Research Article https://doi.org/10.12973/eu-jer.11.2.949

European Journal of Educational Research

Volume 11, Issue 2, 949 - 963.

ISSN: 2165-8714

https://www.eu-jer.com/

Student Performance Prediction Model for Predicting Academic Achievement of High School Students

Pratya Nuankaew*
University of Phayao, THAILAND

Wongpanya Sararat Nuankaew
Rajabhat Mahasarakham University, THAILAND

Received: October 11, 2021 ▪ Revised: December 8, 2021 ▪ Accepted: February 19, 2022

Abstract: Modern technology is necessary and important for improving the quality of education. While machine learning algorithms to support students remain limited. Thus, it is necessary to inspire educational scholars and educational technologists. This research therefore has three main targets: to educate the holistic context of rural education management, to study the relationship of continuing education at the upper secondary level, and to construct an appropriate education program prediction model for high school students in a rural school. The data for research is the academic achievement data of 1,859 students from Manchasuksa School at Mancha Khiri District, Khon Kaen Province, Thailand, during the academic year 2015-2020. Research tools are separated into 2 sections. The first section is a basic statistical analysis step, it composes of frequency analysis, percentage analysis, mean analysis, and standard deviation analysis. Another section is the data mining analysis phase, which consists of discretization technique, XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest), confusion matrix performance analysis, and cross-validation performance analysis. At the end, the research results found that the reasonable distribution level of student achievement consisted of four clusters classified by academic achievement. All four clusters were modeled on predicting academic achievement for the next generation of students. In addition, there are four success models in this research. For future research, the researcher aims to develop an application to facilitate instruction for learners by integrating prediction models into the mobile application to promote the utilization of modern technology.

Keywords: *Data science in education, educational data mining, learning analytics, learning strategies, lifelong learning.*

To cite this article: Nuankaew, P., & Nuankaew, W. S. (2022). Student performance prediction model for predicting academic achievement of high school students. *European Journal of Educational Research*, *11*(2), 949-963. https://doi.org/10.12973/eu-jer.11.2.949

Introduction

The importance of human development is to provide quality education to learners. The key principle for raising the quality of education is to promote learners, apply knowledge, and use innovative technologies that are appropriate for learners (Almetov et al., 2020; Prudnikov, 2020). In addition, the characteristics and limitations of learners' learning behaviors create learning styles that produce different educational achievements (May et al., 2014; Onah & Sinclair, 2017; Pappas & Drigas, 2016). However, there are also different perspectives and perceptions about engineering education (Almetov et al., 2020). It is in principle linked to science, technology, engineering, and mathematics (STEM) theory. STEM is the study of science, technology, engineering, and mathematics in all levels of education. It often serves as the knowledge for science and technology education at the university level (Kersanszki & Nadai, 2020). Engineering education should not aim and promote learning only in engineering but should lead the scientific and engineering process to develop a body of knowledge in educational management.

New technologies and innovations that have arisen violently create a tide of change for every academic area with regards to machine learning algorithms and artificial intelligence technology. It has influenced many dimensions, for example Health Technology Era, Innovation for Business and Society, and analyze learners' instincts and learning behaviors with disruptive technology. However, in the education sector, there is still little impact on attention (Li et al., 2020). It discusses the meaning of technology's role in improving human teaching by using engineering processes and systems to facilitate learners. It focuses primarily on hardware. In later eras, machine learning technology was seen as a support to build learners' knowledge through testing and learning management systems known as intelligent tutoring system (ITS). It also has a lot of ITS that researchers are interested in (Onah & Sinclair, 2017; Pappas & Drigas, 2016).

* **Corresponding author:**
Wongpanya Sararat Nuankaew, Faculty of Information Technology, Rajabhat Mahasarakham University, Maha Sarakham, Thailand.
🖂 wongpanya.nu@rmu.ac.th

However, the current primary focus of educational technologists focuses on learning behaviors and learning styles, by promoting the quality of learning according to the interests and needs of the learners (Nuankaew & Nuankaew, 2019). Therefore, this research is an intense study on the development of learning styles that are appropriate and consistent with the learner's behavior. There are three goals: The first goal was to educate the students' context of the lower and upper secondary levels at a rural school in Thailand. The second goal was to study the relationship of continuing education at the upper secondary level of a rural school in Thailand. Finally, the third goal was to construct an appropriate education program prediction model for high school students in a rural school in Thailand.

