Vol. 11, No. 3, pp. 650-658, 2021 DOI: 10.23960/jpp.v11.i3.202116

Jurnal Pendidikan Progresif

e-ISSN: 2550-1313 | p-ISSN: 2087-9849

http://jurnal.fkip.unila.ac.id/index.php/jpp/

Predicting On-time Graduation based on Student Performance in Core Introductory Computing Courses using Decision Tree Algorithm

Jeffrey Co [1], Niel Francis Casillano[1*]

[1]Eastern Samar State University, Borongan City, Philippines

*Corresponding email: nfcasillano@gmail.com

Received: 11 November 2021 *Accepted: 21 December 2021* *Published: 30 December 2021*

Abstract: **Objectives**: This study primarily aimed at developing a model that will predict whether a student will graduate on time based on their academic performance in their respective core introductory computing courses. **Methods:** The educational data mining process was employed in the conduct of this research. The process commenced with the collection of educational data and culminated with the evaluation of the developed model. This research utilized the decision tree algorithm. **Findings:** The model evaluation resulted to an 88.9% classification accuracy where the total number of actual "Yes" (students who graduated on-time) is 52.49 were classified correctly and 3 were misclassified as "No" in the prediction and the total number of actual "No" (students who did not graduated on-time) is 20.15 of which were classified correctly and 5 were misclassified in the prediction. **Conclusion:** Results of the study can be used as inputs in the crafting of new resource materials and an improved curriculum that will help improve the performance of students in the database management course. The model can also be used as a tool to help students graduate on-time.

Keywords: decision tree, prediction, on-time graduation.

Abstrak: ***Tujuan:*** *Studi ini ditujukan untuk mengembangkan model yang akan memprediksi apakah seorang siswa akan lulus tepat waktu berdasarkan performa akademik mereka dalam mata kuliah pengantar komputasi.* ***Metode:*** *Proses data mining pendidikan digunakan dalam penelitian ini. Prosesnya dimulai dengan pengumpulan data pendidikan dan diakhiri dengan evaluasi model yang dikembangkan. Penelitian ini menggunakan decision tree algorithm.* ***Temuan:*** *Evaluasi model menghasilkan akurasi pengklasifikasian hingga 88,9% di mana jumlah total jawaban "Ya" (siswa yang lulus tepat waktu) adalah 52,49 yang diklasifikasikan dengan benar dan 3 salah diklasifikasikan sebagai "Tidak" dalam prediksi dan jumlah total jawaban "Tidak" (siswa yang tidak lulus tepat waktu) adalah 20,15 di antaranya diklasifikasikan dengan benar dan 5 salah diklasifikasikan dalam prediksi.* ***Kesimpulan:*** *Hasil penelitian dapat digunakan sebagai masukan dalam penyusunan bahan ajar baru dan perbaikan kurikulum yang akan membantu meningkatkan kinerja mahasiswa pada mata kuliah manajemen basis data. Model juga dapat digunakan sebagai alat untuk membantu mahasiswa lulus tepat waktu.*

Kata kunci: *decision tree, prediksi, lulus tepat waktu.*

To cite this article:

Co, J. & Casillano, N, F. (2021). Predicting On-time Graduation based on Student Performance in Core Introductory Computing Courses using Decision Tree Algorithm. *Jurnal Pendidikan Progresif*, *11*(3), 650-658. doi: 10.23960/jpp.v11.i3.202116.

INTRODUCTION

Higher education is an important and crucial component of human resource development, as it improves people's knowledge and abilities while also ensuring a country's economic growth. The major goal of a higher education institution, which is also one of its difficulties, is to deliver quality education to its students (Tampakas et al., 2018). Most students attend college with the hopes of laying the groundwork for a successful life or learning a skill that will help them land a decent career. A large number of students enroll in college each year, yet many of them fail or leave out in less than three years (Orion, Forosuelo & Cavalida, 2014). Understanding the paths students take to complete their degrees can assist instructors and administrators better serve student populations and help them achieve their educational objectives (Aiken et al., 2020). Universities must be concerned with the creation of effective mechanisms for monitoring students' progress and identifying crucial components of their performance in order to reach a higher level of education quality (Tampakas et al., 2018). One modality to monitor and predict salient educational outcomes is through data mining. Data mining is the process of extracting key patterns from a database, making it a useful tool for turning data into useful information. Marketing, finance, educational research, surveillance, telecommunications fraud detection, and scientific discovery are just a few of the domains where data mining can be used. Data mining, in particular, can be used to uncover buried information, relationships, and inform decision-making in a variety of fields. One of these fields is education, where the key aim is the evaluation and, as a result, improvement of educational organizations (Tekin, 2014; Han & Kamber, 2008).

