Analysis and Prediction of Student Performance Based on Moodle Log Data using Machine Learning Techniques

https://doi.org/10.3991/ijet.v18i10.35841

Chayaporn Kaensar(✉), Worayoot Wongnin
Ubon Ratchathani University, Ubon Ratchathani, Thailand
chayaporn.k@ubu.ac.th

Abstract—During the COVID-19 Pandemic, many universities in Thailand were mostly locked down and classrooms were also transformed into a fully online format. It was challenging for teachers to manage online learning and especially to track student behavior since the teacher could not observe and notify students. To alleviate this problem, one solution that has become increasingly important is the prediction of student performance based on their log data. This study, therefore, aims to analyze student behavior data by applying Predictive Analytics through Moodle Log for approximately 54,803 events. Six Machine Learning Classifiers (Neural Network, Random Forest, Decision Tree, Logistic Regression, Linear Regression, and Support Vector Machine) were applied to predict student performance. Further, we attained a comparison of the effectiveness of early prediction for four stages at 25%, 50%, 75%, and 100% of the course. The prediction models could guide future studies, motivate self-preparation and reduce dropout rates. In the experiment, the model with 5-fold cross-validation was evaluated. Results indicated that the Decision Tree performed best at 81.10% upon course completion. Meanwhile, the SVM had the best result at 86.90% at the first stage, at 25% of the course, and Linear Regression performed with the best efficiency at the middle stages at 70.80%, and 80.20% respectively. The results could be applied to other courses and on a larger e-learning systems log that has similar student activity conditions and this could contribute to more accurate student performance prediction.

Keywords—student performance prediction, machine learning, Moodle LMS, database system subjects

1 Introduction

The education industry was impacted by the COVID-19 pandemic, with institutions changing platforms toward online learning. In these times, the demand for digital and innovative technology to support teaching assignments, manage classes and track learners, have become a crucial part of education [1].

Similarly, in Thailand, the government decided to close all educational institutes in 2021 and transitioned to online learning. Although online classes were portable, easily accessed, and increased adult learning opportunities, all schools and universities faced multiple challenges in requiring the adoption of online teaching programs [2–3].

Ubon Ratchathani University (UBU), located in Ubon Ratchathani Province in Northeast Thailand and established in 1990, exposed students to new research and technology. With over 13,000 students per year, many are encountering problems due to the transition to online learning, leaving most teachers and students facing the difficult task of continuing their coursework. For example, students miss laboratory work and are given less feedback because they are frequently too shy to ask questions during the course, while teachers lack direct classroom contact with students, can neither observe them nor explain the content, and cannot immediately notify students who may be at risk. Many such key factors increase the likelihood of many students failing, dropping out, and withdrawing early before graduation.

However, with increased technology, the trend towards student-centered and life-long learning that responds to students' needs, increased substantially [4]. Hence, teachers and universities must provide online learning tools to support education efficiently. In this regard [4–5], Moodle LMS offers a learning environment with digital software that has been used in many universities worldwide. This is evidence that students turn to Moodle LMS for rapid access to group activities and classroom interaction.

Fortunately, we have found that accessing online courses in the daily life of UBU students is not much of a problem. The reason for this is that our university previously provided the UBU Moodle LMS as a supplemental learning environment since 2017, serving more than 25,110 users among 2,258 courses, for 99,930 activities [6]. This allowed students to interact with and access lessons more easily, while teachers can use it as an efficient tool to manage the classroom via quizzes, assignments, test examinations, and other activities. In addition, the UBU LMS can collect online routine student activity data traced from LMS resulting in a large log file. This LMS advantage is used for analyzing and forecasting online student habits to better understand teacher perceptions [7].

This benefit is related to the research in [8–10] that studied and implemented a predictive model with different Machine Learning techniques from a dataset collected from student LMS interaction both to help teachers depict patterns, predict student academic success, and for the use of students to self-identify their own risk of failure. Furthermore, many recent studies such as [11–13] have explored and compared the accuracy of various Machine Learning algorithms with a model trained to find the best one for predicting student performance. However, few researchers have compared various models created through different stages of course duration to detect student performance early.

Therefore, our work fills these gaps by focusing on the question: "How may one predict student performance using the database system subjects employing six models while training at four stages of 25%, 50%, 75%, and 100%, of course completion?". Overall, we highlight both identifying patterns of UBU student behavior by applying Predictive Analytics and comparing six Machine Learning algorithms, including Neural Network (NN), Random Forest (RF), Decision Tree (DT), Logistic Regression (GR), Linear Regression (LR), and the Support Vector Machine (SVM), with a model focused on different stages of analysis of the course.

In this study, we used a sample UBU Moodle log of database system subjects (DBS) comprising approximately 54,803 student activities, containing 8 predictor variables, such as the course, assignments, forums, files, quizzes, labels, videos, and attendance, for students studying in the Data Science and Software Innovation Program (DSSI) at the Faculty of Science during 2021–2022. Hence, it was important to investigate this and conduct DBS because it is a prerequisite course required before taking other subjects at the next level in the program, such as Web Programming, Software Engineering, and particularly, Cooperative Education, which is to prepare individuals to be successful for internship experiences before graduation. Specifically, this subject has two lectures and two laboratory hours a week per semester, which reveals challenges to online teaching in laboratory-based disciplines. Furthermore, it was reported that this subject suffers from many failing students, for more than 25% of students failed to pass from 2019–2020. Given that poor or failing grades for DBS are on the rise through the years likely influences enrollment plans and dropout behavior for the next course.

This study contributes to the research in two ways. First, we developed and compared six prediction models from the Moodle log using Machine Learning techniques for early prediction of student educational performance at four stages identified best in the terms of the Confusion Matrix (Accuracy, Precision, Recall, F1-measures). Second, based on the result of this study, all stakeholders in education benefited from the prediction model. These included students becoming aware of their learning status, while teachers and universities who had to present online lectures and laboratory exercises, monitored student performance to take precautions for students at risk, adjusting learning strategies, and making education more efficient.

1.1 Objectives

1. To explore student behavior through Moodle log of database system subjects.
2. To compare six Machine Learning classification models of LMS log to predict student performance early at different stages of 25%, 50%, 75%, and 100% of the course completion.
3. To investigate the best prediction model upon course completion.

1.2 Research questions

1. It is possible to identify student performance by analyzing their interactions with the LMS log at different stages of 25%, 50%, 75%, and 100% of the course?
2. Which Machine Learning algorithm can perform best at predicting student performance at different stages of 25%, 50%, 75%, and 100%?
3. How effective are Machine Learning algorithms for predicting the best model upon course completion?

The remainder of this paper is organized as follows: Section 2 offers a literature review, section 3 presents concepts, section 4 details the research methodology, and section 5 explores the research results. Finally, section 6 discusses conclusions and plans.

2 Literature review

This section reviews student performance predictions from their behavior log data using Machine Learning techniques. The three areas investigated for performance prediction in this study are discussed below.

