International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I4P231

Volume 71 Issue 4, 354-358, April 2023

Original Article

Forecasting Graduation Schedule Model of Higher Education Learners using Feature Selection Techniques

Wongpanya S. Nuankaew[1], Tipparat Sittiwong[2], Sittichai Bussaman[3], Patchara Nasa-Ngium[4], Pratya Nuankaew[5]

[1,5]*School of Information and Communication Technology, University of Phayao, Phayao, Thailand*
[3,4]*Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand*
[2]*Faculty of Education, Naresuan University, Phitsanulok, Thailand*

[5]*Corresponding Author : pratya.nu@up.ac.th*

Received: 02 March 2023 Revised: 19 April 2023 Accepted: 21 April 2023 Published: 25 April 2023

Abstract *- The biggest concern for learners in the 21st century is that graduation does not fit their educational plan. This research aims to study factors affecting students' academic achievement in the Faculty of Education during the COVID-19 crisis in Thailand. The data was on the academic achievement of 90 students from the Bachelor of Art Program in Educational Technology and Communications at the Faculty of Education, Naresuan University, Phitsanulok, Thailand. The research process was analyzed using data mining techniques, including CRISP-DM procedures, decision tree algorithm, forward selection analysis, cross-validation techniques, and confusion matrix performance. This research found that the course 001211 Fundamental English was the most significant subject for delayed graduation, where the developed model has a very high level of accuracy (89.00%). Researchers can use such a model to create effective planning strategies for preventing graduation failures.*

Keywords *- Academic achievement, Educational data mining, Students dropout, Student model.*

1. Introduction

Higher education is an advanced education that is integral to the development of learners in both developing and developed countries. This is because higher education institutions guide young people to experience the current social situation and understand the problem around them to cope with and deal with various knowledge. Therefore, the quality of educational programs and institutes is important in promoting learners' quality. In this regard, every bachelor's degree program in Thailand must be certified by an educational institution named “CHE Curriculum Online: Thailand”.

Naresuan University was established on July 29, 1990, and has 22 organizations under management. The Faculty of Education of Naresuan University was founded together with the University, which has a total period of 33 years. They have produced quality graduates in more than 20 academic programs in the past.

However, during the year 2019-2021, the Faculty of Education was affected by the COVID-19 epidemic, which directly affected the teaching and learning process in the institution. It inevitably had a latent effect on students. Moreover, the research team found that the student's behaviours and the length of time they completed the educational program had changed. Learners are more trends and chances to delay graduating later than usual. It is, therefore, the root of the problem that led to this research. This research has two main objectives. The first objective is to study the overview of the graduation situation of undergraduate students of the Faculty of Education, Naresuan University, during the period affected by the COVID-19 epidemic. The second objective is to develop a model and identify significant factors for the delayed graduation of students in the Faculty of Education.

Naresuan University has authorized the data collected through the research ethics process. The data consisted of academic reports of 90 undergraduate students from the Bachelor of Art Program in Educational Technology and Communications at the Faculty of Education, Naresuan University, Phitsanulok, Thailand. The research tool consists of three parts. The first part is fundamental data analysis. It was a preliminary statistical analysis. The second part is the application of machine learning tools consisting of decision tree algorithms and feature selection by forward selection techniques. The third section aimed to test the model's performance, consisting of the cross-validation methods and the confusion matrix performance.

The expectations and goals of this research were to study and design learning strategies for learners in the 21st century living in the digital age by adjusting the context of the

organization and the curriculum following the learning behaviours of the learners to learn and develop the potential of learners sustainably.

2. Literature Reviews

One of the educators' main goals is to develop learners' quality in all dimensions so that they can bring out their potential in future careers. Many educators try to study and develop learning styles consistent with student behaviours: Multiple Intelligences theory [1], Bloom's Taxonomy theory [2], The Schema and Constructivism theory [3], and so on.

However, nowadays, the context and environment of the learning society are suddenly changing, creating a learning style that is more consistent with the context of the learners. Educational theories that have emerged over the past decade include distance learning styles [4], [5], problem-based learning styles [6], self-regulated learning styles [7], blended learning styles [8], [9], hybrid learning styles [10], [11], and so on.