Indeed, the development of educational database systems resulted in the creation of a significant number of educational databases, allowing data mining to extract important information from them (Alyahyan & Düºtegör, 2020). The staggering amount of data produced in educational institutions has led to the emergence of an independent research field called Educational Data Mining (EDM) (Liñan & Perez, 2015). Data mining has been used to predict a wide range of important educational outcomes, such as student success and performance (Martins, Miguéis, Fonseca, & Alves, 2019; Xing, 2019). The ability to accurately forecast students' future performance is seen as critical for carrying out appropriate pedagogical interventions in order to assure students' on-time and successful graduation. By analyzing students' development, relevant measures and strategic programs in an institution can be carefully designed in order to reduce students' graduation time and lessen student dropout (Tampakas et al., 2018). Xu, Moon & Van Der Schaar (2017) mentioned that Students may take a variety of courses, but not all of them are equally useful in forecasting future success. They further explained that it's critical to figure out the underlying correlation between courses in order to provide reliable performance forecasts. With this in mind, the researchers aimed to develop a model that will aim to predict on-time graduation based on student performance in core introductory computing courses using a predictive data mining technique called Decision Tree Algorithm. The research specifically aimed to [1] Develop a model on predicting the on-time graduation using Decision Tree Algorithm and [2] Measure the accuracy of the developed model using a confusion matrix and accuracy measurement.

METHODS

Research Design

This research followed the Educational Data Mining Process (Alyahyan & Düºtegör, 2020). The process contains the step-by-step procedure

to successfully develop a model (See Figure 1). Patterns and trends that go beyond simple analysis are discovered using this method. To segment the data and forecast future events, data mining employs powerful mathematical calculations and algorithms. Its concepts and methods can be used in a variety of fields, in this case, education. It contains the following steps: 1) data collection, 2) data initial preparation, 3) statistical analysis, 4) data preprocessing, 5) data mining implementation, and 6) result evaluation.

Data Collection, Preparation, and Pre-processing

Data will be collected from the College of Computer Studies student grade archives of BS Information Technology Students who graduated in the years 2017, 2018, 2019, and 2020. Grades from introductory computing courses will be harnessed and will be transformed to digital from (spreadsheet file) for processing and analysis. The initial collected data (also known as raw data) is frequently not ready for analysis

Figure 1. Educational data mining process (Alyahyan & Dütegör, 2020)

and modeling in its original form. Missing data, inconsistent data, erroneous data, miscoded data, and duplicate data can all be found in datasets created by combining information from several sources. This is why the raw data must go through some preliminary processing (Alyahyan& Düºtegör, 2020). After Collection and preparation, data goes through data cleaning where missing data, data noise and inconsistency are treated so ensure that the quality of prediction is not compromised. Although it should be noted that missing data can be handled by models like random forests and decision trees (Aleryani, Wang, De, & Iglesia, 2018).

Data Description

To convert the grades of the students to categorical data, the researchers utilized the equivalent adjectival rating of student grades as outlined in form ESSU-ACAD-712.b (Grade sheet form). The data that will be harnessed will be fed to a decision tree model using the following conventions:

Predictive Model

To Predictive and descriptive data mining models are often employed in EDM applications for success prediction (Kantardzic, 2003). This research will utilize the decision tree model. A decision tree is a basic algorithm that divides data into nodes based on the purity of the classes. It serves as a forerunner to Random Forest. Tree in Orange is a custom-built program that can handle both discrete and continuous information (Orange Documentation, 2015). Tree parameters:

a. Induce binary tree: build a binary tree (split into two child nodes)
b. Mininum number of instances in leaves: if checked, the algorithm will never construct a split which would put less than the specified number of training examples into any of the branches.
c. Do not split subsets smaller than: forbids the algorithm to split the nodes with less than the given number of instances.
d. Limit the maximal tree depth: limits the depth of the classification tree to the

