Behavioral Analytics of a Freshmen Small Private Online English Course

Kingkan Luenpan[1], Sotarat Thammaboosadee[2*] and Rojjalak Chuckpaiwong[3]

[1,2*,3]Faculty of Engineering, Mahidol University, Thailand

(Received: August 23, 2022; Revised: October 6, 2022; Accepted: October 10, 2022)

Abstract

SPOC (Small Private Online Course) is an online learning platform combining classroom and online lessons. For the development of the SPOC, it is worth looking at learning behaviors in-depth throughout the course. As Higher Education courses in Thailand focus on English language benchmarking, this research decided to choose English Level 1 as a content course for the reading and grammar skills from August 2020 to June 2021. The descriptive analytics study conducted by Visual Analytics and K-Means Clustering, both student-level and lesson-level learning behavior. The result found four types of learners in Student-level learning behavior: 1) Intelligent, 2) Weak-cognitive, 3) Inattentive 4) Unenthusiastic. Moreover, Predictive Analytics, for predicting learning quality through learning behaviors of each cluster by four machine learning models, were comparatively experimented with: Generalized Linear Model, Decision Tree, Random Forest, and Gradient Boosted Trees. The Optimization method is used for tuning the optimum parameter of each method. For student-level behavior prediction, the Unenthusiastic Decision Tree was 0.0449, and Lesson-level Weak-cognitive Gradient Boosted Trees was 0.0371 relative error. Additionally, the factor of importance in quality prediction was found that amount of quizzes was the essential variable among all clusters. The result of this research is that the instructors can further develop content and teaching methods in the course to truly meet the learners' needs.

Keywords: 1) SPOC 2) Online Learning 3) Descriptive Analytics 4) Predictive Analytics

[1] Student, Faculty of Engineering Mahidol University; E-mail: kingkan.lun@student.mahidol.ac.th

[2*] Lecturer, Faculty of Engineering Mahidol University; E-mail: sotarat.tha@mahidol.ac.th (Corresponding Author)

[3] Lecturer, Faculty of Engineering Mahidol University; E-mail: rojjalak.chu@mahidol.ac.th

Introduction

Education is an enormous influence on the lives of learners. In particular, higher education drives the country (Ministry of Higher Education, Science, Research, and Innovation, 2021, pp. 33-35). Technology has played a considerable role in creating education courses (Maslin, et al., 2010, pp. 33-35). Higher education institutions' SPOC is an online lesson that creates a novelty for the education industry (Lu, 2018, pp. 157-169) by simulating virtual classroom conditions (Prates, Garcia and Maldonado, 2019, pp. 129-138). SPOC is becoming increasingly popular with learners as its speeds up accessing lesson content and interaction between learners and instructors.

The researcher considered that Thailand had focused on learning English because of the English language benchmark for the institute. Therefore, the students must pass the English test to graduate since the English Level 1 course is mandatory for first-year students studying Thai language programs. It contains reading and grammar skill tests. Both are essential learning skills because students need reading skills to analyze the articles in the lesson and grammar skills to analyze and synthesize the content of the study or exams. Both skills could lead to academic writing skills (Patterson and Duer, 2006, pp. 81-87).

Research motivation thought English that important in the education age based on the English language proficiency criteria. In addition, the student's problems with the teaching content are expressed through learning behaviors. Strategies should be adjusted by content and courses developed to the real needs of the learners. The research objective is to analyze the relationship Descriptive from the grouping of the behavior of the learners by segmentation in both individual (student-level) and content (lesson-level). Moreover, the resulting Predictive model will be used for a content course in the developer's system. And the strategic deployment looks at problems in each type of learner to close the flaws leading to the enhancement of the potential of both learners and instructors for maximum efficiency. The main contribution is the multi-dimensional analysis of student behavior in the foundation English course, consisting of student and learning behavior levels. The result of this research should motivate the full-stream analytics value chain using the combination of clustering techniques and predictive modeling and strategically deployed in the education processes.

Literature Review

Nguyen, Chen and Saravanarajan (2022, pp. 1-5) proposed the most recent related research. Their motivation is that the pandemic situation opened up the understanding of students' behavior and evaluating their performance in an e-learning environment. The data were collected from a Brazilian University and secondary education at two Portuguese schools. The Feed-forward Neural Network, Perceptron, and Self-organizing Algorithms are used to predict students' behavior. The finding shows that the highest accuracy of the Perceptron method is 80 percent. Examining students' behavior is based on reactions from the assessment of learning outcomes and the

usage of in-classroom social features.

Suleiman and Anane (2022, pp. 1480-1485) investigated the institutional data of a Nigerian university and the predictive impact on student performance, measured in terms of cumulative grade point average (CGPA). The results showed that age and marital status have no significant relationship with final CGPA. The gender and pre-entry score have a weak relationship but are not good predictors. The application of four machine learning algorithms, Linear Regression, Support Vector Regression, Decision Tree, and Random Forests, indicate that the third year CGPA is a good predictor of final year CGPA. Support vector regression has the best performance in predicting the final CGPA.