In addition, several data scientists have good intentions to improve the quality of students, and organizations related to education have applied artificial intelligence and machine learning technology to enhance the education industry. Many researchers referenced and are interested include matching learners and instructors with different contexts [12], recommending educational programs suitable for learners [13], students' dropout crisis in higher education [14], [15], and so on.

Other contributions include the reviews of the use of analytical tools for machine learning techniques for educational technology [16], [17], the reviews of researchers involved in educational data mining [18]–[21], the reviews of a replacement for the technology to be used in current, and so on. From the areas of interest and research that researchers have already carried out have pushed and created awareness for researchers to conduct this research. The researcher has a great expectation that this research will continue to drive and promote the education industry.

3. Research Methodology

The research methodology in this research was meticulously designed and based on academic principles, which consisted of four main parts: The first part is to define the population and sample. The second part is model development with data mining techniques. The third part is model performance testing, and the fourth part is output design and interpretation.

3.1. Population and Sample

The population was students from the Bachelor of Art Program in Educational Technology and Communications at the Faculty of Education, Naresuan University, Phitsanulok, Thailand.

The research sample consisted of 90 students from the Bachelor of Art Program in Educational Technology and Communications at the Faculty of Education, Naresuan University, Phitsanulok, Thailand, who obtained academic results from their enrolment results during the academic year 2019-2021. Table 1 presents the sample data classified by curriculum and academic year.

Table 1 found that 80 students completed the program as scheduled, representing 88.89%, and ten graduated as delayed, representing 11.11%.

3.2. Modelling

Model development at this stage comprises the two method's essential machine learning tools. The first part is to develop a decision tree model. The decision tree algorithm is used that it is a machine learning technique that is easy to understand and can be used in practice [16], [22], [27]. The functional structure of the decision tree algorithm is the upside down of the actual tree structure.

The internal structure consists of nodes and branches. Nodes act as variables or attributes describing the structure of the decision tree model, with the essential node being called the "root node" at the top of the model and the latest node, called the "leaf node", serving as the model's answer. Simultaneously, the branches serve as options and alternative conditions that connect to the next node.

The second tool, called "Forward Selection" [23], [24], is an optimization method for selecting variables and optimizing the model. The principle of forward selection method is a method of selecting independent variables into the equation one by one in order of correlation. In addition, the forward selection method is a selection of features by modelling, where a classification model is derived from a defined set of features. It measures the model's performance and selects the set of features that make the model most efficient, such as the one that gives the most accuracy.

In this research, the two techniques were combined to develop the most efficient model by referencing and selecting the model with the highest accuracy.

Table 1. Data Collection

Academic Year	Number of Graduates		
	Scheduled	Delayed	Total
2019	23 (25.56%)	4 (4.44%)	27 (30.00%)
2020	30 (33.33%)	5 (5.56%)	35 (38.89%)
2021	27 (30.00%)	1 (1.11%)	28 (31.11%)
Total	80 (88.89%)	10 (11.11%)	90 (100%)

3.3. Model Performance

Two essential techniques were selected to test and determine the model's efficiency for this research.

The first technique divides the data for model development and model testing. This research divided the data into two parts according to the principle of cross-validation methods [26], [28]. The first piece of data is called the training dataset, which is used to develop the model. The second piece of data left over from the first is used to test the resulting model, called testing the dataset. This method is accepted because it uses actual data to test the model itself.

The second technique to test model performance is a tool to indicate model performance, and it consists of three parts: accuracy, precision, and recall. The accuracy value was used to determine the model's validity, divided by the total number of data predicted correctly by the model. Precision measures how accurately the model can predict each response class. The final metric, recall, was used to determine the validity of the model responses classified by class.

Both tools were used to find the model with the highest accuracy for selection as a prototype.

3.4. Research Results and Interpretation

Interpretation of model results and findings, researchers used the best model results based on the model with the highest accuracy to discuss results classified by disciplines and academic year. The factors to be derived from this study are nodes that emerge from each model, with the researcher proceeding to summarize and discuss the factors involved and occurring in all models.