2.1 Using Moodle LMS and learning analytics in education

During the COVID-19 pandemic, many global universities transformed traditional face-to-face learning into virtual platforms such as the web, video, online conferencing tools, intelligent tutoring system (ITS), and especially the Moodle LMS, which is one of the most popular LMS in the world, including Thailand [3, 11, 12].

The Moodle LMS is an open-source platform that provides features that easily support learning, evaluation, and collaboration, as well as instructional communication, also enabling teacher interaction with students [1]. Mostly, scientific and other fields in education learning have been employed in the environment to receive attention and improve the quality of education. For example, [14], a tool was developed for measuring student perception of the use of ICT tools during the COVID-19 crisis. This clarifies factors that affect learning performance such as student skills, social interactions, and the cost of learning.

Correspondingly, [15] factors were detected that most influenced 258 students in their perception of and satisfaction with e-learning during the COVID-19 situation. The results indicated that student factors and software quality have become the dominant issues, characterized by the ease of interaction and learning practice. We observed that [5] investigated student perception of the use of LMS features which was carried out on 122 students at Crete University. The authors noted that students preferred accessing learning material on their more convenient smartphones. However, they suggested that future studies should concentrate on the relevant data sources and measure learning objectives through pattern usage.

Likewise, [16] surveyed the review of adaptive digital learning systems and techniques to describe the utilization of the LMS, and the results support the idea that learning style has a benefit in encouraging them to become more proactive, and providing information to the teacher. As noted above, although much research has proved that the LMS online learning process is based on several factors such as system quality, content quality, student skills, perceived usefulness, and so forth, the task of analyzing students' usage patterns and tracking their performance in online learning is still discussed.

In the last few years, Moodle 3.4 released the Learning Analytics feature to explore some predictions [11]. For example, the conditions of analysis to improve reliable prediction were investigated in [17], while [18] also used the Learning Analytics to predict which student's online activities will improve learning outcomes, which proved this exercise activity returns the most significant results, but suspicions about experimental performance remain.

Thus, to implement and validate a more reliable model, Machine Learning is widely applied for analyzing student behavior data and finding the specific hidden knowledge related to student performance. This finding was confirmed by the study of [16] that established Machine Learning is to effectively extract, analyze and process, data that can predict learning outcomes for more accurate performance. It is for these reasons that Machine Learning has been employed in a variety of studies.

2.2 Applying Machine Learning for performance prediction

Applying Machine Learning in education has become a major direction of Educational Data Mining applications [8] such as identifying student behavior, assessing pass-fail grading, and especially, predicting student performance. Therefore, the development and combination of education and machine learning is the current trend at this time [11].

Most published works concentrated on using a single classification algorithm. For example, in [19] Logistic Regression was used to analyze courses of LMS predictor variables and predict students' grades over 10 weeks. The results indicated that assessment grades related to final grades, while interaction events like discussions, forums, or wiki usage, are a less reliable predictor of the final grade.

In [20], a Random Forest algorithm was applied to build a model using log events (lectures, quizzes, labs, videos) to predict student failure with high accuracy at 96.3%. They revealed that the results of lab scores are the strongest predictor in this study. Likewise, [21] uses 4-course data such as weekly material, videos, lectures, quizzes, and exercises, from Van Lang University in Vietnam, by applying a Linear Regression classifier to forecast the risk of failing the course. Students with less than 37% interaction were found to be at risk of failing. An analysis of 30,000 student records was conducted [22] which included five indicators such as academic performance, assignments, access, social aspects, and quizzes, which revealed that after the Regression Tree was applied, the implemented model had an accuracy rate of 89.70%.

Following that, the capacities of the Neural Network were enhanced by integrating feature selection to create a prediction model for spotting a student who may be on the failing list [23]. This study employs Multiple Linear Regression to investigate key features with a 97% accuracy rate in predicting student performance. Meanwhile, a Decision Tree classifier framework has been proposed to predict student outcomes at risk at three levels (high, average, and low) [24]. Many feature details (gender, session, class duration, GPA, major, degree, year, attendance, and midterm score) were collected from students who studied in the Introduction to Programming course at Buraimi University College in Oman. The accuracy of the model was 87.88%, demonstrating its effectiveness.

Similarly, the Decision Tree model was proposed to predict student dropout rates at Phayao University in Thailand [3]. A dataset of 397 students was analyzed and the dropout causes were considered. The result produced six courses, such as Fundamental Information Technology, Introduction to Programming, Thai Language Skills, Principles of Marketing, Life, and Health, and Principles of Management, that affect student achievement. Overall precision was around 87.21%. This was close to [25] that collected a dataset of University Malaysia Pahang (UMP) between 2002–2015 to classify student performance of those who enrolled in chemistry subjects using multilayer perceptron (MLP), which reported an accuracy of 92.23%.

Although prediction performance works with a single highly accurate algorithm, many researchers explored and compared different models to discover the best performance.

2.3 Comparing multiple machine learning for performance prediction

In this section, several prediction models with Machine Learning techniques were studied to predict student performance and were compared with a popular algorithm to establish model effectiveness. For example, there is an experiment [12] that uses three models: Cluster Analysis, Multiple Linear Regression, and Correlation, by constructing four attributes (age, task, questionnaire, and access) to predict student academic performance. The results indicated that there is an inverse relationship between student performance and age, but student performance is related to the amount of Moodle interaction.

Another study [26] compared three algorithms for predicting student performance and found that the Decision Tree algorithm outperforms the others with a precision rate of about 95.78%. In addition, four Machine Learning classifiers such as Support Vector Regression, Random Forest, Linear Regression, and Decision Tree, were employed [11] to indicate the best performance for predicting the final GPA. They discovered that the GPA of the third-year student is the most important predictor for graduation, while marital status and age are irrelevant.

In [27], four Machine Learning techniques were employed, such as K-Nearest Neighbor (KNN), XGboost, Random Forest, and Support Vector Machine. This predicts student grades for Object-Oriented Programming Subject (OOP), using 7,057 event records from Moodle, consisting of five variables (ex. attendance, lectures, codes, exercises, and assignments). Although this experiment extracted data from Moodle logs and Zoom reports, the size of the dataset and the number of event activities are known to be less. The accuracy rate achieved with Random Forest was 78%.

Likewise, [28] also implemented and compared sets of methods such as Bayesian network, Logistic Regression, SVM, and Random Forest, to predict learner motivation on a MOOC platform. The results indicate that Random Forest was most accurate at 95% compared to other techniques. While [13], they used four Machine Learning methods including Naïve Bayes, Decision Tree, MLP, and Logistic Regression to predict student performance at 10%, 25, 33%, and 50% of course duration to detect high-risk students. The results showed that MLP offers the best performance at 80% accuracy.