Online Learning Behavioral Data was proposed for research on Clustering Mining and Feature analysis (Zhang, Zhang and Ran, 2018, pp. 629-634). This work created a model to find a relationship between behavior and academic performance in SPOC, which consists of the duration of each class, the number of admissions, the quality of the study, the status of the study, quality inspection of teaching, and peer review. These elements can be indicators of student performance which is divided into two levels: 1) Low-dimension is the indicator of learner participation in learning, such as duration of learning and number of classes, and 2) High-dimension is the indicator of how knowledge is measured and responded to online learning, such as assessments by learners and instructors in measuring their behavior. In this research, existing cluster data should be used to predict learner behavior using Predictive Analytics to benefit learners' future content development.

Han, et al., (2017, pp. 1-7) proposed a paper supporting quality teaching by educational data mining. The research focuses on learning behaviors to predict the essential qualifications in learning and quality assessments to improve the quality of teaching on OpenEdx by entering a predictive model. The Gradient Boosting Decision Tree (GBDT) creates decision trees by randomly selecting a variety of replicas to get the best model. The results are evaluated from the predicted model learner's performance based on the weekly variable and the ROC-AUC value. Students are more likely more enthusiastic about handling problems than during the first problem period in class, which is approximately 0.18.

Additionally, for variables that have an essential effect on model prediction, the maximum value of a problem completed before 96 hours at approximately 0.38. In research, applications showing two or more correlations will increase results in other dimensions. Moreover, analyze which variables are essential for prediction. The weakness of this research is the use of algorithms only for the Gradient Boosting Decision Tree (GBDT). More algorithms will produce results with a variety of dimensions. Due to the diversity of data, there may be more than one algorithm to use appropriate algorithms for the best results by finding from Evolutionary optimization.

Xu, et al. (2017, pp. 11-13) proposed research to predict the final cumulative grade point average in a 3-year course at UCLA using four machine learning algorithms: k-Nearest

Neighbors, linear regression, logistic regression, and random forest, the mean square errors in the ensemble predictor appeared that random forest had the best performance in MAE 182A at 0.45 and EE110L at 0.29. This research should build on the results of the predictions to be applied as a strategy and a guideline for teachers to develop knowledge for learners to achieve maximum efficiency. This research can serve as a teaching guide to assess performance and recommend courses to students. Additionally, design other education policies.

Yanta, et al. (2021, pp. 567-573) proposed research to predict course performance in the most effective adjustment of admission criteria and strategies of undergraduate students by using algorithms to predict performance. The performance evaluation methods are the relative error. The root means square error and the absolute error. Three experimented models are Generalized Linear, Gradient Boosted Tree, and Deep Learning. The best model is the Gradient booted tree had the lowest relative error of 0-0.4%. Moreover, prescriptive analytics is used to optimize the number of students admitted in each round, and recruitment using evolutionary optimization must prepare a strategic plan. This research should design a strategic plan to improve curriculum quality and close learners' gaps.

According to the above-related works, the researcher found an opportunity to create quality research to solve problems by learning from data preparation. Both descriptive and predictive analytics meet the goals that have been set. Notably, the conducted research must benefit the people involved in a better direction.

Methods

The research methodology of the Descriptive and Predictive Analytics Process is shown in Figure 1.

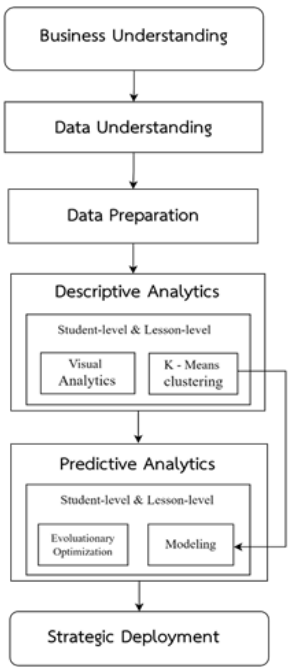

Figure 1 Research methodology

1. Business Understanding: The two main problems with learners in the SPOC system:

1.1 The results score criteria is greater than or equal to 0.5. (total score 1.0) is considered passing status. For learners who do not pass (lower than 0.5) the course, depending on the main elements: homework and quizzes in lessons and the number of times learners can know the responsibility and set of study subjects.

1.2 The status of the non-finished course in the system will show up as learning. The learner has not finished because of the learner's behavior. Not attending school short learning period, and the test scores in the lessons were lower than the criteria (lower than 0.5).

2. Data Understanding: Data on the SPOC behavior from the Level-1 English course in University's database was collected. The collected data include academic statistics for course learners, including Time-in, Timeout, Grade, Quiz, and Number of attendance. The researcher collected data on English learning behavior Level 1 from August 2020 to June 2021. Total 2,120 students and 25,123 learning transactions.

3. Data Preparation: The raw data have been processed for analysis in the following steps:

3.1 Data transformation: Data is transformed into a suitable format and imported into the model. This step transforms the Time-in and Timeout variables from date to time and displays each class's Time-in and Timeout. Transform the display data of the Duration variable from the "Hour" unit to the "second" unit to calculate the duration of the study covering all student data.

3.2 Missing value handling: data with the method of replacing the missing values. Eliminate the missing value in the Duration variable by finding the difference between the lower Time-in variable - Upper Time-in will result in the duration of the learning value.

4. Descriptive Analytics

4.1 Relationships in variables are displayed through visual analytics results

Figure 2(a) shows the relationship of learners' status by applying the status variables 'Passed' and 'Failed' to the number of students in the course. Since there are about 2,000 people who have passed, it shows that students have a great intention to study, representing 89.53%. It intends to show the learners' quality in the course that the overall quality of learners is excellent.