4. Research Results

4.1. Model Results

The model results show the result of selecting the model with the highest accuracy, as shown in Table 2 and Table 3.

Table 2 shows the four-criterion analysis of the decision tree model. The model with the highest accuracy is the 50-Fold cross-validation test data model, which uses the information_gain criterion with an accuracy of 86.67%. The detailed model performance results are presented in Table 4.

Table 3 shows the results of the analysis of model optimization by forward selection technique, given the parameters of the maximal number of attributes equal to 10 and stopping behaviour without signification increase with alpha equal to 0.05. It found that the model with the highest accuracy is the 50-Fold cross-validation test data model and uses the information_gain criterion with an accuracy of 89.00%. The detailed model performance results are presented in Table 5.

Table 2. Model Results from Decision Tree

Criterion	Cross-validation / Accuracy		
	10-Fold	50-Fold	Leave-one-out
gain_ratio	84.44%	85.56%	83.33%
information_gain	83.33%	86.67%*	85.56%
gini_index	83.33%	80.00%	80.00%
accuracy	83.33%	84.44%	84.44%

Table 3. Model Results from Forward Selection Analysis

Criterion	Cross-validation / Accuracy		
	10-Fold	50-Fold	Leave-one-out
gain_ratio	84.44%	85.56%	83.33%
information_gain	88.89%	89.00%*	88.89%
gini_index	83.33%	80.00%	80.00%
accuracy	83.33%	84.44%	84.44%

Table 4. Model Performance from Decision Tree

Accuracy: 86.67%	Actual		Class Precision
	True Scheduled	True Delayed	
Pred. Scheduled	77	9	89.53%
Pred. Delayed	3	1	25.00%
Class Recall	96.25%	10.00%	

Table 5. Model Performance from Forward Selection Analysis

Accuracy: 89.00%	Actual		Class Precision
	True Scheduled	True Delayed	
Pred. Scheduled	80	10	88.89%
Pred. Delayed	0	0	0.00%
Class Recall	100.00%	0.00%	

4.2. Model Performance

The model performances show the result of selecting the model with the highest accuracy, as shown in Table 4 and Table 5.

Table 4 shows the model efficiency from Table 2. Overall, the model predicted accuracy and performance at a high level, with accuracy equal to 86.67%. However, the researchers believe that more high-accuracy models can be developed, as the findings in the analysis are shown in Table 3 and Table 5.

Table 5 shows the analysis of the model performance by using the forward selection technique. It found that the model's accuracy was increased and thus discovered the most significant factor in meeting deadlines and delays: Course 001211 Fundamental English.

5. Research Discussion

The results of the research were achieved under all research objectives. This section of the discussion discusses the goals that have been established.

From the first objective, the researchers found that learners in the Bachelor of Art Program in Educational Technology and Communications at the Faculty of Education, Naresuan University, Phitsanulok, Thailand, did not suffer from delayed completion, as shown in the data gathered in Table I. It, therefore, concludes initially that the curriculum has effective management of learners. However, there was an epidemic situation of COVID-19.

For the second objective, the researchers developed an artificial intelligence and machine learning model, which found that the developed model had a very high level of accuracy and discovered factors that affected the delay in completion of students. The obtained model has an accuracy value of 89.00%, as shown in Table III and Table V. The factor that affects the delayed completion is the course 001211 Fundamental English. These findings are in line with other research showing that learners in Thailand still lack a crucial skill, namely English language skills, a key factor for 21st-century learners.

The suggestion of this research is that when discovering the causes and what might affect learning success, that is, English language skills. Executives and related persons at all levels should continue to raise awareness and develop English language skills for learners in Thailand.

6. Conclusion

This research is concerned with students during the COVID-19 pandemic situation that may directly affect students, and two important research objectives were established. The first objective is to study the overview of the graduation situation of undergraduate students of the Faculty of Education, Naresuan University, during the period affected by the COVID-19 epidemic. The second objective is to develop a model and identify significant factors for the delayed graduation of students in the Faculty of Education. This research has achieved both objectives. The researchers knew that the learners in the curriculum were slightly affected.