The aforementioned [29], analyzed five algorithms, including Logistic Regression, Naïve Bayes, MLP, Random Forest, and Support Vector Machine. They used various criteria like activity types, timing statistics, and peripheral activity count, to predict student performance providing high precision at 97.4% with the Random Forest. Additionally, [30] compared six algorithms (Gradient Boost Trees, Naïve Bayes, Decision Trees, Random Forest, Support Vector Machine, and Logistic Regression) with data collected at percentages of 20, 40, and 60 of course completion. The results showed Random Forest has a high accuracy (84.47%).

Using data from Moodle and Blackboard [31], some researchers applied nine classifier algorithms, including Decision Table, Random Forest, Decision Tree, KNN, Naïve Bayes, Support Vector Machine, One Rule, MLP, and JRip. Furthermore, feature selection techniques such as filter-based and wrapper-based were applied for choosing the finest feature of the dataset that correlated to student performance. The results showed that Random Forest and KNN algorithms outperformed the others. One sees that [32] investigated seven algorithms of CART, J48, C5.0, KNN, Naïve Bayes, Support Vector Machine, and Random Forest. They considered varying parameters on the various classifiers collecting data from school, college, and e-learning platforms. The result showed that C5.0 and Random Forest are more accurate than the others.

However, the purpose of the previous work was concerned with applying Machine Learning classifiers to predict student performance in Moodle LMS. Based on this, some studies analyzed and compared several classification algorithms to indicate the best predictive model. Few works provide and compare different models at different stages of the class. In addition, none of them used various input predictors in terms of the early prediction of student performance.

3 Concept

In this work, we describe the concept of the Machine Learning algorithm as follows:

3.1 Linear Regression

Linear Regression (LR) is a widely used statistical method in Machine Learning techniques. These regression estimates explain the correlation between dependent and independent variables. The regression's simplest form is defined by Equation (1), in which the coefficients can be computed as estimates of some model parameters, defining the relationship between two entities (predictor value and response). Given Y is a dependent variable, a is a constant, b is a regression coefficient, X is an independent variable, and ε is a random error value [33].

There are three major types of regression analysis, including forecasting an effect, determining the strength of predictors, and trend forecasting.

$$Y = a + bX + \varepsilon \tag{1}$$

3.2 Logistic Regression

Logistic Regression (GR) plays a major role in Machine Learning techniques. It is often applied for classification and Predictive Analytics, analogous to the LR regardless of how they are used. Indeed, LR and GR are used for solving regression and classification problems, respectively. This algorithm can handle not only fitting a regression line but also fit an "S" shaped-logistic method predicting two maximum statuses (0 or 1). Thus, this method will be employed to analyze observations by applying different types of data, and clearly reveal the most effective variables [34].

3.3 Decision Tree

Decision Tree (DT) is one of the most effective learning techniques which are widely used in several areas such as education, statistics, banking, and recognition, among others. The tree structure is constructed from a root node, internal nodes, and leaf nodes. In DT, when testing an attribute on all internal nodes, the output of the test is on a "branch", while each "leaf" node is assigned a class label. The main aspect of DT is to minimize the number of tests equal to the number of nodes that are not leaves. So, solving classification problems is an important task for this technique. In addition, the DT classifier has a high accuracy rate and can deal with large and complicated datasets [33].

3.4 Random Forest

Random Forest (RF) is a commonly-used supervised Machine Learning method playing a significant role in classification and regression problems. The RF is an extension of the bagging function as it combines feature randomness and bagging on samples to construct an uncorrelated model [34]. Thus, RF is a group of Decision Tree but RF only selects a subset of those features, while DT considers all the possible feature splits. It provides many key features such as producing a reasonable prediction without hyperparameter tuning, reducing the risk of overfitting, providing flexibility, and easily determining feature importance [33].

3.5 Artificial Neural Network

An artificial Neural Network (NN) aims to recognize underlying relationships in a set of networks comprised of an input layer, a hidden layer, and an output layer that attempts to mimic the operation of the human brain. The process of NN is divided into two parts: network topology and learning adjustment [33]. NN was designed as a network topology, which involves arranging a network with its nodes and connecting links such as feedforward, recurrent, multi-layered, convolutional, or single-layered, while learning adjustment is a method of training the neural network. Gradient descent and Backpropagation algorithm are two common methods for training a neural network [34].

3.6 Support Vector Machine

The Support Vector Machine (SVM) is an adaptable supervised learning method for problems involving classification and regression. Each data point in n-dimensional space corresponds to the value of each feature. The classification is then carried out by selecting the hyperplane which best distinguishes the two classes. SVM can conduct linear classification efficiently by mapping the given input set into higher dimensional spaces. SVM is grouped into two different types, non-linear and linear SVM [34].

From an analysis of the literature and the concept, it becomes clear that several Machine Learning algorithms have been set up primarily to predict student academic performance based on Moodle LMS log files. However, lacking analysis data and comparing the different predictive models of student performance at various stages of the online learning process may be explained and turned into a sophisticated research method, which will be discussed in the next section.

4 Methodology

This study presents the use of Predictive Analytics techniques to analyze data concerning student behavior, using 54,803 records obtained from the UBU Moodle environment of the DBS, to predict student academic performance. In addition, we also compared the performance of the different classification methods (Neural Network, Random Forest, Decision Tree, Logistic Regression, Linear Regression, and the Support Vector Machine) in constructing prediction models.

We addressed the student performance prediction effect of models generated from six supervised algorithms at four different periods at stages of 25%, 50%, 75%, and 100%, of course completion. The research process was separated into four main stages as shown in Figure 1.

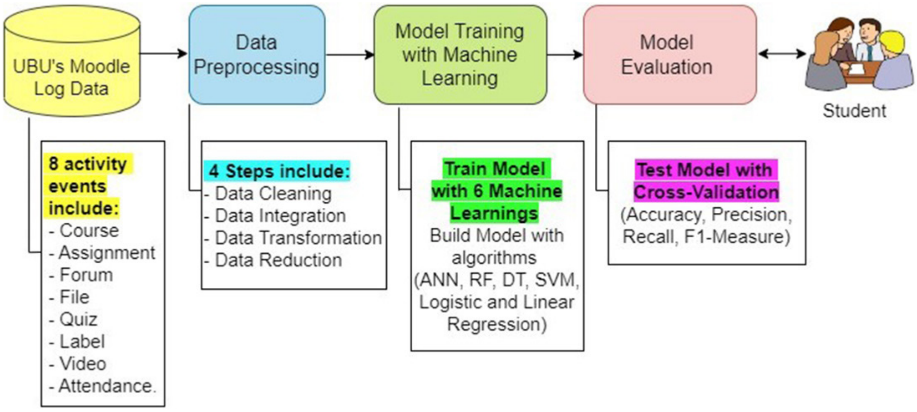
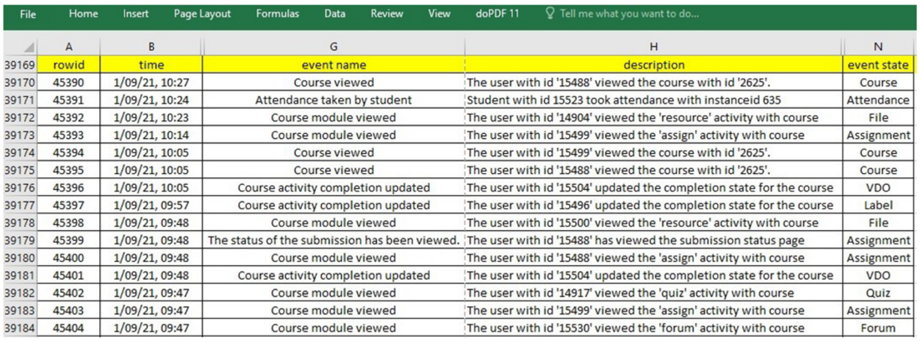