The research scope is limited to a rural school from Khon Kaen Province, Thailand: Manchasuksa School at Mancha Khiri District. Manchasuksa School currently offers two levels comprising of the lower and upper secondary level. The data used in the research were the academic results of students from the academic year 2015-2020, which had a total of 1,859 students, as detailed in Table 1.

Table 1. Student Information at Manchasuksa School in the Academic Year 2020

Educational Level	Gender		Total	Classroom
	Female	Male		
Lower Secondary				
Mathayom 1	213	183	396	10
Mathayom 2	179	187	366	10
Mathayom 3	173	140	313	10
Total Lower Secondary:	*565*	*510*	*1,075*	*30*
Upper Secondary				
Mathayom 4	93	167	260	8
Mathayom 5	118	131	249	8
Mathayom 6	95	180	275	8
Total Upper Secondary:	*306*	*478*	*784*	*24*
Total:	*816*	*1,043*	*1,859*	*54*

Table 1 shows the student information at Manchasuksa School in the academic year 2020. Overall, Manchasuksa School had more male students than female students. The ratio is 1,043 males to 816 females, or approximately 56.11% per 43.89%. It has a 16.93% reduction in the proportion of continuing education. In addition, the school has managed 54 classrooms. Whereas the school divides the classrooms into two levels, which are classified into 30 classrooms in the lower secondary level, and 24 classrooms in the upper secondary level. Full details and the number of samples used in the research are presented in Tables 2 to Table 4.

Additionally, the research prepared an outline that contain of 5 sections: The first section is to provide an overview of research needs and global issues. The second section describes the scientific research process in the research materials and research methods. The third section is the research results from the research framework that has been designed. The fourth section is an analysis of the research results. Finally, the last section is to summarize the research results and present research guidelines in the next work. In the end, the researcher firmly believed that this study will be of immense benefit.

Methodology

Research materials and research methods are carried out according to the scientific procedure of data mining analysis. It is classified into two layers of data and application: the data mining layer and the user interface layer, according to CRISP-DM: Cross-industry standard process for data mining (Huber et al., 2019; Schröer et al., 2021; Skarpathiotaki & Psannis, 2022). The data mining layer section consists of the first three phases: business understanding, data understanding, and data preparation. While the user interface layer section consists of the last three phases: modeling, evaluation, and deployment.

Business Understanding

The business understanding (BU) stage is the starting point for defining a research solution. While a good understanding of a problem is to create a learning process of the research problem. The main problem of this research is to design a learner's learning model that is appropriate for the social context learners exist in Thailand.

The context of Manchasuksa School is an agricultural community. The main occupation of the people in Mancha Khiri District, Khon Kaen Province, Thailand is agriculture. There are 13,815 farming families, representing 87.94% of the total households. The major cash crops are rice, sugarcane, and cassava. In terms of livestock, farmers have the largest number of cows and buffaloes. The second category is pigs and chickens. Therefore, most of the students of Manchasuksa School have the learning objective according to the laws and regulations of the government to continue living in the community.

From Table 1, data presented to learners in Manchasuksa School in the 2020 academic year, it has been shown that the number of lower secondary school learners was greater than the number of upper secondary school learners (1,075 students per 784 students, approximately 27.07%). Moreover, the vast majority of learners graduating from high school want to apply for jobs in major cities in Thailand, with relatively low prospects for university education. In addition, the selected education program of the learners is mainly based on the recommendations of the guidance teachers. As a consequence, the education program is inconsistent with the learner's potential and does not influence the learner's genuine learning needs for improvement.

In conclusion, understanding this problem provides an opportunity to design an appropriate learning path to improve the learners' academic achievements according to the social context as shown in Figure 1 to Figure 3.

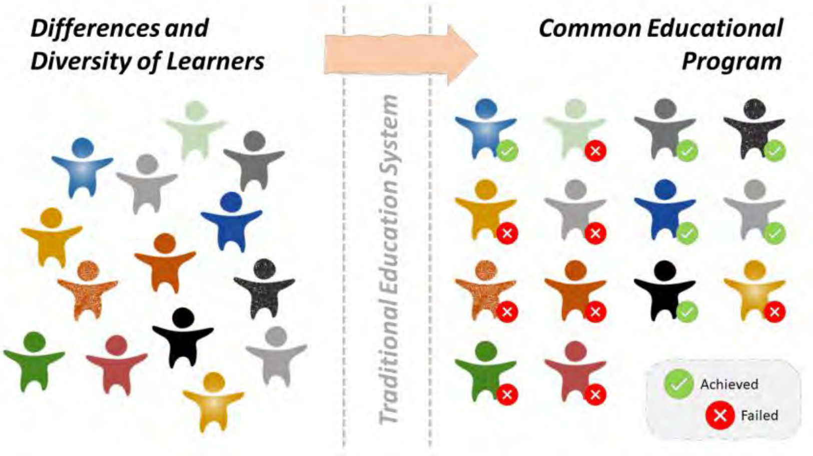

Differences and Diversity of Learners
Traditional Education System
Common Educational Program
Achieved
Failed

