range of machine learning methods have been used for predicting student performance. Traditional regression models have been used to predict student dropout (Hippel and Hofflinger, 2021). Traditional machine learning methods like Support Vector Machines (SVM) and Boosted Trees with ensemble methods were used to achieve accuracies ranging from 71% to 93.5% depending on the term in which prediction was made; earlier terms yielding lower accuracies (Hannaford et al., 2021). Boosted trees were also used achieve an accuracy of 91% to predict student dropout (Oreshin et al., 2020). Similarly, AutoML was used with ensemble models on admission data of students to achieve accuracy rates of about 75% (Zeineddine et al., 2021). Similarly, SVM and Random Forest were used to predict first year student dropout with an accuracy of 85% (Del Bonifro et al., 2020). Other traditional machine learning techniques like Bayesian Belief Network (BBN) have been used to predict student performance with an accuracy of 76% (Delen et al., 2020). Finally, Naïve Bayes was used to achieve an accuracy of 76% across two cohorts.

Neural networks have also been used to predict performance. For example, multi-layered perceptron was also used to predict student failure with an F1-Score of 0.83 (Karimi-Haghighi et al., 2021). Similarly, neural networks were used to predict students at risk with an accuracy of 83.7%. At the course level a neural network outperformed traditional machine learning methods like SVM, K-NN etc. and achieved an F1-Score of 0.96 (Tomasevic et al., 2020). In order to explicitly cater for time-nature of performance (activity on a campus) recurrent networks and SVM have been used to predict student performance (Wang et al., 2020). One problem with performance data is the small size of these educational data sets. The ICGAN-DSVM algorithm combined Generative Adversarial Networks (GANs) with SVM to achieve better performance than supervised learning methods alone (Chui et al., 2020).

3. METHDOLOGY

3.1 Data

The data used for this paper was drawn from the registration system of an engineering program of a college of engineering at an American University. The data ranged from 2014 to 2019. Only the courses taught in the engineering program were considered. For example, humanities, Math, and other courses were excluded. The data included accumulated hours, grade point average (GPA) in each course taken, the instructor for each

course, academic status, and the admitting cohort of the student. The data consisted of 789 students who had registered in a total of 10,508 individual courses over this time. Many students had only taken one course so far with a maximum of 43 courses taken and a mean of 13.318 courses attempted per student. The total number of courses available to students in this engineering program was 64. Students could register for multiple courses in a term and four terms (including the summer terms) were counted per year. Regardless of cohort, the data was normalized for each student to start in term 1.

3.2 Neural Network Models

Since the purpose of this paper was to use the timed sequence of student performance to predict future performance, three commonly used time-based neural network prediction models were used. Each is described below.

3.2.1 Long Short-Term Memory Networks (LSTM)

Since their proposal in 1995, recurrent neural networks with long short-term memory (LSTM) have been successful in handling sequential data in a variety of domains (Greff et al., 2017). LSTM architecture is based on a memory cell that maintains its state over time. LSTM uses nonlinear gating units that regulate the flow of information in and out of the cell. An LSTM can take an *n* sized input sequence $x^1, x^2, .. x^n$ where each element of the sequence x^i can be represented by a fixed set of features $f_1^i, f_2^i, .. f_k^i$. In this paper, since the input to LSTM consisted of a sequence of terms with associated GPA in each course, x^i represented the *i*th term while the GPA in each course was represented by a feature f_l^i. For example, the engineering program considered had a total of 64 upper-level courses (i.e., $k = 64$) which were typically attempted by each student. The second term (i.e., x^2) for a particular student was hence represented by the feature vector $f_1^2, f_2^2, .. f_{64}^2$ where each f_j^2 either represented the GPA in a course that was attempted in this term or a very small number ($\varepsilon = 0.0001$) indicating that course was not attempted in this term.