Figure 2(b) shows the relationship between grades' status and learning duration. To demonstrate the duration of attendance to the effect of 'Pass' and 'Fail,' it can be seen that most learners in 'Pass' status study duration (hours) are approximately 3 -28 hours to a maximum of 64 hours, and for most learners in the 'Failing' state, the learning duration (hours) is about 15 - 47 hours to a maximum of 94 hours, can be seen that studying 1 hour of lessons is still not possible to pass this course.

Figure 2(c) shows the relationship between academic performance and the number of learners to demonstrate the aggregated scores by intervals. It shows that most learners who received grades of about 0.60 - 0.64 and 0.94 - 0.99 considered passing.

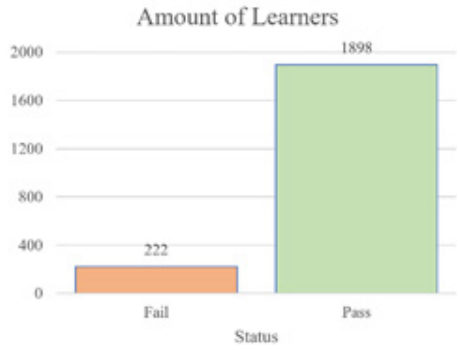

Figure 2(a) Visual analytics of learner status

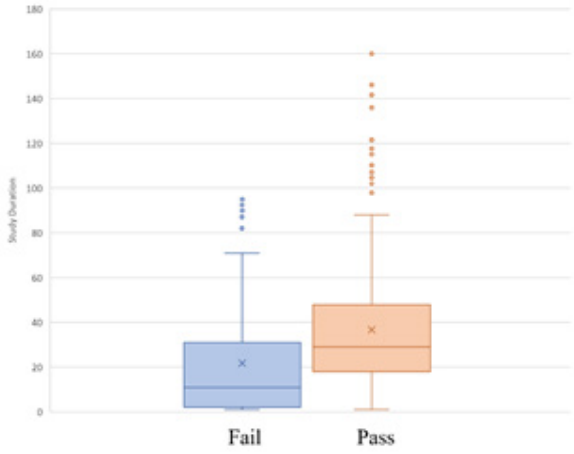

Figure 2(b) Visual analytics of learner status and durations

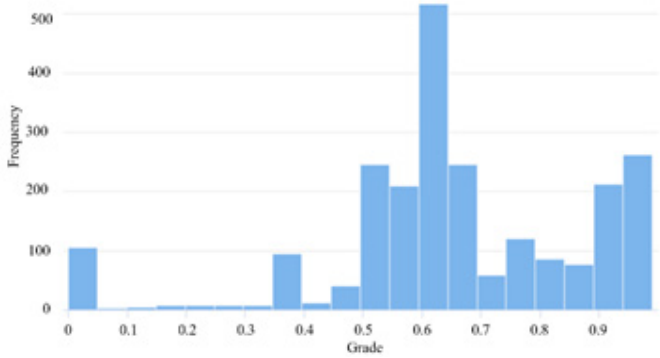

Figure 2(c) Visual analytics of learners' grades

The visualization discovers the correlation between various variables, including the number of passers and fails, grade ranges, and learning durations of a pass and fail status, to show the learners' quality in the course and the initial relationship affecting variables leading to the next level of data analysis. It categorizes the learners in the course to find the relationship of various variables to delve more deeply.

4.2 K-means clustering

Cluster Analysis (Omran, Engelbrecht and Salman, 2007, pp. 583-605) groups data with multiple dimensions by measuring similar data sets (homogeneous) properties and distances in the same cluster. The K-means Algorithm (Hamerly, 2003, pp. 8-9) segmented K centroids to find the best K value to group data sets. Running the data until the best segmentation is obtained increases the data

density in the group. It also increases the distance between the data groups to be grouped according to similar properties.

In order to perform the clustering properly, the attendance and quiz data have to be joined and preprocessed, as shown in Figure 3.

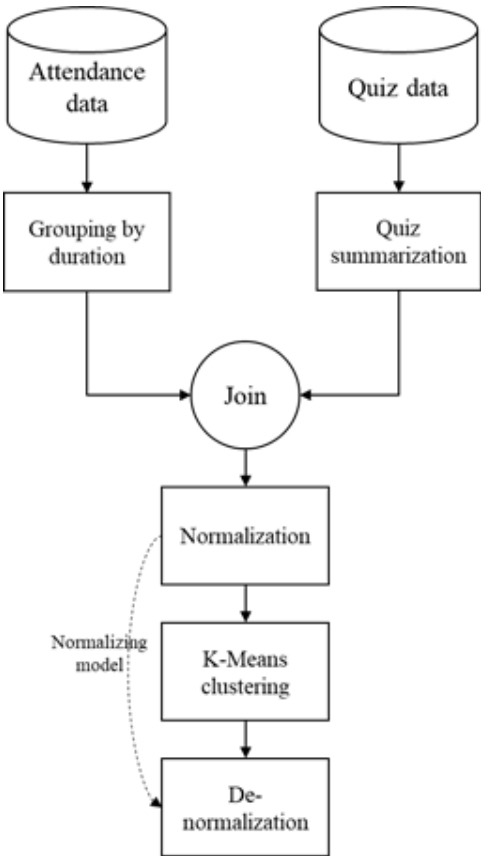

Figure 3 Data preprocessing for clustering

Data input: The attendance transaction and quiz records are the input of this process. The attributes of both data are already described in the Data Understanding section.