Figure 1. Traditional Education System

Figure 1 shows the production of learners according to traditions and methods of creating a body of knowledge at present. Most of the problem is that learners are produced to standards that do not correspond to their potential. Most of the learners' achievement assessments did not meet or pass the minimum threshold. Therefore, the concept of learner development in accordance with their potential is reasonable to support the quality of learners as presented in Figure 2.

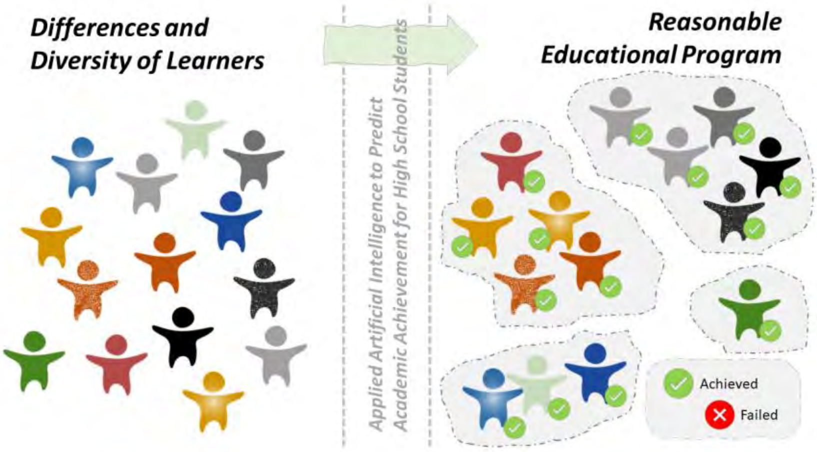

Differences and Diversity of Learners
Applied Artificial Intelligence to Predict Academic Achievement for High School Students
Reasonable Educational Program
Achieved
Failed

Figure 2. Applied Machine Learning Technology to Predict Academic Achievement

Figure 2 illustrates the concept of applying machine learning technology to create alternatives in the management of coherent group education programs as detailed in Figure 3.

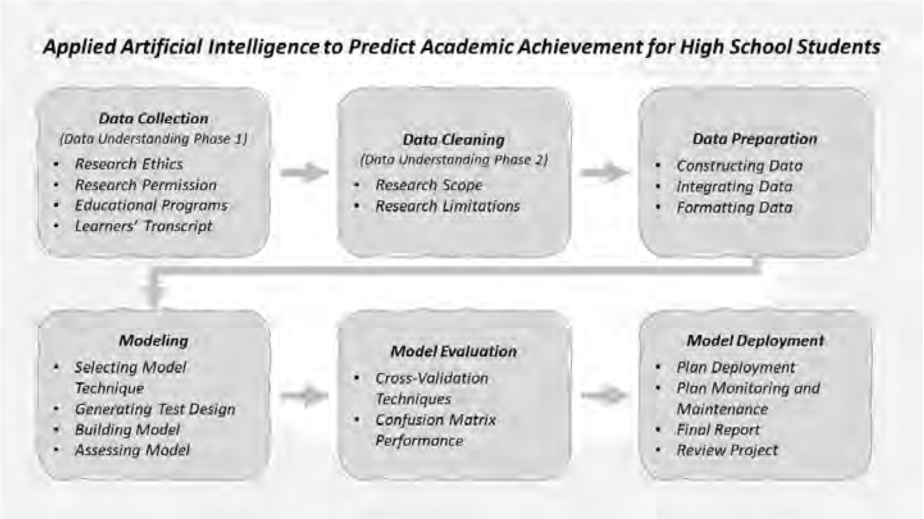


Figure 3. Research Methodology

Figure 3 shows the steps of the concept of applying machine learning technology to create alternatives in the management of coherent group education programs. It consists of six procedures: collecting data, cleaning data, preparing data, modeling, model evaluation, and model deployment which correspond to the concept of data mining development (CRISP-DM). Each step expands the understanding of the different phases of the research methodology.