Within this formulation, the LSTM was trained using variable length sequences $x^1, x^2, .. x^l$ as input and performance to be predicted in future terms $p(x^l), p(x^{l+1}), p(x^{l+3})$,.. etc. For example, one could use the data from the first term only (i.e., x^1) to predict performance in the third term (i.e., $p(x^3)$). Performance can either by the actual cumulative GPA of the student in the said term or a classification into a performance category based on the cumulative GPA. For a term x, this paper defined *p* as shown in Eq. (1). In other words, if the student's cumulative GPA in term x was less than 2 out of 4, then they were classified as weak with a code of 1, and so on; the student's performance was categorized in one of the three categories depending on their cumulative GPA.

$$p(x) = \begin{cases} 1 & 0 <= cgpa(x) <= 2 \\ 2 & 2 < cgpa(x) <= 3 \\ 3 & 3 < cgpa(x) <= 4 \end{cases} \quad (1)$$

The training data for the LSTM consisted of batches of variable length sequences of terms (e.g., first two terms) followed by a predicator based on a subsequent term. For example, one set of training data consisted of $< x^1, p(x^2) >$ in predicting student performance in the second term based on the GPA of courses taken in the first term (i.e., x^1) only. Similarly, $< x^1, x^2, .. x^p, p(x^q) >$ represents a training set that used the data from the first *p* terms to predict performance in the *q*th term where $q > p$. All such valid sequences were used to train a single LSTM for each *p* and *q*.

After the LSTMs were trained, given the performance of a student $x^1, x^2, .. x^p$ as input the respective LSTM could predict the performance $p(x^q)$ in the *q*th subsequent term.

Figure 1 shows the number of terms completed verses the number of batches when predicting performance in the next term. Each batch consists of data of the form $< x^1, x^2, .. x^p, p(x^q) >$ for some *p* and *q*. For example, for each *p* terms (e.g., one term) completed, Figure 2 has data in the form of $< x^1, x^2, .. x^p, p(x^{p+1}) >$ as we are trying to predict the GPA in the next (i.e., *p+1*) term. As the figure shows, the number of batches is reduced as we get closer to higher terms because fewer students have taken that many courses.

Figure 2 shows graphical representation of various inputs $< x^1, x^2, .. x^p, p(x^{p+1}) >$ for p = 15. Each colored dot represents a grade point average (GPA) in a course. As the Figure shows, lighter colors mean higher GPA. The Figure also demonstrates that there does not seem to be an obvious pattern on how the various students choose to attempt the courses within the same engineering program.