Grouping by duration: This step groups the learning duration by the summation and sequence variables number of attendances by counting function.

Quiz summarization: This step finds the number of times students have completed all quizzes.

Join: This step links two tables using id as joined keys by the inner join method.

Normalization: This step normalizes the duration by the z-transformation method (Patro and Sahu, 2015, pp. 20-22). It scales the data to an equal range using mean ad standard deviation. The final z values are approximately in the range of -3 to 3. This process benefits K-Means clustering since it is needed to compare and compute variables' differences.

K-Means Clustering: This step optimizes K to achieve the best K-Means clustering result. The evaluation of the model was measured using the Sum Square Error (SSE).

De-Normalization: This step changes the attribute's value to the original scale before normalization for the clustering results explanation advantage.

5. Predictive Analytics

Predictive analytics (Kumar and Garg, 2018, pp. 31-37) aims to predict the possible data trends or behaviors based on historical data leading to the predictive modeling process from requirements gathering, deployment, and data quality checking, and modeling by Machine Learning algorithms. In this research, the input data are the same as in the previous clustering process, enhanced by adding the student-level and lesson-level segmentation results. In this paper, four Machine Learning methods are selected for the experiments:

5.1 Generalized Linear Model (GLM) (Dobson and Barnett, 2018, pp. 1-20) assign independent and dependent variables to find their relationship. Reducing the original constraint only for continuous outcome variables makes data prediction covering both continuous and discrete variables and modeling more efficient.

5.2 Decision Tree (DT) (Kumar and Garg, 2018, pp. 31-37) in decision-making for data classification. The structure is divided into branches according to its properties until they are leaves that represent decision-making. Depending on the variable to set the data values in small batches, such models can help make analytical decisions.

5.3 Gradient Boosted Tree (GBT) (Kumar and Garg, 2018, pp. 31-37) is like a tree to make decisions with the resampling method, randomly creating several decision trees together until the best Gradient Boosting with the balance of the data set. The result is the average of the data sets and weight criteria on each side closest to each other.

5.4 Random Forest (RF) (Cutler, Cutler and Stevens, 2012, pp. 157-175) can predict classification and regression trees by creating new training sets. Times based on the existing database randomly from the predicted variables that were the most voted trees. Furthermore, the optimal division of data in subsets is a model that improves data classification for outstanding prediction performance.

Additionally, the Evolutionary Optimization method (Blum, et al., 2012, pp. 1-29) is used to find the best parameter set for each predictive model by generating random and selecting patterns by evaluating appropriate data. The research methodology of the Predictive Analytics

The process by Duration, Amount of Attendance, and Amount of Quiz data imported into the algorithm to predict learners' quality is shown in Figure 4.

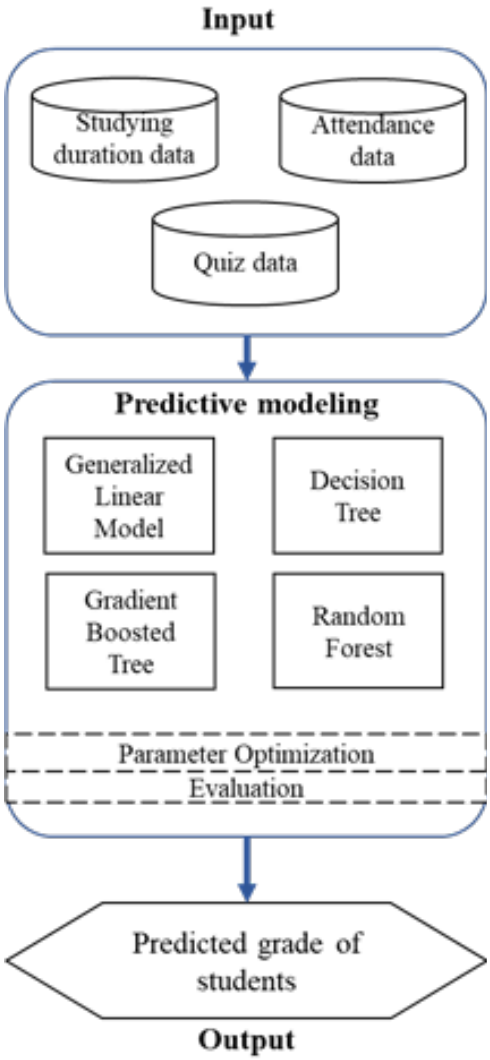

Figure 4 Predictive analytics process

6. Strategic Deployment

According to the goals set, strategic plans have been created to improve teaching efficiency and learning in the course to be of the highest quality. As a result of Descriptive Analytics on the learner, classification is to close the gap for each learner to improve their learning and encourage more motivation to study. As for Predictive Analytics, predicting learners' quality can be improved from the problems of former learners for future learners to have learning efficiency according to the course criteria.