Data Understanding

The Data Understanding (DU) stage is the phase of understanding the data that corresponds to the research problem. It begins with collecting preliminary data and conducting activities to be familiarized with the data, to identify the integrity and quality of the data, to determine the preliminary data before it detects interesting data, or to find hypotheses for the hidden data.

This section was divided into two phases as shown in Figure 3. The data understanding phase 1 is the collection of data with four sub-steps. The first step is the research ethics. This research had been approved by the University of Phayao as research code: UP-HEC 1.3/045/64. The second step is research permission. This research has been accepted to provide information and conduct research in Manchasuksa Schools in the files https://bit.ly/3kN50Jv, and https://bit.ly/3x4JRgh.

The third step is the education programs. In this section, the researchers studied the education programs of the Manchasuksa School. The secondary level consists of two programs, regular education program, and specialized education program. It is the same at the upper secondary level that consists of two types: regular education program, and specialized education program. Finally, the last step is the learners' transcript. At this step, the researcher was assisted by the Manchasuksa School to collect data through the school's system. Another phase of data understanding is data cleaning. It has two components. The first component is the research scope. The data provided by the school is the student data from the academic year 2015 to the academic year 2020. The data obtained are shown in Table 2. While the second component is the selection of data that will be used to develop the model as shown in Table 3.

Table 2. The Data Obtained From Manchasuksa School

Education Level	**Academic Year**					
	2015	*2016*	*2017*	*2018*	*2019*	*2020*
M.1	392/388	394/393	356/352	353/342	386/381	388/383
M.2	400/395	374/364	376/367	339/325	328/321	365/347
M.3	391/386	364/359	353/347	354/344	317/315	314/304
M.4	282/281	243/241	299/291	299/293	275/266	261/248
M.5	289/288	270/266	227/222	281/277	281/281	252/249
M.6	299/299	279/276	264/264	218/218	275/271	275/274
Total	2053/2037	1924/1895	1875/1844	1844/1799	1862/1838	1855/1805

M = Mathayom

Table 2 shows the data provided by Manchasuksa School. It is classified according to education level and academic year. It contains data of students from Mathayom 1 to Mathayom 6 during the academic year 2015-2020. Overall, the data obtained is that the number of students has a tendency to decline as observed from the total number of students; in academic year 2015, the 1st semester had 2,053 students, while in the academic year 2020, semester 1 had 1,855 students; 198 dropped which was approximately 9.64 percent. However, the data used in the research to analyze the model were data of learners who continued their education from Mathayom 1 to Mathayom 6 as shown in Table 3.

Table 3. Data Gathering

Education Level	**Academic Year 2015 – 2020**					
	Program 1*		***Program 2*****		***Overall***	
	S_1	***S_2***	***S_1***	***S_2***	***S_1***	***S_2***
Mathayom 1	35	35	357	353	392	388
Mathayom 2	35	35	339	329	374	364
Mathayom 3	33	32	320	315	353	347
Mathayom 4	32	32	267	261	299	293
Mathayom 5	31	30	250	251	281	281
Mathayom 6	30	30	245	244	275	274

S = Semester, Program 1* = Specialized Education program, Program 2** = Regular Education program

Table 3 shows the data used in the research. It consists of two types of education programs: regular education program, and specialized education program. The regular education program data consisted of 244 students who graduated from Mathayom 6, and the specialized education program data consisted of 30 students who graduated from Mathayom 6.

Please note that there are only 211 students who have continued their studies from Mathayom 1 and completed their Mathayom 6 from the regular education program. Therefore, the analysis in every model for the regular education program used only 211 samples. At the same time, only 20 students enrolled in specialized education programs continued their education from Mathayom 1 and completed their Mathayom 6. Thus, the analysis in every model for the specialized education program used only 20 samples. All these analyzes were presented in the research report section. Likewise, the structure of the education program at each level of each program is shown in Table 4.

Table 4. The Structure of Each Education Program

Education Level	**Academic Year 2015 – 2020**							
	Program 1*				***Program 2*****			
	S_1		***S_2***		***S_1***		***S_2***	
	n_Course	***n_Credit***	***n_Course***	***n_Credit***	***n_Course***	***n_Credit***	***n_Course***	***n_Credit***
Mathayom 1	17	16.00	15	13.50	18	15.50	15	13.50
Mathayom 2	17	16.00	17	16.00	14	13.50	14	13.50
Mathayom 3	16	16.00	16	16.00	15	13.50	15	13.50
Mathayom 4	21	16.50	20	16.50	20	16.00	19	16.00
Mathayom 5	19	16.50	19	16.50	18	16.00	18	16.00
Mathayom 6	19	16.50	19	16.50	20	16.00	19	16.00
Total:	109	97.50	106	95.00	105	90.50	100	88.50

S = Semester, Program 1* = Specialized Education program, Program 2** = Regular Education program

Table 4 concludes that the first education program has 215 courses, 192.50 credits, while the second education program has 205 courses, 179.00 credits. The first education program had 10 more courses than the second education program, it is a 4.65 percent difference. While the difference in credits is 13.50 credits, it is a 7.01 percent difference. It was evident that the first education program studied harder than the second education program.