Table 1. Data Description

Variable (Course)	Possible Values	Description
IntCom	1.0 (O) Outstanding 1.1-1.5 (E) Excellent 1.6 - 2.0 (VG) Very Good 2.1 - 2.5 (G) Good 2.6 - 3.0 (Fair) Fair 3.1 - 3.5 (Con) Conditional 3.6 - 5.0 (Failed) Failed INC (Inc) Incomplete Dr (Dr) Dropped WP (WP) Withdrawn with Permission IP (IP) In Progress	Grade in Introduction to Computing/IT Fundamentals
Prog1		Grade in Computer Programming 1
Prog2		Grade in Computer Programming 2
DiscM		Grade in Discrete Mathematics
DB1		Grade in Database management Systems 1/ Fundamentals of Database Systems
DB2		Grade in Database management Systems 2/ Advanced Database Systems
OTGS	Yes/no	On-Time Graduation Status

specified number of node levels.
*Stop when majority reaches [%]: stop splitting the nodes after a specified majority threshold is reached (Orange Documentation, 2015)

Model Evaluation

To evaluate the model, the Test and score widget on Orange will be utilized. The widget has two functions. It starts with a table that lists various classifier performance metrics, such as classification accuracy and area under the curve. Second, it generates assessment data that can be used by other widgets to analyze classifier performance, such as ROC Analysis and Confusion Matrix (Orange Documention, 2015).

Machine Learning Tool Used

This study utilized the Orange Data Mining Software to do the data preprocessing, predictive modelling and model evaluation. Orange is an open source machine learning and data visualization software. Build data analysis workflows visually, with a large, diverse toolbox (Demsar et al., 2013).

Ethical Consideration and Research Reflexivity

Conducting a research specifically a research that involves personal information from individuals must always follow ethical standards. The names of the students were transformed to code (stud1, stud2, stud3, stud4….) before preprocessing of their respective grades. Furthermore, only adjectival ratings (O, E, VG, G, F, and P) were used instead of their numeric grades. The sole intention of the researchers in conducting this research is only to develop a model on predicting the on-time graduation using Decision Tree Algorithm.

RESULT AND DISCUSSIONS

The Model

After applying the decision tree algorithm using the Orange Data Mining software, a decision tree for predicting on-time graduation was developed. Figure 2 shows the different possibilities of whether a student will graduate on-time or not based on their performance in their core introductory computing courses. The decision tree implies that students who fail or get a conditional grade in Database Management 2 will most likely fail to graduate on time. Furthermore, students who pass Database Management 2 but got a fair or failing grade in Database Management 1, Programming 1 and Programming 2 will also most likely fail to graduate on time. The figure also revealed that students who both pass in their DB1 and DB2 subjects will most likely graduate on-time. Suhaimi et al. (2019) mentioned that several researchers used grades or GPA as a way to analyze student performance and that it has a high influence in predicting students 'graduation on time. They explained further that, if students have a low GPA in their first and second years of study, they will most likely be unable to complete their studies within the study schedule. Once a student fails a subject that is a prerequisite of a future subject, the student has to wait for its offering thereby extending his study time. According to Nikula et al. (2011), more than 30% of computer science students worldwide discontinued or failed their initial computing courses in 1999. This staggering number can be attributed to several factors such as teaching and learning methodologies (Hoda & Andreae, 2014), the innate difficulty and complexity of computing courses (Nikula et al., 2011), the inability of the student to complete their computing tasks (Casillano, 2019) and student engagement and motivation (Corney et al., 2010). The study of Alturki (2016) revealed that students with high intrinsic motivation had an easier time finishing the course than other students; the issue arose with students who had low motivation, and it was discovered that these students were enrolled in the course haphazardly and faced few consequences if they failed.