UBU's Moodle Log Data
Data Preprocessing
Model Training with Machine Learning
Model Evaluation
Student
8 activity events include:
- Course
- Assignment
- Forum
- File
- Quiz
- Label
- Video
- Attendance.
4 Steps include:
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
Train Model with 6 Machine Learnings
Build Model with algorithms (ANN, RF, DT, SVM, Logistic and Linear Regression)
Test Model with Cross-Validation
(Accuracy, Precision, Recall, F1-Measure)

Fig. 1. The research process

4.1 Collecting Moodle log data

The sample data set was collected from the UBU LMS, the university's online learning management system, for students who studied DBS at UBU during the academic year 2021–2022. This file can be shown in Figure 2.

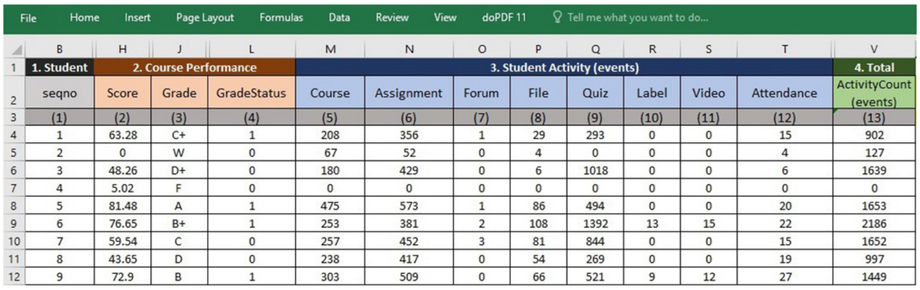
In Figure 2, this file is about 54,803 event records which contained student interaction data according to eight various attributes such as the courses, assignments, forums, files, quizzes, labels, videos, and attendance.



	A	B	G	H	N
39169	rowid	time	event name	description	event state
39170	45390	1/09/21, 10:27	Course viewed	The user with id '15488' viewed the course with id '2625'.	Course
39171	45391	1/09/21, 10:24	Attendance taken by student	Student with id 15523 took attendance with instanceid 635	Attendance
39172	45392	1/09/21, 10:23	Course module viewed	The user with id '14904' viewed the 'resource' activity with course	File
39173	45393	1/09/21, 10:14	Course module viewed	The user with id '15499' viewed the 'assign' activity with course	Assignment
39174	45394	1/09/21, 10:05	Course viewed	The user with id '15499' viewed the course with id '2625'.	Course
39175	45395	1/09/21, 10:05	Course viewed	The user with id '15488' viewed the course with id '2625'.	Course
39176	45396	1/09/21, 10:05	Course activity completion updated	The user with id '15504' updated the completion state for the course	VDO
39177	45397	1/09/21, 09:57	Course activity completion updated	The user with id '15496' updated the completion state for the course	Label
39178	45398	1/09/21, 09:48	Course module viewed	The user with id '15500' viewed the 'resource' activity with course	File
39179	45399	1/09/21, 09:48	The status of the submission has been viewed.	The user with id '15488' has viewed the submission status page	Assignment
39180	45400	1/09/21, 09:48	Course module viewed	The user with id '15488' viewed the 'assign' activity with course	Assignment
39181	45401	1/09/21, 09:48	Course activity completion updated	The user with id '15504' updated the completion state for the course	VDO
39182	45402	1/09/21, 09:47	Course module viewed	The user with id '14917' viewed the 'quiz' activity with course	Quiz
39183	45403	1/09/21, 09:47	Course module viewed	The user with id '15499' viewed the 'assign' activity with course	Assignment
39184	45404	1/09/21, 09:47	Course module viewed	The user with id '15530' viewed the 'forum' activity with course	Forum

Fig. 2. Log file of database system subjects on UBU LMS during 2021–2022

In the detailed outline of the dataset, we next collected and summarized the dataset from student activities and their performance. That is, each row represents student information which is divided into four parts: First, column (1) "rowid" identifies the individual student, Second, column (2)–(4) denote course performance where status 1 is a passing grade (score is greater than or equal to 60%) and 0 is failing grade (score is less than 60%). Third, columns (5)–(12) contain eight student activities such as the courses, assignments, forums, files, quizzes, labels, videos, and attendance which store many student views or submissions for each activity. Last, column (13) keeps an "ActivitiyCount," which refers to the total interaction events for an individual student. This can be viewed in Figure 3.



	B	H	J	L	M	N	O	P	Q	R	S	T	V
1	1. Student	2. Course Performance			3. Student Activity (events)								4. Total
2	seqno	Score	Grade	GradeStatus	Course	Assignment	Forum	File	Quiz	Label	Video	Attendance	ActivityCount (events)
3	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
4	1	63.28	C+	1	208	356	1	29	293	0	0	15	902
5	2	0	W	0	67	52	0	4	0	0	0	4	127
6	3	48.26	D+	0	180	429	0	6	1018	0	0	6	1639
7	4	5.02	F	0	0	0	0	0	0	0	0	0	0
8	5	81.48	A	1	475	573	1	86	494	0	0	20	1653
9	6	76.65	B+	1	253	381	2	108	1392	13	15	22	2186
10	7	59.54	C	0	257	452	3	81	844	0	0	15	1652
11	8	43.65	D	0	238	417	0	54	269	0	0	19	997
12	9	72.9	B	1	303	509	0	66	521	9	12	27	1449

Fig. 3. Sample of student activities and the student performance dataset

4.2 Data preprocessing

The purpose of data preparation is to screen data for analysis and modeling. The process of cleaning and converting raw data leading up to processing and analysis is known as "data preparation". It is a vital phase before processing the data that typically involves reformatting data, performing data corrections, and integrating data sources to clarify data. It is divided into four steps as follows:

Data Cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset. It is a necessary process for removing imperfections, inaccuracies, and distorted data. For example, to clean up UBU LMS logs, and irrelevant details of the teachers and the admin in Moodle will be removed because they were found to have no impact on the process.

Data Integration is the process of merging data from numerous source systems to create unified sets of information for analytical purposes. Its purpose is to create clean and consistent data sets that meet the information needs of users. To prepare data for model development, we merge student grades and activity logs into our work.