However, the overall outcome of the first education program was the high-level of the second education program, while the researcher discussed and presented the information on the topic of data preparation further.

Data Preparation

The Data Preparation (DP) phase is to manipulate data in preparation for model development. It covers the entire process of converting the initial raw data to the final dataset for modeling. It includes elements such as data selection, data cleanup, data generation, data merging, and data transformation. The data provided herein is described in Table 3 and Table 4. They are the individual grades by course and semester, with the raw data gathered, coding files, and other documents is shared on the link: https://bit.ly/3CGBE4X. The examples of the data collected is shown in Table 5.

Table 6. Samples of Data Gathering

No.	S_ID	Credit	GPA	C_1	C_2	C_3	...	Course_n
1	58xxxx01	16	3.85	3.50	4.00	3.50	...	n
2	58xxxx02	16	3.60	3.50	4.00	2.50	...	n
3	58xxxx03	16	2.82	4.00	2.00	3.00	...	n
4	58xxxx04	16	2.64	3.50	3.00	2.00	...	n
5	58xxxx05	16	3.87	4.00	4.00	2.50	...	n
6	58xxxx06	16	3.82	4.00	4.00	2.50	...	n
7	58xxxx07	16	3.93	4.00	4.00	3.00	...	n
8	58xxxx08	16	3.89	4.00	4.00	2.00	...	n
9	58xxxx10	16	0.999	1.00	2.00	3.00	...	n

S_ID = Student ID, Program 1* = Specialized Education program, Program 2** = Regular Education program

With concerns on the privacy of the data, the researcher conducted the research with regards to the ethics procedure by the University of Phayao: Code ID UP-HEC 1.3/045/64. The data used in the analysis cannot be referred to an individual. It will not have any impact on the contributors.

Modeling

At the heart of the Modeling (M) phase is to create a reasonable model for use in solving research problems. Usually, this phase is implemented closely during data preparation. This is due to the selection of good tools in relation to the objectives and data that the research can be collected. There are generally four steps in modeling: select modeling technique, generate test design, build model, and assess model. All four processes were performed in this research.

The tools used to develop the model include the discretization technique, and the XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest). Discretization technique is the process of manipulating continuous data as discrete data to define as a feature. It also aims to optimize data mining algorithms. The discretization method criterion contains of three parts. The first part is an accuracy rate, which is an assessment of the learning and testing accuracy of classifiers used to predict and model results. The second part is the number of intervals, which is an estimate of the number of data intervals obtained after the intervals were split. An important goal is that the data range must be appropriate and describe the characteristics of the data group. The last part is the time require. Whereas the higher the number of data intervals, it will result in more time spent learning. On the other side, if the number of data intervals is small, it takes less time to learn.

An application of discretization techniques in research is to limit the distribution of student achievement. The reason is that the grade point average (GPA) is sparsely distributed. Therefore, the researcher defined the discrete of the data in four characteristics according to the regulations of the educational institutions: excellent, good, normal, and failed as detailed in Table 6.

Table 6. Four Characteristics of Student Achievement

Category	GPA	Meaning
Excellent	3.01 – 4.00	Learners have achieved academic performance. Learners are also in the highest potential category.
Good	2.01 – 3.00	Learners have achieved academic performance. Learners are also in the high potential category.
Normal	1.01 – 2.00	Learners have achieved academic performance. Learners are also in the normal potential category.
Failed	0.01 – 1.00	Learners did not pass the learning achievement criteria.

The second tool is the XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest). XGBoost (Extreme Gradient Boosting) is an ensemble learning method. Its principle is to use a variety of tools to support analytics, which is multiple learners learning (Chen & Guestrin, 2016; Pan et al., 2022; Zhang et al., 2018). Usually, it is classified into 3 types: Bagging (Bootstrap Aggregating), Boosting, and Stacking.

Bagging technique is creating multiple learners and giving each learner a subset of all the data. After that, set a vote. If there are questions, let's see what most answers or what should be answered on average. With this technique, it can reduce variance and reduce the overfit itself. An example of a model is Random Forest.