The model that was developed confirmed the research of Maulana (2021) and Riyanto et. al (2019) that decision tree algorithm can be utilized to predict students on –time graduation. The development of a predictive model as mentioned by Suhaimi (2019) will benefit a variety of stakeholders, including the university's academic administration, academicians, and students, because it will alert them to students who are most likely to not graduate on time and what steps can be made to correct the problem. Furthermore, by reducing the number of students who are unable to graduate on-time, this technique can improve the university's academic quality. This study can be improved in the future by incorporating more datasets from various programs, notably from non-science and technology domains.

Model Evaluation Results

A five-fold cross validation using the test and score widget of orange data mining software was done to evaluate the model. The evaluation results, as shown in figure 3, revealed that the decision tree model has a classification accuracy (CA) of 88.9%. This percentage pertains to the proportion of the correctly classified data instances. The classification accuracy result is clearly outlined in figure 4 which is the confusion matrix for the model. The confusion matrix gives the number/proportion of instances between the predicted and actual class. In this case, the total number of actual "Yes" (students who graduated on-time) is 52, 49 were classified correctly and 3 were misclassified as "No" in the prediction. On

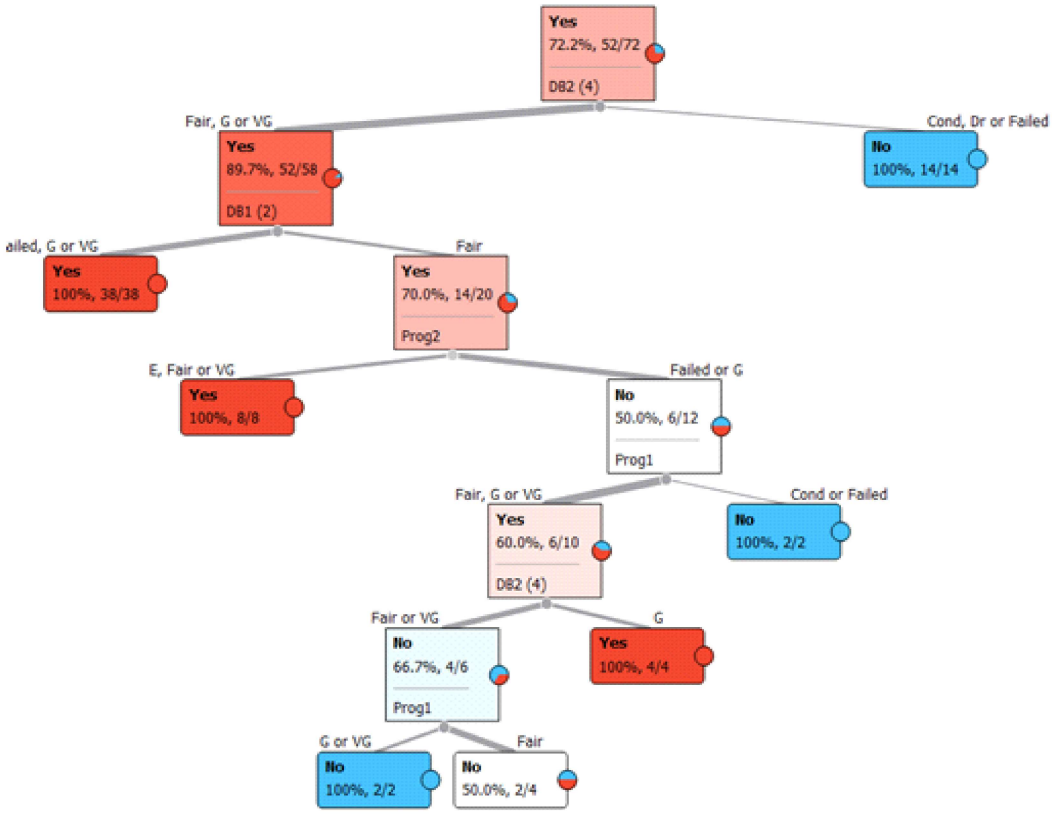

Figure 2. Decision Tree

the other hand, the total number of actual "No" (students who did not graduated on-time) is 20, 15 of which were classified correctly and 5 were misclassified in the prediction. The total number of correctly classified instances is 64 which is 88.9% of the total number of instances (72). The study of Riyanto et. al (2019) which aimed to determine the best algorithm to predict on time graduation resulted to an accuracy rate of 83.1% for the decision tree algorithm. Results of this study also confirmed the study of Suhaimi(2019) which revealed that decision tree was one of the algorithms that dominated accuracy results.