Results

1. Student-level segmentation

A total of 2,120 learners were grouped by behaviors in the Duration, Amount of Attendance, Grade, and Amount of Quiz variables, divided into four clusters from the centroid point in each cluster, including Weak-cognitive, Inattentive, Intelligent, and Unenthusiastic

Weak-cognitive is a learner who is vulnerable to learning. The amount of attendance followed by Low-grade results but contrary to the behavior shown, causing this group of learners to have little understanding of the

content they are learning. Although it takes much time to study, Grades are not much according to the action. Note that the highest value is Duration, indicating a longer duration of the learning.

Inattentive is the learner who ignores learning. Show that they have no interest in studying in the course as they should. The amount of attendance, grade, and quiz times were at the lowest level. Although the duration of the learning was more significant than the other variables, all variables were below the mean line.

Intelligent is a learner with fast learning. Due to receiving a grade in the high range with pass status, despite the minimum duration, attendance is high. Show that they can recognize content quickly. It makes it possible to do a great quiz.

Unenthusiastic is a learner with the same behavior throughout the class at a moderate level. No variables pass, meaning this group is not as enthusiastic about studying as they should be. The result is that the behavior shown, therefore, receives a low grade according to the behavior shown in the Parallel chart as in Figure 5.

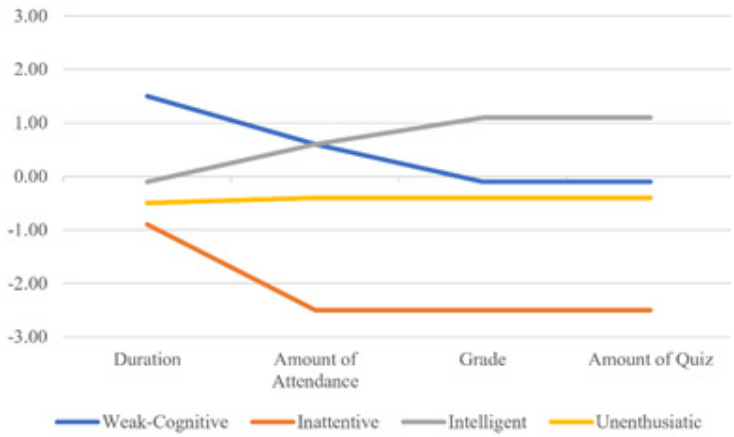

Figure 5 Centroid Chart for Student-level behavior clustering with a normalized value

2. Lesson-level segmentation

Classifying learners based on their learning behavior in that lesson by selecting the top nine most popular lesson out of 15 lessons from the duration of Learning and the amount of attendance is shown in Figure 6. K- means clustering was used to classify learners, divided into three types of lessons, consisting of 1) Lessons 1-3, 2) Lessons 4-6, and 3) Lessons 8, 10, and 15, which are grouped from popular groups with similar lesson ranges from Beginning, Middle, and Terminal. The learning behavior of different types of learners is as follows: Intelligent a learner with fast learning, Learnings have a high level of focus at a low grade that most Beginning Lessons have two Weak-cognitive lines, meaning that the beginning of lessons in the beginning Learners is in the process of adjusting to understand the content.