Data Transformation is the method of converting data from one format to another, typically from the format of a source system to the required format of a destination system. To build classification models, we convert numerical grades into categories with pass and fail criteria.

Data Reduction is the technique of converting data to obtain a reduced representation of the dataset while maintaining the integrity of the original data.

We employed six classifier algorithms to analyze and train the model which will be explored in the next topic.

4.3 Model training with Machine Learning

In this step, we perform gathering, modeling, and analyzing, data to extract insights that can be used to make decisions. Raw data were used in the preprocessing phase which was cleaned, integrated, and transformed, into data entries for the training and testing process. Then, we start with building the model using Python programming through Google Colab, an open-source platform that supports many popular Machine Learning libraries [34].

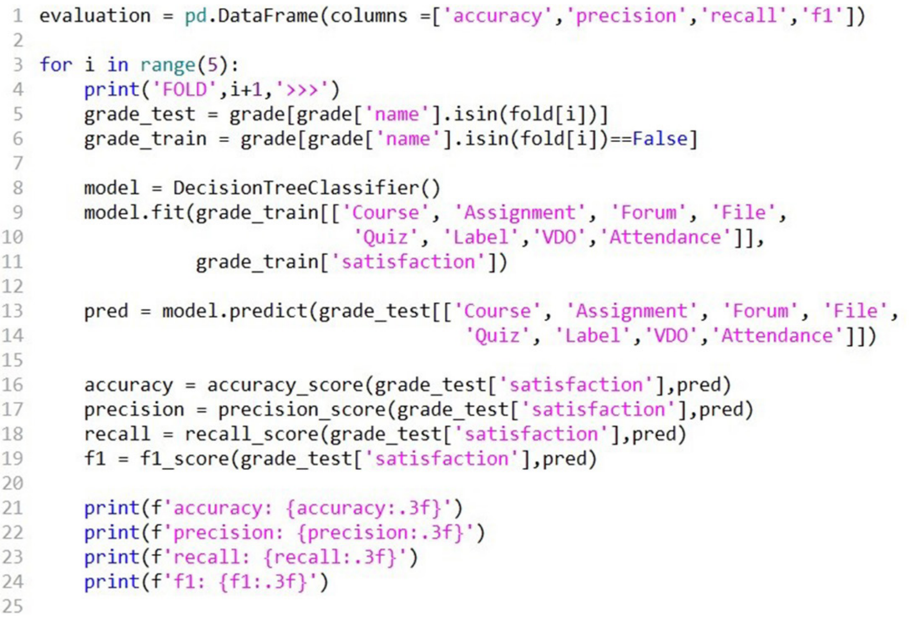
The implementation of codes in a training process was executed in three steps: split the data set, select the algorithm, fit, and check the model, and the following details could be established:

First Stage: we gathered data by splitting a previous preparation of 54,803 interactive records of students into two sets, with 80% of the first set serving as training data and the remaining 20% serving as testing data.

Second Stage: we selected algorithms based on those most commonly used to analyze such as Artificial Neural Network, Random Forest, Decision Tree, Logistic Regression, Linear Regression, and Support Vector Machine. Then, Scikit-learn (Sklearn), the most useful library for machine learning, was applied to create the models [35].

Third Stage: the output models were trained with training data to fit the model by using the method “fit()”.

For example, Figure 4 illustrates the sample Python code creating a training model for a Decision Tree with the class “DecisionTreeClassifier()”. Then, the output model was fit and assessed to measure the model which will be discussed under the topic of model evaluation.



```
evaluation = pd.DataFrame(columns =['accuracy','precision','recall','f1'])

for i in range(5):
    print('FOLD',i+1,'>>>')
    grade_test = grade[grade['name'].isin(fold[i])]
    grade_train = grade[grade['name'].isin(fold[i])==False]

    model = DecisionTreeClassifier()
    model.fit(grade_train[['Course', 'Assignment', 'Forum', 'File',
                             'Quiz', 'Label','VDO','Attendance']],
              grade_train['satisfaction'])

    pred = model.predict(grade_test[['Course', 'Assignment', 'Forum', 'File',
                                      'Quiz', 'Label','VDO','Attendance']])

    accuracy = accuracy_score(grade_test['satisfaction'],pred)
    precision = precision_score(grade_test['satisfaction'],pred)
    recall = recall_score(grade_test['satisfaction'],pred)
    f1 = f1_score(grade_test['satisfaction'],pred)

    print(f'accuracy: {accuracy:.3f}')
    print(f'precision: {precision:.3f}')
    print(f'recall: {recall:.3f}')
    print(f'f1: {f1:.3f}')

```

Fig. 4. Decision tree implementation in Google Colab using Python

4.4 Model evaluation

The experiment was divided into two steps for the model evaluation phase: k-fold cross-validation, and model evaluation. First, six classifiers (NN, RF, DT, GR, LR, and SVM) were evaluated using 5-fold cross-validation, one of the most commonly used techniques in measuring the model in Machine Learning. In this step, data was split into five parts, the first four parts for training (80%) and the remaining for testing (20%). As a consequence, an evaluation model was constructed and composed of five iterations, each iteration used a different dataset for testing. Second, we assessed the performance of models by calculating the average performance score across all folds. Each duration, such as 25%, 50%, 75%, and 100%, of the course completion, was applied by this approach in the same manner.

Performance evaluation was carried out with the Confusion Matrix, widely used for the evaluation of classification quality [33]. We used a variety of four common measures Accuracy, Precision, Recall, and F1-Measure [35]. All forms are defined in Equations (2)–(5):

- **Accuracy** is defined as the proportion of the total number of correct classifications to the sum of classifications, the formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- **Precision** is the proportion of positive test cases that are correctly classified, the formula is:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- **Recall** is the proportion of true positive test cases that were correctly classified, the formula is:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

- **F1-Measure** is a compromise of both precision and recall. This score is typically an effective metric for measuring the quality of an approach, the formula is:

$$F1 - measure = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

5 Results and discussions

This study aimed to create and compare different Machine Learning algorithms based on Moodle logs to analyze student behavior and predict their performance at different stages of the database system subjects during 2021–2022. To address the research question, a two-part experiment was performed. First, we wanted to explore the performance of six classification models trained at four stages of 25%, 50%, 75%, and 100%, of the course. Second, we wanted to show details of the best algorithm performed when the course has been completed.

In the first experiment shown in Figure 5, we created and executed six binary classifiers with default parameters, such as NN, RF, DT, GR, LR, and SVM. These models are trained at different moments in the course duration. Figure 5 shows a different perspective of six predictive models in terms of four measures such as accuracy, precision, recall, and the f1-measure which can conclude in the following cases:

First, at 25% of course progress, as shown in Figure 5a, the highest f1-measure for this stage averaged 0.602 (60.20%). This starts gradually, but these values increase by the next period. However, SVM achieved the highest f1-measure in the three examples analyzed for precision, accuracy, and f1-measure, with the highest score at 0.869 (86.90%). Second, the middle stage at 50% of course duration is depicted in Figure 5b. The plot shows that measurements of precision, accuracy, and f1-measure of LR are greater than the other classifiers, but the recall value is quite low. However, they had an f1-measure with a score of 0.708 (70.80%) and an almost perfect precision score of 0.996 (99.60%).