However, in the second objective, the researchers discovered that learners were still missing critical skills for 21st-century learning. It is an essential English language skill for sustainable learning and student development.

In addition, the results of this research use data between 2019-2021, which the Faculty of Education, Naresuan University can compare with the normal situation that is not affected by COVID-19 for a comprehensive study. Lastly, researchers have great expectations that this research will create sustainable value for the Thai educational industry.

Funding Statement

This research project was supported by the Thailand Science Research and Innovation Fund and the University of Phayao (Grant No. FF66-UoE002)

Acknowledgments

This research was supported by many advisors, academics, researchers, students, and academic staff from three organizations: the Faculty of Education at the Naresuan University, Phitsanulok, 65000 Thailand; Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, 44000 Thailand, and the School of Information and Communication Technology at the University of Phayao, Phayao, 56000 Thailand. The authors would like to thank all of them for their support and collaboration in making this research possible.

References

[1] Howard Gardner, and Seana Moran, "The Science of Multiple Intelligences Theory: A Response to Lynn Waterhouse," *Educational Psychologist*, vol. 41, no. 4, pp. 227–232, 2006. [CrossRef] [Google Scholar] [Publisher link]

[2] David R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory into Practice*, vol. 41, no. 4, pp. 212–218, 2002. [CrossRef] [Google Scholar] [Publisher link]

[3] I.L. McCann, and L.A. Pearlman, "Constructivist Self-Development Theory: A Theoretical Framework for Assessing and Treating Traumatized College Students," *Journal of American College Health*, vol. 40, no. 4, pp. 189–196, 1992. [CrossRef] [Google Scholar] [Publisher link]

[4] Sawsan Abuhammad, "Barriers to Distance Learning during the COVID-19 Outbreak: A Qualitative Review from Parents' Perspective," *Heliyon*, vol. 6, no. 11, p. e05482, 2020. [CrossRef] [Google Scholar] [Publisher link]

[5] Roberto D. Costa et al., "The Theory of Learning Styles Applied to Distance Learning," *Cognitive Systems Research*, vol. 64, pp. 134–145, 2020. [CrossRef] [Google Scholar] [Publisher link]

[6] Saiful Amin et al., "The Effect of Problem-Based Hybrid Learning (PBHL) Models on Spatial Thinking Ability and Geography Learning Outcomes," *International Journal of Emerging Technologies in Learning*, vol. 15, no. 19, 2020. [CrossRef] [Google Scholar] [Publisher link]

[7] J. Broadbent, and W. L. Poon, "Self-regulated Learning Strategies & Academic Achievement in Online Higher Education Learning Environments: A Systematic Review," *The Internet and Higher Education*, vol. 27, pp. 1–13, 2015. [CrossRef] [Google Scholar] [Publisher link]

[8] Sura I. Alayed, and Ahmad Adnan Al-Tit, "Factors Affecting the Adoption of Blended Learning Strategy," *International Journal of Data and Network Science*, vol. 5, no. 3, pp. 267–274, 2021. [CrossRef] [Google Scholar] [Publisher link]

[9] Wirachat Banyen, Chantana Viriyavejakul, and Thanin Ratanaolarn, "A Blended Learning Model for Learning Achievement Enhancement of Thai Undergraduate Students," *International Journal of Emerging Technologies in Learning*, vol. 11, no. 4, 2016. [CrossRef] [Google Scholar] [Publisher link]

[10] Agoritsa Konstanti, and Antonia Moropoulou, "Hybrid Educational Methodology for the Cognitive Domain of Built Heritage Protection Interconnecting Secondary with Tertiary Level Education," *International Journal of Engineering Pedagogy*, vol. 3, no. 4, 2013. [CrossRef] [Google Scholar] [Publisher link]

[11] Ariel Siegelman, Blended, Hybrid, and Flipped Courses: What's the difference?, Center of the Advancement of Teaching, 2019. [Online]. Available: https://teaching.temple.edu/edvice-exchange/2019/11/blended-hybrid-and-flipped-courses-what%E2%80%99s-difference