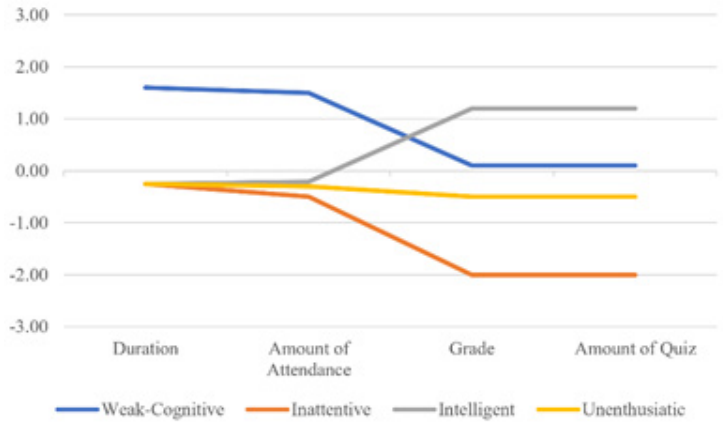

Figure 6(a) Centroid chart for lesson-level clustering of beginning lessons

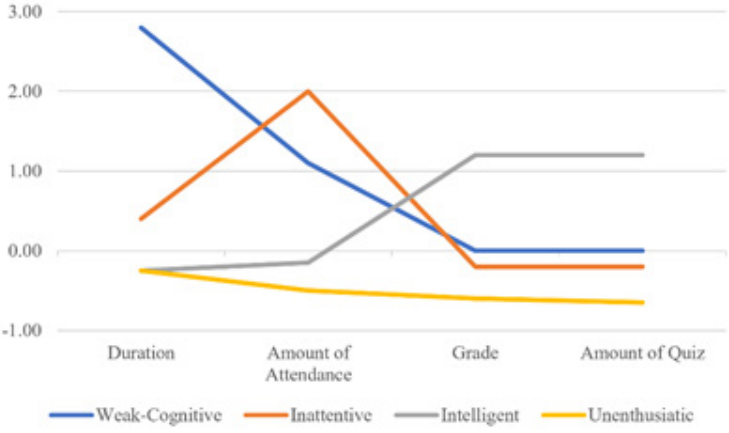

Figure 6(b) Centroid chart for lesson-level clustering of middle lessons

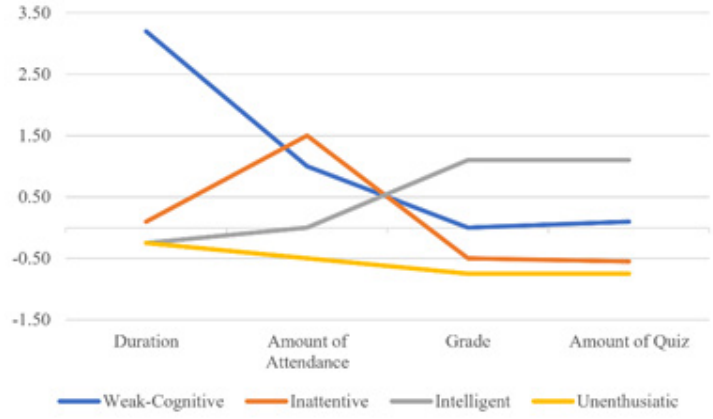

Figure 6(c) Centroid chart for lesson-level clustering of terminal lessons

3. Prediction Performance

1) Grade prediction by Student-level behavior and Lesson-level behavior

Model Performance to measure the best modeling efficiency based on the lowest Relative Error (RE) for predicting models with the least amount of quizzes. As a result, the grade is high. The unenthusiastic learner who has the same behavior throughout the course depends on the learners' behavior. An inattentive learner with the highest value is the amount of attendance.

On the contrary, the learning duration is short-lived due to a lack of concentration.

Weak-cognitive who are vulnerable to learning takes a long time to understand the content. Student-level and Lesson-level results in various best models, as shown in Figure 7. The best models of each group are marked with *.

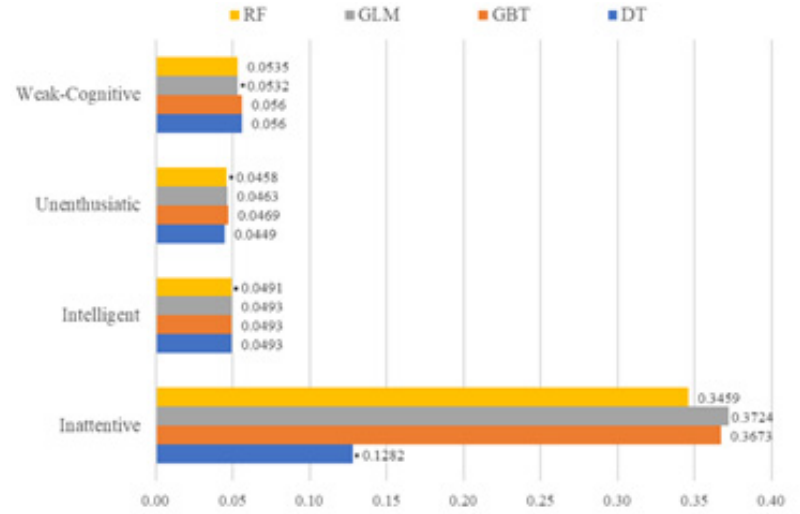

Figure 7(a) Relative error of student-level prediction model

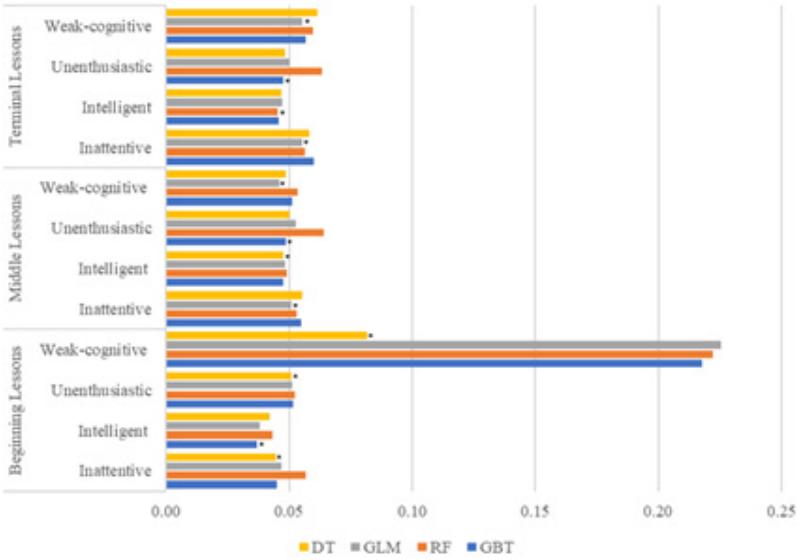

Figure 7(b) Relative error of lesson-level prediction model

2) Factors Importance affecting the prediction of learners' final grades for Amount of Quiz, Amount of Attendance, and Duration variables were related to predicting learning quality through grades as variables with high to low influence.

The importance of the model DT, RF, and GBT is calculated based on Information Gain Ratio, while the GLMs are picked from the coefficient of each variable. The most important factor in all clusters for predicting student performance, both Student-level and Lesson-level, is the amount of quizzes, while the Duration and Amount of Attendance had different rankings, as shown in Figure 8.