Boosting is a sequential learning process where the learner takes the previous learner and takes the "error" modifier to reduce the error from the previous learner. It results in better accuracy than bagging technique but makes it easy to overfit.

Stacking is similar to Boosting, but it divides the learner into several groups and takes all the data for the first group to learn and then combines the "answers" of the first group and then passes them on to the next group.

This research uses a gradient boosted model (Decision Tree, Gradient Boosted Trees, and Random Forest) to apply XGBoost technique in research. A gradient boosted model is a collection of a combination of regression or classification models. Both of these approaches are forward group learning methods that produce predictive results through gradual estimation. Boosting is a flexible nonlinear regression process that improves the accuracy of the decision tree model. While boosting a tree increases the accuracy. But it also reduces the speed and ability of human interpretation. The structure of the model development is demonstrated in Figure 4.

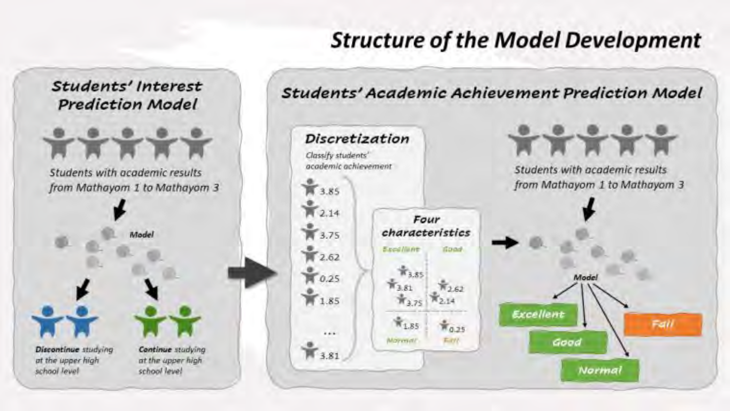

Structure of the Model Development
Students' Interest Prediction Model
Students with academic results from Mathayom 1 to Mathayom 3
Model
Discontinue studying at the upper high school level
Continue studying at the upper high school level
Students' Academic Achievement Prediction Model
Discretization
Classify students' academic achievement
3.85
2.14
3.75
2.62
0.25
1.85
...
3.81
Four characteristics
Excellent
Good
3.85
3.81
3.75
2.62
2.14
1.85
0.25
Normal
Fail
Students with academic results from Mathayom 1 to Mathayom 3
Model
Excellent
Good
Normal
Fail

Figure 4. The Structure of the Model Development

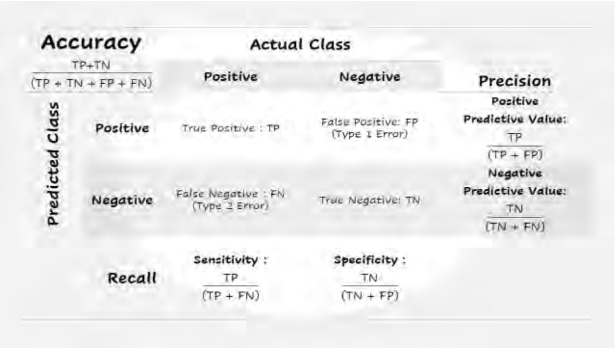
From Figure 4, the model development structure consists of three goals. The first goal was to predict students' interest in further education at the upper high school level from two types of education programs: Regular education program, and specialized education program. The tools used in this section include XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest), confusion matrix performance analysis, and cross-validation performance analysis.

The second goal was to study the distribution of student achievement at the high school level. The tool in this section is discretization technique. The researcher classified four characteristics according to the regulations of the educational institutions: excellent, good, normal, and failed as detailed in Table 6.

The third goal was to predict academic achievement from students interested in entering the upper high school level using the four characteristics analyzed in the second goal. The tools used in this section include XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest), confusion matrix performance analysis, and cross-validation performance analysis.

Evaluation

Evaluation phase (E) aims to test the results or models obtained from the modeling phase. The tools used to test each type of model differ depending on the model being developed. The key components of the evaluation phase contain of 3 parts: evaluate, review, and determine. In this research, the tools used in the model analysis were aimed to determine the efficiency of the models. It consists of two techniques include the confusion matrix performance analysis, and the cross-validation performance analysis. The confusion matrix is an essential tool for measuring machine learning capabilities. It is used to determine the composition and performance of the developed model. The confusion matrix results are concluded in Figure 5.