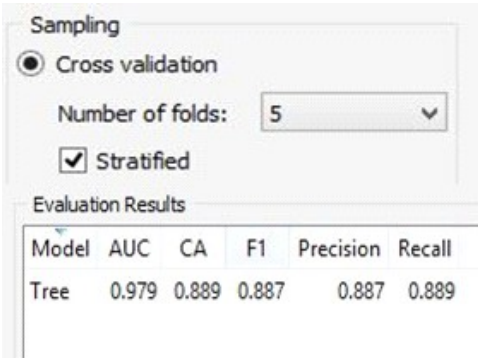

Figure 3. Classification accuracy results

		Predicted		
		No	Yes	Σ
Actual	No	15	5	20
	Yes	3	49	52
	Σ	18	54	72

Figure 4. Classification accuracy results

To further visualize the accuracy of the model, a receiver operating characteristic (ROC) curve was utilized. The ROC curve plots the true positive rate against a false positive rate of the model. Figures 4 and 5 shows the result of the ROC analysis of the model. As can be gleaned in the figures, both true positive rate of "Yes" and "No" results, represented with a green line, are above the 0.5 threshold, represented by a red dotted line. This implies that that the model has more true positive results versus its false positive counterpart and further implies that the model performed well.

■ CONCLUSIONS

This study primarily aimed at developing a predictive model for determining whether a student will graduate on time based on their

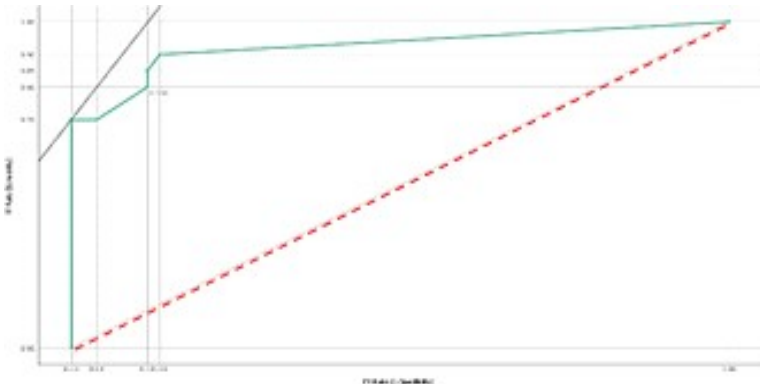

Figure 5. ROC curve for "No"

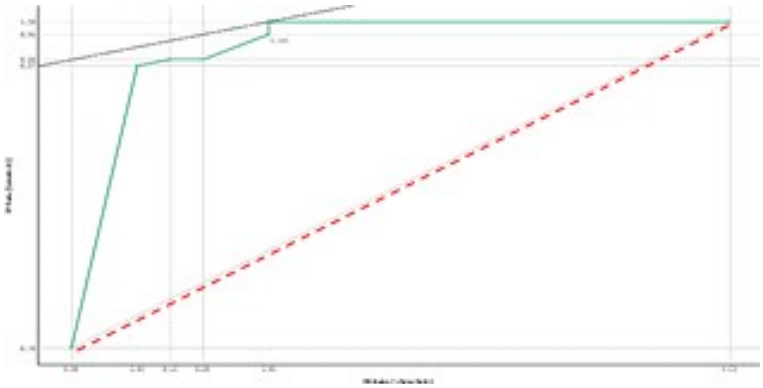

Figure 6. ROC curve for "Yes"

performance in their introductory computing courses. Decision tree algorithm was used to develop the model and it was evaluated through a 5-fold cross validation. The model that was developed revealed that the introductory computing course that greatly predicts the on-time graduation of the student is the Database Management 1 and Database Management 2 (DB2) course. The predictive model evaluation resulted to an 88.9% classification accuracy. This result can be used as inputs in the crafting of new materials that will help improve the performance of students in the database management course. The model can also be used as a tool to help students graduate on-time.