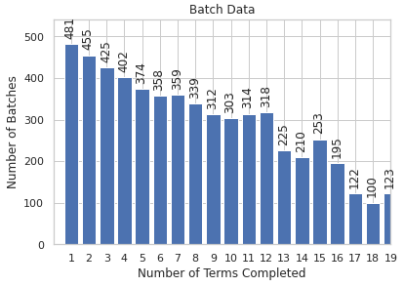

Figure 1. Number of batches completed to predict performance in the next term

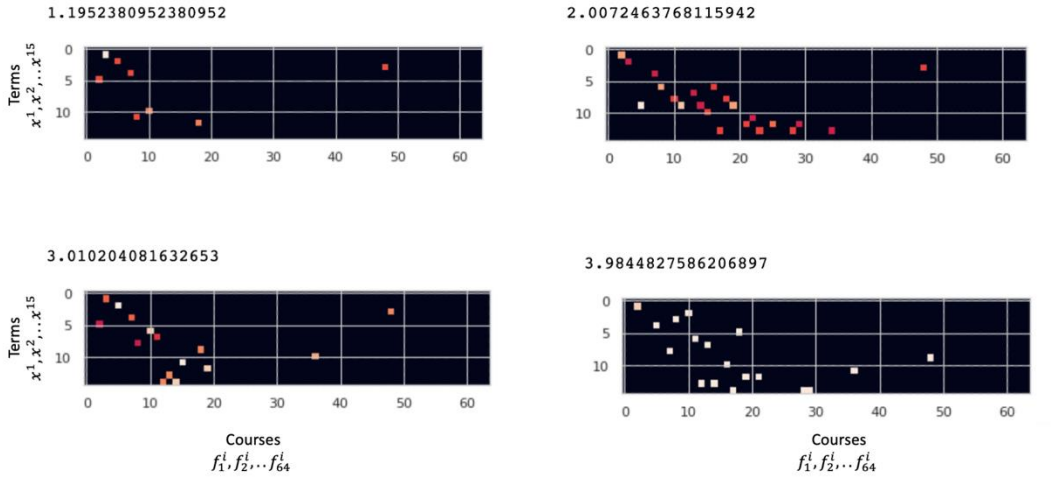

Figure 2. Randomly selected students with different cumulative GPAs for batch size = 15 terms

3.2.2 Gated Reference Units (GRU)

Recurrent networks with gated recurrent units (GRU) emerged in 2014 as a simpler form of LSTM and are more efficient (Chung et al., 2014). Since the basic formulation is the same as the LSTM, the same data was used for training as the LSTM but replacing the LSTM with a GRU.

3.2.3 Bi-Directional LSTM

Bi-directional LSTM is another version of an LSTM that processes the input sequence both forward and in backwards direction to make the predictions (Graves and Schmidhuber, 2005). Since the the basic formulation is the same as a normal LSTM, the same data was used for training as an LSTM but the LSTM was replaced with a bi-directional LSTM.

3.3 Training

One model of each type (e.g., LSTM) for each forward prediction capability was trained on the data set. The data set used an 80:20 split for training/testing and a 70:30 split of training data into training/validation sets.

In each case a two-level stacked model with 16 internal nodes was found to work best. The learning rate was 0.001. 20% dropout was used between the two stacks to prevent overfitting. Each model was trained for 50 epochs and a batch size of 16 was used. Generators were used to dynamically feed batches to the models for training. For example, if the model was to predict *n* terms ahead, then the model was trained on various batches $< x^1, x^2, .. x^p, p(x^{p+n}) >$ where p could vary from 1 to the maximum number of terms available that allowed prediction to $p(x^{p+n})$. For example, since only data for 20 terms was available, one could only use data from the first 4 terms to be able to predict 16 terms forward. Hence, the models that were trying to predict farther in the future had lesser data available for training.

4. RESULTS

Figure 3 shows the multiple performance metrics for the various models trained for each predict forward capability. As the Figure shows, the LSTM model to predict performance one term ahead was quite reasonable with a macro F1 Score of 0.70. However, as the prediction horizon becomes longer the performance deteriorates and the macro F1 Score drops to drops to 0.19 for predicting 17 terms ahead. Similarly, GRU performed similarly to the LSTM with a macro F1 Score of 0.72 for one term look ahead. The performance dropped to a macro F1 Score of 0.22 for predicting 17 terms ahead. Finally, Bidirectional LSTM models also performed similarly with an F1-score of 0.72 for one term look ahead. The performance drops to 0.24 for predicting 17 terms ahead.

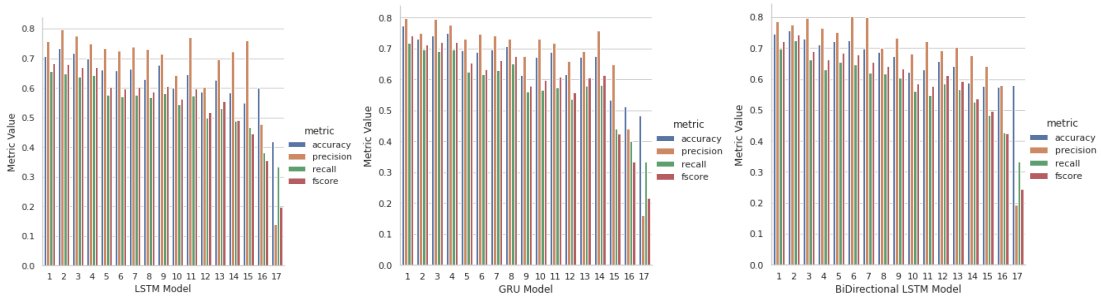

Figure 3. Performance metrics for various models with different forward predictions

As Figure 4 shows, the LSTM models are robust under overfitting as the validation and the training loss follow each other. However, overfitting begins to occur when the number of data points for training become quite small as shown in Figure 6 (f) where validation loss diverged from the training loss.