Similarly, 75% of course progress is depicted in Figure 5c. Overall, LR still provides better results than the other measurement classifiers. Specifically, the precision value reached 0.995 (99.50%), or close to 100%, and also had an f1-measure score of 0.802 (80.20%). However, this trend illustrates that their recall value is lower than DT and GR.

Lastly, when the course was completed as shown in Figure 5d, DT out performed the rest at the highest score of 0.811 (81.10%) and also presented the best performing model among the four classifications. The results are shown in Figure 5.

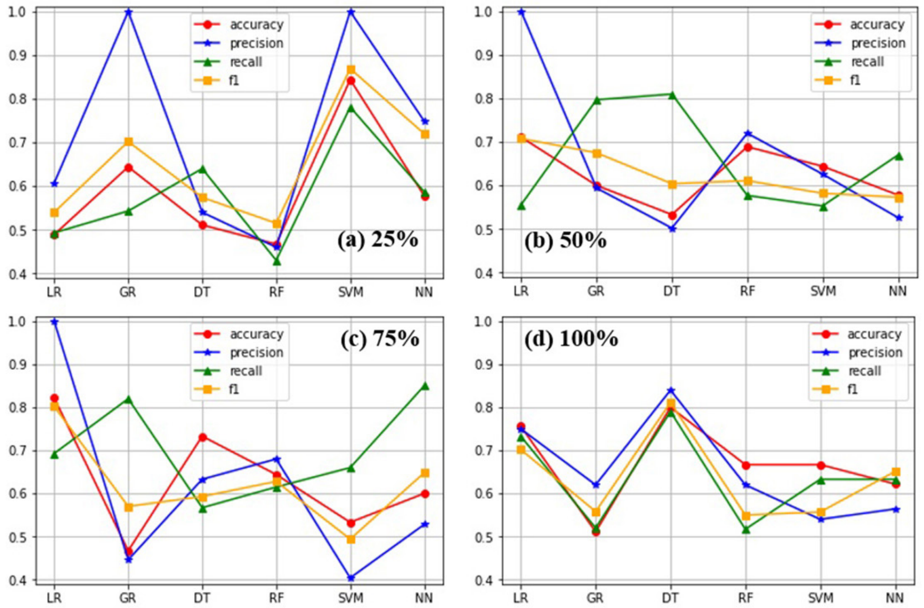


Fig. 5. Measurement classifications at different stages of prediction

Taken together, we showed the DT algorithm, the most efficient classifier, to predict student success or failure for the database system subjects which is summarized in Table 1. However, the score of accuracy, precision, and f1-measure, are high values and greater than 80% compared to the other classifiers in spite of the recall value being less, but it is still 79.00%.

Table 1. Comparison of six classification algorithms when course completion (100%)

Algorithm	Accuracy	Precision	Recall	F1-measure
LR	75.60%	75.00%	73.30%	70.20%
GR	51.10%	62.00%	52.00%	55.80%
DT	**80.00%**	**84.00%**	**79.00%**	**81.10%**
RF	66.70%	62.00%	51.70%	55.00%
SVM	66.70%	54.00%	63.30%	55.70%
NN	44.40%	27.30%	50.00%	34.90%

In the second experiment, we investigated and detailed the highest f1-measure belonging to DT to identify pattern knowledge for predicting student performance at course completion, which depicts visualized splitting and labeling of data given in Figure 6.

In the task of model prediction, we used eight various input variables such as the course, assignments, forums, files, quizzes, labels, videos, and attendance, extracted from the UBU LMS log to predict and distinguish between passing and failing students. As a result, Figure 6 found that students pass the course when the checked attendance rate is greater than 16 events and when they handed in assignments between 186–411 events, and had a high quiz instance greater than 217 events. Conversely, students who failed the course when they hit the attendance limits or undertook fewer than 16 events, and viewed course contents at fewer than 186 events, it was determined that variables such as attendance, assignments, quizzes, the course, and files, have more of an impact on student academic performance as we saw when interpreting the results of the DT model used in this study. The result is shown in Figure 6.

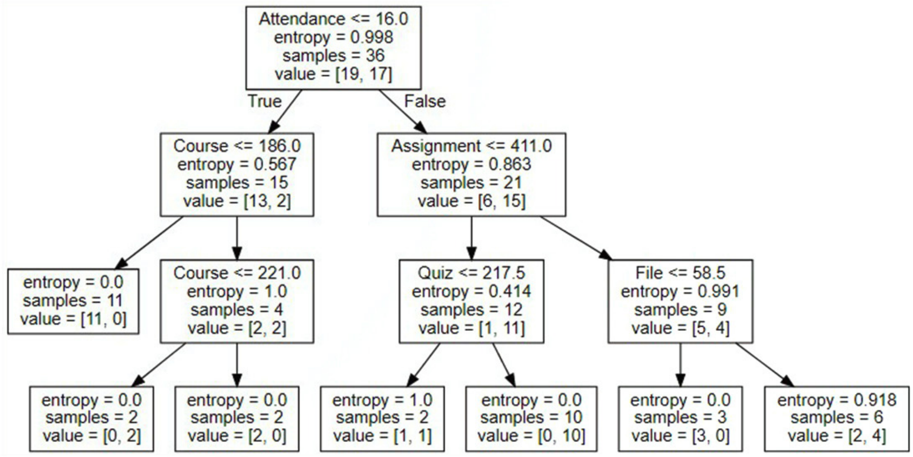


Fig. 6. A result of the decision tree for a database system subject

Each stage of the course has a different performance prediction model, and this is because the number of student activities is not uniform. Students' interaction levels are low at the early stages of the course, while Moodle provides a high density of activities at later stages. Furthermore, students will be aware of their situation in terms of the number of activities on UBU LMS based on the appropriate model for prediction at each stage.

These experimental results are consistent with the results of similar published work such as [26, 36, 37]. However, this study can be shown not only by analyzing and comparing several Machine Learning methods to predict student academic performance for which DT and RF are known to be highly precise, but also exploring the different prediction models created by the supervised algorithm at different course learning stages.

This could help both teachers and students in the future identify or filter students who are at risk at each learning stage. Moreover, our study captures and transforms student activity data into eight event variables, which enable users to specify necessary student activities in greater detail. Additionally, the result of this study may be applied in other online courses that have similar activities or instructional conditions.