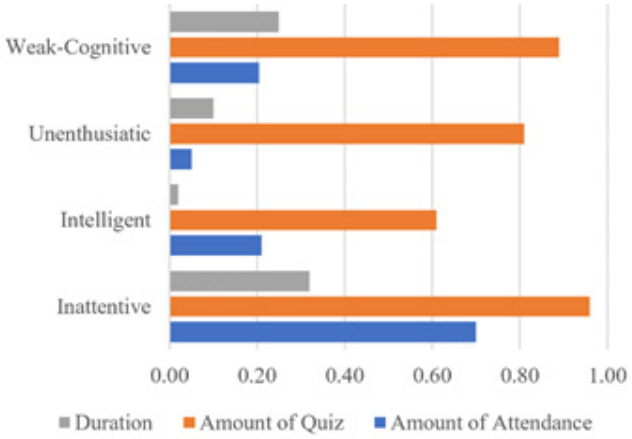

Figure 8(a) Factor Importance of student-level predictive models

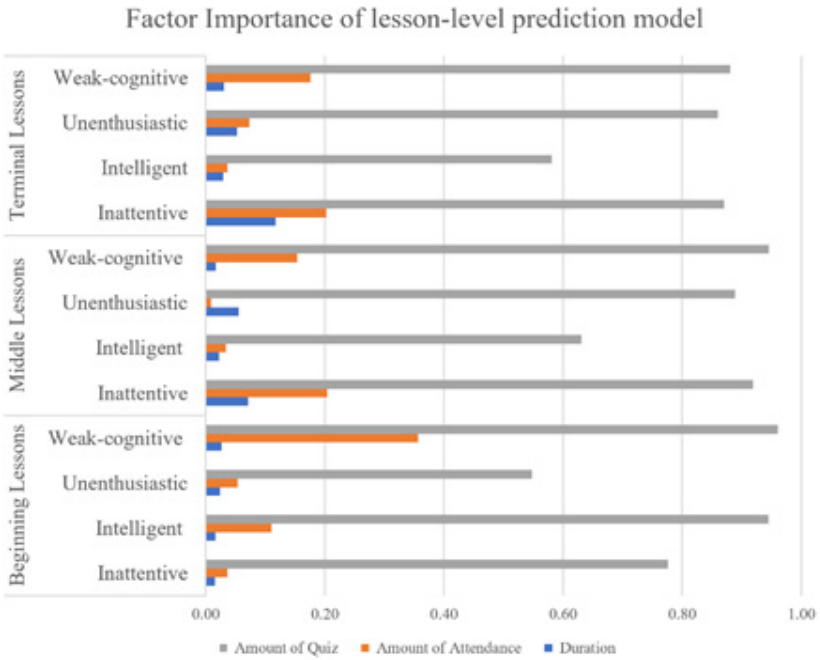

Figure 8(b) Factor Importance of lesson-level predictive models

Finally, the research implementation is compared with the current works. This proposed research is more comprehensive than the work of Yanta, et al. (2021, pp. 567-573) since this work provides a strategy of educational deployment that could be more beneficial for the management level. However, the proposed work still has a gap in prescriptive optimization results, e.g., what characteristics details of the student can achieve the study goal. This work also advances the research of Zhang, Zhang and Ran (2018, pp. 629-634) since they focused on segmentation. However, we provided the advantage of segmentation that leads to predictive modeling. Moreover, compared to Nguyen, Chen, and Saravanarajan (2022, pp. 1-5), Suleiman and Anane (2022, pp. 1480-1485), Han, et al. (2017, pp. 1-7), and Xu, et al. (2017, pp. 11-13), our work analyzes the importance of factors to make the deployed strategy more actionable and explainable. However, our work still lacks individual perfor-

mance prediction, e.g., GPA or CGPA of each student, which Suleiman and Anane (2022, pp. 1480-1485) proposed. More related factors like the social interaction feature that Nguyen, Chen and Saravanarajan (2022, pp. 1-5) used may be more beneficial for educators due to more comprehensive input factors.

4. Strategic Deployment

The strategic deployment plan that aims to improve learning efficiency is shown in Table 1. The decision-making steps of the strategy are shown in Figure 9. All learners must pass and attend more than 75% of total class times. All learners could be monitored for their study activities to achieve that goal.

Conclusions

This research analyzes the relationship between learners' grouping in learning behavior and predicts learner quality through learning from English Level 1, a course for first-year students on Small Private Online Courses (SPOC). The result comes from analyzing the correlation from the classification of learners, both student-level and lesson-level, to classify learners from similar behavior into the same category. As a result, the teacher or the lesson developer recognizes the learning behavior to adjust the teaching methods and content for the learners to have better behavior and learning efficiency. They also predicted learning quality by selecting the best model for the most accurate predictions across clusters.

For K-Means Clustering, there are four types of learners in the course: 1) Intelligent, 2) Weak-cognitive, 3) Inattentive, and 4) Unenthusiastic. Additionally, four machine learning models were used to predict learning quality through learning behaviors: Generalized Linear Model, Decision Tree, Gradient Boosted Trees, and Random Forests, which selected the best model by evaluating relative error. Furthermore, the most important factor affecting academic performance is the number of quizzes, as the Quiz score is included as an average in the final score. The deployment has a strategic plan to improve learning efficiency from the problem management of the learner type and adopt different strategies. For example, the university can improve the content to be diversified in the course, regulations for determining the number of study hours, etc. All of this is towards the set goals.