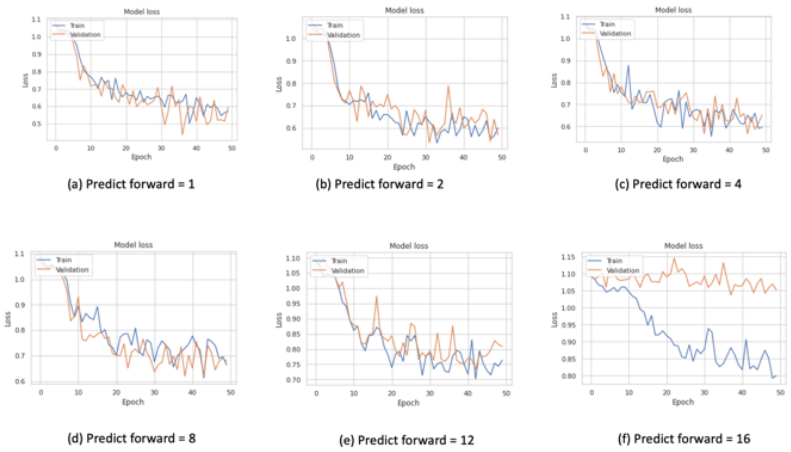

Figure 4. Loss while training for various LSTM models with different forward predictions

Similarly, as Figure 5 shows, like the LSTM, the GRU models also did not overfit. In fact, in this particular training regimen the GRU models did not seem to overfit even for very low amount of data. This can perhaps be attributed to the fact that GRU models generally require much lesser number of parameters as opposed to an LSTM. Finally, as Figure 8 shows, like the normal LSTM, the Bi-Directional LSTM also begins to overfit when the amount of training data becomes very small. However, it tends to not overfit until even when trying to predict 12 terms ahead.

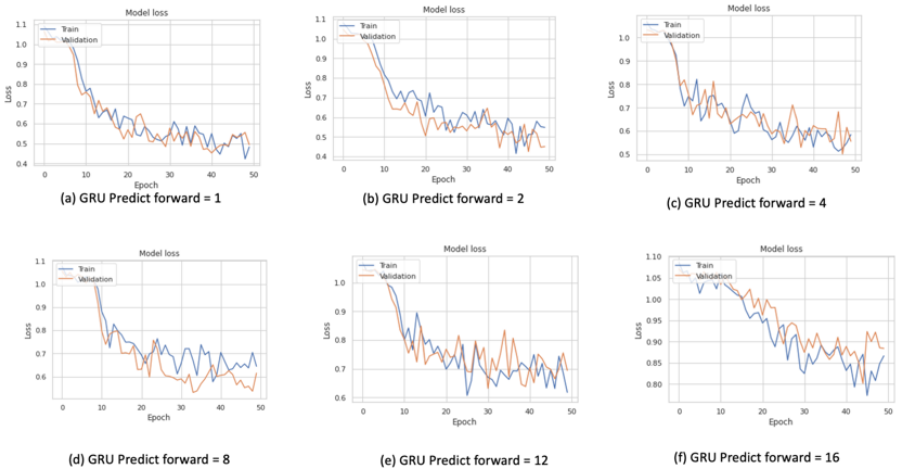

Figure 5. Loss while training for various GRU models with different forward predictions