Accuracy $\frac{TP+TN}{(TP + TN + FP + FN)}$		Actual Class		
		Positive	Negative	Precision
Predicted Class	Positive	True Positive : TP	False Positive: FP (Type 1 Error)	Positive Predictive Value: $\frac{TP}{(TP + FP)}$
	Negative	False Negative : FN (Type 2 Error)	True Negative: TN	Negative Predictive Value: $\frac{TN}{(TN + FN)}$
	Recall	Sensitivity : $\frac{TP}{(TP + FN)}$	Specificity : $\frac{TN}{(TN + FP)}$	

Figure 5. The Confusion Matrix

Figure 5 shows the elements of the confusion matrix. Its main components are classification accuracy, classification precision, and classification recall. Where classification accuracy is the proportion of the number of correctly predicted data divided by the total amount of data. Classification precision is the accuracy of the result in the correct prediction of each category. Classification recall is the accuracy value that the predictive model considers for each category. The equations for calculating the accuracy, precision, and recall values are presented in Figure 5.

Cross-validation (CV) method is one of the essential model testing tools for machine learning that has been developed. The main principle is to divide the data into two parts. The first part is used to develop the model is known as the training dataset. The rest is called the testing dataset. It is used to test a model developed from the data of the first part, as illustrated in Figure 6.

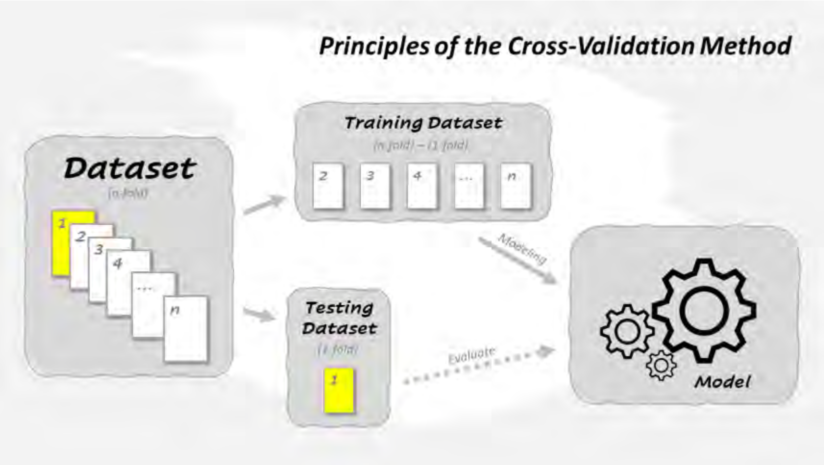


Figure 6. Principles of the Cross Validation Method

Figure 6 shows the principles of the cross-validation method. It represents the process of testing the model being created. The process of cross-validation is to split data into fragment. For example, split the data into 10 fragments (10-fold). Then extract a piece of data (1-fold) to keep for testing the model. From this process it will get the model test results from the actual data available.

Deployment

The phase of deployment (D) is the applying process into applications. It could happen at the conceptual level or lead to a practical approach. It normally composes of 4 main steps: deployment plan, monitoring and maintenance plan, produce report, and project review. Initially, the researcher consulted and planned a research development project with related organizations, comprising administrators of Manchasuksa School, and administrators of the School of Information and Communication Technology, University of Phayao to create research collaborations and apply these findings in the future.

Findings / Results

Research results are classified according to the structure of the model development as explained in Figure 4, which can be summarized in three main points: students' interest prediction model, students' achievement distribution cluster, and students' academic achievement model. It discovered that the special classrooms were not different for learners. While a typical classroom has the variety of success in a variety of educational management processes. The learning achievement of the learners is therefore very important and necessary to study the facts in the behavior of each individual learner. The results of the study appear as follows.

Students' Interest Prediction Model

There are two models results according to two education programs: regular education program, and specialized education program. The tools used in this section include XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest), confusion matrix performance analysis, and cross-validation performance analysis. The results of the analysis are detailed as follows.

Regular Education program

The predictive model of students' interest from the regular education program is summarized in Table 7. It shows analytical results from all eight cross-validation models: 5-fold, 10-fold, 15-fold, 20-fold, 25-fold, 50-fold, 100-fold, and leave-one-out cross-validation. All techniques set a maximal model depth at level 10. Ultimately, the selection result of the most effective model is summed up by selecting from the models with the highest accuracy. It can be summarized as demonstrated in Table 8.