REFERENCES

Aiken, J. M., De Bin, R., Hjorth-Jensen, M., & Caballero, M. D. (2020). Predicting time to graduation at a large enrollment American university. Plos one, 15(11), e0242334.

Aleryani, A., Wang, W., De, B., & Iglesia, L. (2018). Dealing with missing data and uncertainty in the context of data mining. In International Conference on Hybrid Artificial Intelligence Systems.

Ali, R. (2020, September 23). Predictive Modeling: Types, Benefits, and Algorithms. Retrieved from Oracle Netsuite: https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml

Alturki, R. A. (2016). Measuring and improving student performance in an introductory programming course. Informatics in Education-An International Journal, 15(2), 183-204.

Alyahyan, E., & Dü°tegör, D. (2020). Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1), 3.

Casillano, N. F. (2019). Unraveling Views of Students Towards Computer Programming A Sentiment Analysis and Latent Semantic Indexing Application. IJRTE, 8(1) 543-456.

Chaurasia, P. (2020). CONFUSION MATRIX. Retrieved from MGCUB: http://www.mgcub.ac.in/pdf/material/20200429020322e5dac20f58.pdf

Corney, M., Teague, D., Thomas, R.N. (2010). Engaging students in programming. In: Twelfth Australasian Conference on Computing Education – Volume 103. 63–72.

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.

Gupta, P. (2017, May 18). Decision Trees in Machine Learning. Retrieved from Towards Data Science: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Han, J. Kamber, M. (2008). Data Mining: concepts and techniques. 2nd Edition, Morgan Kaufmann publishers.

Hand, D. J., & Adams, N. M. (\2014). Data mining. Wiley StatsRef: Statistics Reference Online, 1-7.

Hoda, R., Andreae, P. (2014). It's not them, It's us! Why computer science fails to impress many first years. In: 16th Australasian Computing Education Conference. 159–162.

Kantardzic, M. (2003). Data mining : concepts, models, methods, and algorithms. Wiley Interscience. Retrieved from https://

ieeexplore-ieeeorg.library.iau.edu.sa/book/5265979.

Liñán, L. C., & Pérez, Á. A. J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, 12(3), 98-112.

Martins, M. P. G., Miguéis, V. L., Fonseca, D. S. B., & Alves, A. (2019). A data mining approach for predicting academic success – A case study, (pp. 45–56). Cham: Springer.

Maulana, A. (2021, February). Prediction of student graduation accuracy using decision tree with application of genetic algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 1073, No. 1, p. 012055). IOP Publishing.

Nikula, U., Gotel, O., Kasurinen, J. (2011). A motivation guided holistic rehabilitation of the first programming course. ACM Transactions on Computing Education. DOI: 10.1145/2048931.2048935

Nurafifah, M. S., Abdul-Rahman, S., Mutalib, S., Hamid, N. H. A., & Ab Malik, A. M. (2019). Review on predicting students' graduation time using machine learning algorithms. International Journal of Modern Education and Computer Science, 11(7), 1.

Orange. (2015). Orange Visual Programming . Retrieved from orange3.readthedocs.io/:https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/

Orion, H. C., Forosuelo, E. J. D., & Cavalida, J. M. (2014). Factors affecting students' decision to drop out of school. *Slongan*, 2(1), 16-16.

Riyanto, V., Hamid, A., & Ridwansyah, R. (2019). Prediction of Student Graduation Time Using the Best Algorithm. Indonesian Journal of Artificial Intelligence and Data Mining, 2(1), 1-9.

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. Eurasian Journal of Educational Research, 54, 207-226.

Tampakas, V., Livieris, I. E., Pintelas, E., Karacapilidis, N., & Pintelas, P. (2018, June). Prediction of students' graduation time using a two-level classification algorithm. In International Conference on Technology and Innovation in Learning, Teaching and Education (pp. 553-565). Springer, Cham.

Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11(5), 742-753.

Xing, W. (2019). Exploring the influences of MOOC design features on student performance and persistence. Distance Education, 40(1), 98–113