Finally, this research's main finding is the multi-dimensional analysis of student behavior in the foundation English course. The analysis consists of student-level and learning behavior levels that can motivate future research or real education institutes to apply in their organization. The full-stream analytics value chain uses the combination of clustering techniques and predictive modeling and strategically deploys in the real education processes.

Further research can be developed on the prediction model linked to the system's actual database. So that teachers analyze learning behavior data to adjust the educational strategy with content and teaching methods and control learner behaviors according to the course criteria. As a result, both teachers and students achieve maximum efficiency.

Acknowledgment

This work is partially supported by the Division of Academic Affairs, Mahidol University, for the anonymized historical learning data. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

Table 1 Strategic deployment plan

Segmentation	Strategy
Intelligent	Increase the content to be diversified in the course. (Reading skills about complex texts and specialized articles)
	Intensity level for Quiz from Performance. (Grammar skill in a sentence and interactive quiz)
Weak-cognitive	Pre-test knowledge skills to evaluate the level of learning content.
	Flipped Learning
Inattentive	Divide the content into smaller teaching sections, taking approximately 5 minutes to study per 1 topic to increase learners' ability to focus.
	Adjust the teaching content so that students are more interested in learning.
	Increase the number of hours of study regulations by at least 15 hours.
Unenthusiastic	Thread of comments about content and method of teaching.
	Increase the number of hours of study regulations by at least 15 hours.

Figure 9 Decision flow of the deployed strategy

Bibliography

Blum, C., Chiong, R., Clerc, M., De Jong, K., Michalewicz, Z., Neri, F. and Weise, T. (2012). Evolutionary optimization. In Raymond, C., T. Weise and Z. Michalewicz (Eds.), **Variants of evolutionary algorithms for real-world applications** (pp. 1–29). Germany: Springer.

Cutler, A., Cutler, D. R. and Stevens, J. R. (2012). Random forests. In C. Zhang and Y. Ma (Eds.), **Ensemble machine learning** (pp. 157-175). New York: Springer.

Dobson, A. J. and Barnett, A. G. (2018). **An introduction to generalized linear models** (4th ed.). Boca Raton: Chapman and Hall/CRC.

Hamerly, G. J. (2003). **Learning structure and concepts in data through data clustering.** Doctoral dissertation, Ph.D., University of California, San Diego.

Han, W., Jun, D., Xiaopeng, G. and Kangxu, L. (2017). Supporting quality teaching using educational data mining based on OpenEdX platform. In **The Frontiers in Education Conference** (pp. 1–7). Indiana: Purdue University.

Kumar, V. and Garg, L. M. (2018). Predictive analytics: A review of trends and techniques. **International Journal of Computer Applications,** 182(1), 31–37.

Lu, H. (2018). Construction of SPOC-based learning model and its application in linguistics teaching. **International Journal of Emerging Technologies in Learning,** 13(2), 157–169.

Maslin, N. M., Consultant, P. and Ltd, S. S. (2010). Impact of modern technology. **HF Communications,** 3, 33–35.

Ministry of Higher Education, Science, Research, and Innovation (2021). **Higher education plan to produce and develop the country's people complete version 2021 – 2027.** Bangkok: Ministry of Higher Education, Science, Research, and Innovation.

Nguyen, H. T. T., Chen, L. H. and Saravanarajan, V. S. (lecturer). (January 3-5, 2022). Using feed-forward backprop, perceptron, and self-organizing algorithms to predict students' online behavior. In **The 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)** (pp. 1-5). Seoul: Republic of Korea.

Omran, M. G., Engelbrecht, A. P. and Salman, A. (2007). An overview of clustering methods. **Intelligent Data Analysis,** 11(6), 583–605.

Patterson, J. P. and Duer, D. (2006). High school teaching and college expectations in writing and reading. **English Journal,** 95(3), 81-87.

Patro, S. and Sahu, K. K. (2015). Normalization: A preprocessing stage. **International Advanced Research Journal in Science, Engineering and Technology,** 2(3), 20–22.

Prates, J. M., Garcia, R. E. and Maldonado, J. C. (2019). Small private online courses in computing learning: Evidence, trends and challenges. In **Brazilian Symposium on Computers in Education (SBIE 2019)** (pp. 129–138). Brasília: Brazilian Computer Society.

Suleiman, R. and Anane, R. (lecturer). (May 4-6, 2022). Institutional data analysis and machine learning prediction of student performance. In **The 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)** (pp. 1480-1485), Hangzhou: China.

Xu, J., Moon, K. H., Member, S. and Schaar, M. van der. (2017). A machine learning approach for tracking and predicting student performance in degree programs. **IEEE Journal of Selected Topics in Signal Processing,** 11(5), 1–13.

Yanta, S., Thammaboosadee, S., Chanyagorn, P. and Chuckpaiwong, R. (2021). Course performance prediction and evolutionary optimization for undergraduate engineering program towards admission strategic planning. **ICIC Express Letters,** 15(6), 567–573.

Zhang, G., Zhang, Y. and Ran, J. (2018). Research on clustering mining and feature analysis of online learning behavioral data based on SPOC. In **The 13th International Conference on Computer Science and Education** (pp. 629–634). Colombo: Informatics Institute of Technology.