6 Conclusions

This work aims to investigate and compare the capability of six Machine Learning techniques to predict student performance in online learning early, by applying Predictive Analytics using Moodle logs at different stages of course completion. Student activity data was collected using Moodle logs for those who studied the Database System subjects in the Data Science and Software Innovation Program at Ubon Ratchathani University during 2021–2022.

To this end, the results of the experiment generated a solution to the main research question. As such, this study supports the idea that student behavior in online learning platforms like Moodle affected student performance. Machine Learning classifiers such as Neural Network, Random Forest, Decision Tree, Logistic Regression, Linear Regression, and Support Vector Machine can produce early prediction models at different stages. Additionally, the result demonstrates that the performance of the Support Vector Machine obtained a high result of accuracy during the first stage of the course, and Linear Regression provided high performance in terms of accuracy at the middle stage. Lastly, the Decision Tree outperformed the others at the course completion stage. It is interesting to note that from the entire course period and according to the Decision Tree rule, it can be seen that students should pay more attention, and submit assignments and quizzes, engendering a positive impact on good academic performance. In contrast, if students fail to attend and view fewer courses and files, they may also fail a class.

This study could be applied to other courses from collected data on larger e-learning systems logs that have similar student activity conditions. This could contribute to more accurate student performance prediction. Meanwhile, the output model serves to support the development of tools that can predict student success at different stages of the course for teachers and universities to identify and detect students at risk of failing courses in an educational context early and take appropriate actions to enhance academic performance.

However, without enough interaction, the data usage of students still cannot be measured by such a model, for the effectiveness of prediction would not be at a high rate. As a suggestion to improve the predictive performance with LMS log data for further studies, one needs more insight into the presentation of the LMS, because well-designed courses can increase learner motivation and improve engagement [5, 38, 39, 40, 41].

In future work, we plan to study and focus on validating and extending the models by using multiple heterogeneous courses until graduation and also find interaction patterns that are repeated for better performance in other academic programs.

7 Acknowledgments

We acknowledge and thank the Faculty of Science and Ubon Ratchathani University for supporting this research.

8 References

[1] Zabolotniaia, M., Cheng, Z., Dorozhkin, E., & Lyzhin, A. (2020). Use of the LMS Moodle for an effective implementation of an innovative policy in higher educational institutions. International Journal of Emerging Technologies in Learning, 15(13): 172–189. https://doi.org/10.3991/ijet.v15i13.14945

[2] Wattanakasiwich, P., Suree, N., Chamrat, S., Saengsuwan, W., Suttharangsee, W., Panrat, T., Ruamcharoen, J., Trianpo, W., Amornsamankul, S., & Laesanklang, W. (2021). Investigating challenges of student centered learning in Thai higher education during the COVID-19 pandemic. In 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, pp. 1–7. https://doi.org/10.1109/FIE49875.2021.9637298

[3] Nuankaew, P. (2019). Dropout situation of business computer students, University of Phayao. International Journal of Emerging Technologies in Learning, 14(19): 115–131. https://doi.org/10.3991/ijet.v14i19.11177

[4] Alturki, U., & Aldraiweesh, A. (2021). Application of learning management system (LMS) during the COVID-19 pandemic: A sustainable acceptance model of the expansion technology approach. Sustainability, 13(19), 10991. https://doi.org/10.3390/su131910991

[5] Papadakis, S., Kalogiannakis, M., Sifaki, E., & Vidakis, N. (2018). Evaluating Moodle use via smart mobile phones: A case study in a Greek University. EAI Endorsed Transactions on Creative Technologies, 5(16): 1–10. https://doi.org/10.4108/eai.10-4-2018.156382

[6] Academic Affairs, "UBU LMS" Ubon Ratchathani University, 2022. [Online]. Available: https://lms.ubu.ac.th/ubulms/ [Accessed: Nov. 11, 2022].

[7] Okike, E. U., & Mogorosi, M. (2020). Educational data mining for monitoring and improving academic performance at university levels. International Journal of Advanced Computer Science, Applications, 11(11): 570–581. https://doi.org/10.14569/IJACSA.2020.0111171

[8] Murad, D. F., Heryadi, Y., Wijanarko, B. D., Isa, S. M., & Budiharto, W. (2018). Recommendation system for smart LMS using machine learning: a literature review. In 2018 4th International Conference on Computing, Engineering and Design (ICCED), Bangkok, Thailand, pp. 113–118. https://doi.org/10.1109/ICCED.2018.00031

[9] Oreški, D., & Hajdin, G. (2019). A comparative study of machine learning approaches on learning management system data. In 2019 3rd International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Los Angeles, USA, pp. 136–141. https://doi.org/10.1109/ICCAIRO47923.2019.00029

[10] Hooshyar, D., Pedaste, M., & Yang, Y. (2019). Mining educational data to predict students' performance through procrastination behavior. Entropy, 22(1): 1–24. https://doi.org/10.3390/e22010012

[11] Suleiman, R., & Anane, R. (2022). Institutional data analysis and machine learning prediction of student performance. In 2022 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, pp. 1480–1485. https://doi.org/10.1109/CSCWD54268.2022.9776102

[12] Bravo-Agapito, J., Romero, S. J., & Pamplona, S. (2021). Early prediction of undergraduate student's academic performance in completely online learning: A five-year study. Computers in Human Behavior, 115: 1–11. https://doi.org/10.1016/j.chb.2020.106595

[13] Riestra-González, M., del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. Computers & Education, 163: 1–36. https://doi.org/10.1016/j.compedu.2020.104108

[14] Galarce-Miranda, C., Gormaz-Lobos, D., Kersten, S., & Hortsch, H. (2022). Developing and validating an instrument to measure students' perceptions of the use of ICT and educational technologies in times of the COVID-19 pandemic. International Journal of Emerging Technologies in Learning, 17(22): 186–201. https://doi.org/10.3991/ijet.v17i22.27891

[15] Mohammed, L. A., Aljaberi, M. A., Amidi, A., Abdulsalam, R., Lin, C. Y., Hamat, R. A., & Abdallah, A. M. (2022). Exploring factors affecting graduate students' satisfaction toward e-learning in the era of the COVID-19 crisis. European Journal of Investigation in Health, Psychology and Education, 12(8): 1121–1142. https://doi.org/10.3390/ejihpe12080079

[16] Katsaris, I., & Vidakis, N. (2021). Adaptive e-learning systems through learning styles: A review of the literature. Advances in Mobile Learning Educational Research, 1(2): 124–145. https://doi.org/10.25082/AMLER.2021.02.007

[17] Bognár, L., Fauszt, T., & Nagy, G. (2021). Analysis of conditions for reliable predictions by moodle machine learning models. International Journal of Emerging Technologies in Learning, 16(6): 106–121. https://doi.org/10.3991/ijet.v14i19.11177

[18] Ulfa, S., & Fatawi, I. (2021). Predicting factors that influence students' learning outcomes using learning analytics in online learning environment. International Journal of Emerging Technologies in Learning, 16(1): 4–17. https://doi.org/10.3991/ijet.v16i01.16325