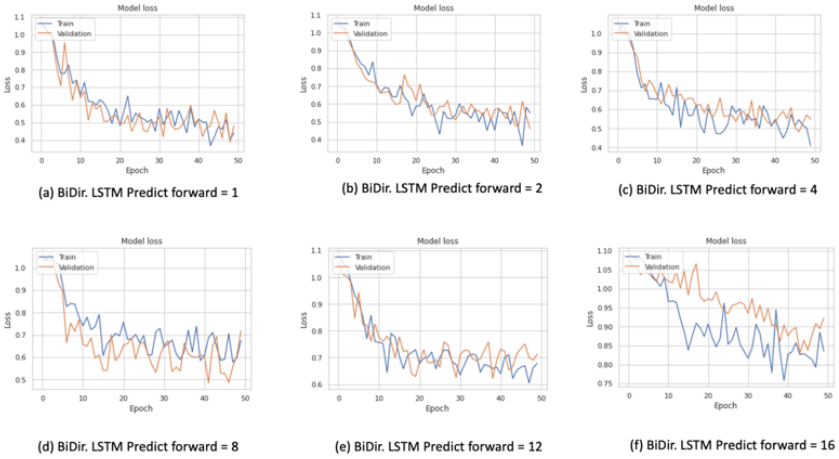

Figure 6. Loss while training for various Bi-directional LSTM models with different forward predictions

5. DISCUSSION

As Figure 7 shows, all three model types performed similarly with respect to the macro F1 Score. For all three model types, the performance dropped significantly when trying to predict beyond seven terms in the future. The results are not spectacular as many have achieved much higher performance by using a variety of data sources. However, given that only performance data was used and that a single model can predict performance in the future regardless of how many terms are counted as input, these initial results are promising.

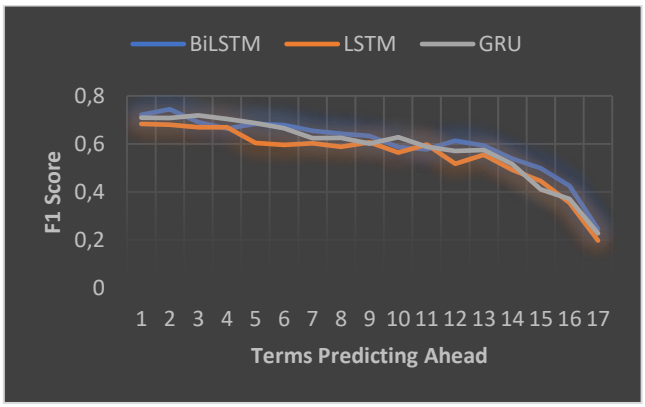

Figure 7. Macro F1 Scores for models trying to predict various terms in the future

6. CONCLUSION

This paper has explored the extent to which course-level GPA alone within an engineering program could be used to predict student performance in future terms. Several models were built based on the how far in the future predictions were to be made. Each model could be used to predict performance even if the student had spent a few terms in the University. The results are reasonable for predictions into up to seven terms in the future. Obviously, the results are limited to one engineering program alone. It will be interesting to compare the results with traditional machine learning techniques like SVM that seemed to have performed well elsewhere using a wider set of data. Similarly, GAN-based architectures to augment data could perhaps also be explored to cater for the low amount of data available in such educational environments

REFERENCES

Casanova, J.R., Gomes, C.M.A., Bernardo, A.B., Núñez, J.C., Almeida, L.S., 2021. Dimensionality and reliability of a screening instrument for students at-risk of dropping out from Higher Education. Studies in Educational Evaluation 68, 100957. https://doi.org/10.1016/j.stueduc.2020.100957

Chui, K.T., Liu, R.W., Zhao, M., De Pablos, P.O., 2020. Predicting Students' Performance With School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine. IEEE Access 8, 86745–86752. https://doi.org/10.1109/ACCESS.2020.2992869

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs].

Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P., 2020. Student Dropout Prediction, in: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (Eds.), Artificial Intelligence in Education, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 129–140. https://doi.org/10.1007/978-3-030-52237-7_11

Delen, D., Topuz, K., Eryarsoy, E., 2020. Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. European Journal of Operational Research, Featured Cluster: Business Analytics: Defining the field and identifying a research agenda 281, 575–587. https://doi.org/10.1016/j.ejor.2019.03.037

Elder, A.C., 2021. Holistic factors related to student persistence at a large, public university. Journal of Further and Higher Education 45, 65–78. https://doi.org/10.1080/0309877X.2020.1722802

Foster, E., Siddle, R., 2020. The effectiveness of learning analytics for identifying at-risk students in higher education. Assessment & Evaluation in Higher Education 45, 842–854. https://doi.org/10.1080/02602938.2019.1682118

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM networks, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. Presented at the Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., pp. 2047–2052 vol. 4. https://doi.org/10.1109/IJCNN.2005.1556215

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems 28, 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Hannaford, L., Cheng, X., Kunes-Connell, M., 2021. Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study. Nurse Education Today 99, 104784. https://doi.org/10.1016/j.nedt.2021.104784

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N., 2018. Predicting academic performance: a systematic literature review, in: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, ITiCSE 2018 Companion. Association for Computing Machinery, New York, NY, USA, pp. 175–199. https://doi.org/10.1145/3293881.3295783

Hippel, P.T.V., Hofflinger, A., 2021. The data revolution comes to higher education: identifying students at risk of dropout in Chile. Journal of Higher Education Policy and Management 43, 2–23. https://doi.org/10.1080/1360080X.2020.1739800

Ifenthaler, D., Yau, J.Y.-K., 2020. Utilising learning analytics to support study success in higher education: a systematic review. Education Tech Research Dev 68, 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Karimi-Haghighi, M., Castillo, C., Hernandez-Leo, D., Oliver, V.M., 2021. Predicting Early Dropout: Calibration and Algorithmic Fairness Considerations. arXiv:2103.09068 [cs].

Namoun, A., Alshanqiti, A., 2021. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Applied Sciences 11, 237. https://doi.org/10.3390/app11010237

Nieuwoudt, J.E., Pedler, M.L., 2021. Student Retention in Higher Education: Why Students Choose to Remain at University. Journal of College Student Retention: Research, Theory & Practice 1521025120985228. https://doi.org/10.1177/1521025120985228

Oreshin, S., Filchenkov, A., Petrusha, P., Krasheninnikov, E., Panfilov, A., Glukhov, I., Kaliberda, Y., Masalskiy, D., Serdyukov, A., Kazakovtsev, V., Khlopotov, M., Podolenchuk, T., Smetannikov, I., Kozlova, D., 2020. Implementing a Machine Learning Approach to Predicting Students’ Academic Outcomes, in: 2020 International Conference on Control, Robotics and Intelligent System, CCRIS 2020. Association for Computing Machinery, New York, NY, USA, pp. 78–83. https://doi.org/10.1145/3437802.3437816

Talaifar, S., Ashokkumar, A., Pennebaker, J.W., Medrano, F.N., Yeager, D.S., Swann, W.B., 2021. A New Pathway to University Retention? Identity Fusion With University Predicts Retention Independently of Grades. Social Psychological and Personality Science 12, 108–117. https://doi.org/10.1177/1948550619894995

Tomasevic, N., Gvozdenovic, N., Vranes, S., 2020. An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & Education 143, 103676. https://doi.org/10.1016/j.compedu.2019.103676

Wang, X., Yu, X., Guo, L., Liu, F., Xu, L., 2020. Student Performance Prediction with Short-Term Sequential Campus Behaviors. Information 11, 201. https://doi.org/10.3390/info11040201

Willoughby, T., Dykstra, V.W., Heffer, T., Braccio, J., Shahid, H., 2021. A Long-Term Study of What Best Predicts Graduating From University Versus Leaving Prior to Graduation. Journal of College Student Retention: Research, Theory & Practice 1521025120987993. https://doi.org/10.1177/1521025120987993

Zeineddine, H., Braendle, U., Farah, A., 2021. Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering 89, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903