[12] Jaclyn Broadbent, "Comparing Online and Blended Learner's Self-regulated Learning Strategies and Academic Performance," *The Internet and Higher Education*, vol. 33, pp. 24–32, 2017. [CrossRef] [Google Scholar] [Publisher link]

[13] Daniel Gallego et al., "A Model for Generating Proactive Context-Aware Recommendations in e-Learning Systems," *2012 Frontiers in Education Conference Proceedings*, 2012. [CrossRef] [Google Scholar] [Publisher link]

[14] Catarina Felix de Oliveira et al., "How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review," *Big Data Cognitive Computing*, vol. 5, no. 4, 2021. [CrossRef] [Google Scholar] [Publisher link]

[15] Marzieh Karimi-Haghighi, Carlos Castillo, and Davinia Hernández-Leo, "A Causal Inference Study on the Effects of First Year Workload on the Dropout Rate of Undergraduates," *Artificial Intelligence in Education*, pp. 15–27, 2022. [CrossRef] [Google Scholar] [Publisher link]

[16] Chitra Jalota, and Rashmi Agrawal, "Analysis of Educational Data Mining using Classification," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019. [CrossRef] [Google Scholar] [Publisher link]

[17] Pratya Nuankaew, Tipparat Sittiwong, and Wongpanya Sararat Nuankaew, "Characterization Clustering of Educational Technologists Achievement in Higher Education Using Machine Learning Analysis," *International Journal of Information and Education Technology*, vol. 12, no. 9, pp. 881–887, 2022. [CrossRef] [Google Scholar] [Publisher link]

[18] Hanan Aldowah, Hosam Al-Samarraie, and Wan Mohamad Fauzy, "Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and Synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019. [CrossRef] [Google Scholar] [Publisher link]

[19] Ryan S. Baker, Taylor Martin, and Lisa M. Rossi, "Educational Data Mining and Learning Analytics," *The Wiley Handbook of Cognition and Assessment*, 2016. [CrossRef] [Publisher link]

[20] Pranav Dabhade et al., "Educational Data Mining for Predicting Students' Academic Performance Using Machine Learning Algorithms," *Materials Today Proceedings*, vol. 47, no. 15, pp. 5260–5267, 2021. [CrossRef] [Google Scholar] [Publisher link]

[21] Fuseini Inusah et al., "Data Mining and Visualisation of Basic Educational Resources for Quality Education," *International Journal of Engineering Trends and Technology*, vol. 70, no. 12, pp. 296-307, 2022. [CrossRef] [Google Scholar] [Publisher link]

[22] Mai Kiguchi, Waddah Saeed, and Imran Medi, "Churn Prediction in Digital Game-based Learning Using Data Mining Techniques: Logistic Regression, Decision Tree, and Random Forest," *Applied Soft Computing*, vol. 118, p. 108491, 2022. [CrossRef] [Google Scholar] [Publisher link]

[23] Gang Chen, and Jin Chen, "A Novel Wrapper Method for Feature Selection and its Applications," *Neurocomputing*, vol. 159, pp. 219–226, 2015. [CrossRef] [Google Scholar] [Publisher link]

[24] Martin Gutlein et al., "Large-Scale Attribute Selection Using Wrappers," *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 2009. [CrossRef] [Google Scholar] [Publisher link]

[25] Vandana Mulye, and Atul Newase, "A Review: Recruitment Prediction Analysis of Undergraduate Engineering Students Using Data Mining Techniques," *SSRG International Journal of Computer Science and Engineering*, vol. 8, no. 3, pp. 1-6, 2021. [CrossRef] [Publisher link]

[26] Cullen Schaffer, "Selecting a Classification Method by Cross-Validation," *Machine Learning*, vol. 13, pp. 135–143, 1993. [CrossRef] [Google Scholar] [Publisher link]

[27] Raza Hasan et al., "Student Academic Performance Prediction by using Decision Tree Algorithm," *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018. [CrossRef] [Google Scholar] [Publisher link]

[28] Christina G. Skarpathiotaki, and Konstantinos E. Psannis, "Cross-Industry Process Standardization for Text Analytics," *Big Data Research*, vol. 27, p. 100274, 2022. [CrossRef] [Google Scholar] [Publisher link]