[19] Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. IEEE Transactions on Learning Technologies, 10(1): 17–29. https://doi.org/10.1109/TLT.2016.2616312

[20] Ljubobratović, D., & Matetić, M. (2019). Using LMS activity logs to predict student failure with random forest algorithm. In 2019 7th International Conference on Future of Information Sciences: Knowledge in the Digital, Zagreb, Croatia, pp. 113–119. https://doi.org/10.17234/INFUTURE.2019.14

[21] Thi-Diem Nguyen, A. (2021). Using machine learning to predict the low grade risk for students based on log file in moodle learning management system. International Journal of Computing & Digital System, 10(1): 1134–1140. https://doi.org/10.12785/ijcds/110191

[22] Maraza-Quispe, B., Damian Valderrama-Chauca, E., Henry Cari-Mogrovejo, L., & Milton Apaza-Huanca, J. (2021). Predictive model of student academic performance from LMS data based on learning analytics. In 2021 13th International Conference on Education Technology and Computers (ICETC), Wuhan, China, pp. 13–19. https://doi.org/10.1145/3498765.3498768

[23] Mi, H., Gao, Z., Zhang, Q., & Zheng, Y. (2022). Research on constructing online learning performance prediction model combining feature selection and neural network. International Journal of Emerging Technologies in Learning, 17(7): 94–111. https://doi.org/10.3991/ijet.v17i07.25587

[24] Khan, I., Ahmad, A. R., Jabeur, N., & Mahdi, M. N. (2021). Machine learning prediction and recommendation framework to support introductory programming course. International Journal of Emerging Technologies in Learning, 16(17): 42–59. https://doi.org/10.3991/ijet.v16i17.18995

[25] Yahaya, C. A. C., Yaakub, C. Y., Abidin, A. F. Z., Ab Razak, M. F., Hasbullah, N. F., & Zolkipli, M. F. (2020). The prediction of undergraduate student performance in chemistry course using multilayer perceptron. In 2020 6th International Conference on Software Engineering & Computer Systems (ICSECS), Pahang, Malaysia. pp. 1–15. https://doi.org/10.1088/1757-899X/769/1/012027

[26] Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. International Journal of Emerging Technologies in Learning, 14(14): 92–103. https://doi.org/10.3991/ijet.v14i14.10310

[27] Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y. K., & Doneva, R. (2022). Exploring online activities to predict the final grade of student. Mathematics, 10(20): 3758. https://doi.org/10.3390/math10203758

[28] Assami, S., Daoudi, N., & Ajhoun, R. (2022). Implementation of a machine learning-based mooc recommender system using learner motivation prediction. International Journal of Engineering Pedagogy, 12(5): 68–85. https://doi.org/10.3991/ijep.v12i5.30523

[29] Brahim, G. B. (2022). Predicting student performance from online engagement activities using novel statistical features. Arabian Journal for Science and Engineering, 47: 10225–10243. https://doi.org/10.1007/s13369-021-06548-w

[30] Tamada, M. M., Giusti, R., & Netto, J. F. d. M. (2022). Predicting students at risk of dropout in technical course using lms logs. Electronics, 11(3): 468–491. https://doi.org/10.3390/electronics11030468

[31] Evangelista, E. (2021). A hybrid machine learning framework for predicting students' performance in virtual learning environment. International Journal of Emerging Technologies in Learning, 16(24): 255–272. https://doi.org/10.3991/ijet.v16i24.26151

[32] Sathe, M. T., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of students' performance. International Journal of Modern Education & Computer Science, 13(1): 124–137. https://doi.org/10.5815/ijmecs.2021.01.01

[33] Faul, A. C. A concise introduction to machine learning. Boca Raton, FL: CRC Press, 2019. https://doi.org/10.1201/9781351204750

[34] Wei-Meng, L. Python Machine Learning, Hoboken, NJ: Wiley Press, 2019.

[35] Rehman, A., Naz, S., Razzak, M. I., & Hameed, I. A. (2019). Automatic visual features for writer identification: A deep learning approach. IEEE access, 7, pp. 17149–17157. https://doi.org/10.1109/ACCESS.2018.2890810

[36] Saleem, F., Ullah, Z., Fakieh, B., & Kateb, F. (2021). Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning. Mathematics, 9(17): 2078–2100. https://doi.org/10.3390/math9172078

[37] Tamada, M. M., Giusti, R., & de Magalhães Netto, J. F. (2021). Predicting student performance based on logs in Moodle LMS. In 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, pp. 1–8. https://doi.org/10.1109/FIE49875.2021.9637274

[38] Lazarinis, F., Karatrantou, A., Panagiotakopoulos, C., Daloukas, V., & Panagiotakopoulos, T. (2022). Strengthening the coding skills of teachers in a low dropout Python MOOC. Advances in Mobile Learning Educational Research, 2(1): 187–200. https://doi.org/10.25082/AMLER.2022.01.003

[39] Abuhassna, H., Yahya, N., Zakaria, M., Al-Maatouk, Q., & Awae, F. (2021). Guidelines for designing distance learning courses via moodle to enhance student satisfaction and achievements. International Journal of Information and Education Technology, 11(12): 574–582. https://doi.org/10.18178/ijiet.2021.11.12.567

[40] Nuankaew, P., Nasa-Ngium, P., Phanniphong, K., Chaopanich, O., Bussaman, S., & Nuankaew, W. S. (2021). Learning management impacted with COVID-19 at higher education in Thailand: Learning strategies for lifelong learning. International Journal of Engineering Pedagogy, 11(4): 58–80. https://doi.org/10.3991/ijep.v11i4.20337

[41] Tawafak, R. M., ALFarsi, G. M., Jabbar, J., Iqbal Malik, S., Mathew, R., AlSidiri, A., Shakir, M., & Romli, A. (2021). Impact of technologies during COVID-19 pandemic for improving behavior intention to use e-learning. International Journal of Interactive Mobile Technologies, 15(1): 184–198. https://doi.org/10.3991/ijim.v15i01.17847

9 Authors

Chayaporn Kaensar received an M.S. degree in Information Technology from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. She is now an Assistant Professor in the Department of Mathematics, Statistics and Computer at the Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani, Thailand.

Her research interests include Artificial Intelligence (AI), Big Data in Education, Business Intelligence (BI), Object-Oriented Technology, Internet and Wireless Technology, DB, and Programming Languages. (email: chayaporn.k@ubu.ac.th).

Worayoot Wongnin received an M.S. degree in Computer Science from Chulalongkorn University, Bangkok, Thailand. He is now a lecturer in the Department of Mathematics, Statistics and Computer at the Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani, Thailand. His research interests include Machine Learning, Computational Theory, Algorithms, Automata Theory, and Big Data in Education. (email: worayoot.w@ubu.ac.th).

Article submitted 2022-10-06. Resubmitted 2023-03-07. Final acceptance 2023-03-07. Final version published as submitted by the authors.