Table 7. Classification Model Analysis

Cross	Decision Tree			Gradient Boosted Trees			Random Forest**		
Validation	Accuracy	Precision*	Recall*	Accuracy	Precision*	Recall*	Accuracy	Precision*	Recall*
5-fold	69.62%	43.55%	30.68%	75.95%	59.09%	44.32%	79.75%	83.33%	34.09%
10-fold	77.22%	63.33%	43.18%	77.53%	61.33%	52.27%	79.75%	81.58%	35.23%
15-fold	75.32%	60.87%	31.82%	79.43%	68.25%	48.86%	80.38%	84.21%	36.36%
20-fold**	75.32%	61.36%	30.68%	75.32%	56.41%	50.42%	**80.70%**	90.91%	34.09%
25-fold	75.10%	58.49%	35.23%	78.80%	64.00%	54.55%	80.06%	85.71%	34.09%
50-fold	73.73%	54.39%	35.23%	79.11%	64.10%	56.82%	80.06%	85.71%	34.09%
100-fold	73.73%	54.90%	31.82%	77.85%	62.16%	52.27%	80.06%	85.71%	34.09%
leave-one-out	72.47%	50.91%	31.82%	76.90%	60.00%	51.14%	79.75%	83.33%	34.09%

* Positive class: Discontinued ** Highest

Table 7 presents the results of classification model analysis. It was concluded that the random forest model at the 10th depth with 20-fold cross-validation was the most accurate. It is equal to 80.70% of accuracy as presented details in Table 8.

Table 8. Model Performance

Accuracy: 80.70%		Actual Class		Class Precision
		True Continued	True Discontinued	
Prediction Class	Pred. Continued	225	58	79.51%
	Pred. Discontinued	3	30	90.91%
Class Recall		98.68%	34.09%	

Table 8 details the most effective model. It has an accuracy rate of 80.70%, a precision rate of 90.91%, and a recall rate of 34.09%. It can be summarized that this model is reasonable for this approach.

Specialized Education program

The predictive model of students' interest from the specialized education program is summarized in Table 9. It shows analytical results from all six cross-validation models: 5-fold, 10-fold, 15-fold, 20-fold, 25-fold, and leave-one-out cross-validation. All techniques set a maximal model depth at level 10. In the end, the selection result of the most effective model is summed up by selecting from the models with the highest accuracy. It can be summarized as illustrated in Table 10.

Table 9. Classification Model Analysis

Cross	Decision Tree**			Gradient Boosted Trees			Random Forest		
Validation	Accuracy	Precision*	Recall*	Accuracy	Precision*	Recall*	Accuracy	Precision*	Recall*
5-fold***	**66.67%**	66.67%	30.00%	42.86%	33.33%	33.33%	48.57%	28.57%	13.33%
10-fold	51.43%	33.33%	13.33%	45.71%	38.89%	46.67%	48.57%	28.57%	13.33%
15-fold	45.71%	30.00%	20.00%	57.14%	50.00%	53.33%	54.29%	44.44%	26.67%
20-fold	48.57%	28.57%	13.33%	57.14%	50.00%	53.33%	48.57%	33.33%	20.00%
25-fold	48.57%	28.57%	13.33%	54.29%	46.15%	40.00%	42.86%	22.22%	13.33%
leave-one-out	48.57%	28.57%	13.33%	57.14%	50.00%	40.00%	45.71%	25.00%	13.33%

* Positive class: Discontinued ** Highest

Table 9 presents the results of classification model analysis. It was concluded that the decision tree model at the 10th depth with 5-fold cross-validation was the most accurate. It is equal to 60.00% of accuracy as presented details in Table 10.

Table 10. Model Performance

Accuracy: 66.67%		Actual Class		Class Precision
		True Continue	*True Discontinue*	
Prediction Class	*Pred. Continue*	15	6	71.43%
	Pred. Discontinue	5	6	54.55%
Class Recall		75.00%	50.00%	

Table 10 details the most effective model. In addition, Table 9 presents a model with 10th depth and tested with 5-fold cross-validation, which is most efficient. It has an accuracy rate of 66.67%, a precision rate of 66.67%, and a recall rate of 30.00%.

Students' Achievement Distribution Group

From Table 6, the results of the high school students' academic achievement are demonstrated in Table 11 and Table 12, where Table 11 shows the summary results from the regular education program, and Table 12 shows the summary results from the specialized education program.

Table 11. Students' Achievement From Regular Education Program

	Level of Students' Achievement			
	Excellent	**Good**	**Normal**	**Failed**
Quantity	95	86	25	5
Percentage	45.02%	40.76%	11.85%	2.37%
Mean of GPA	3.50	2.56	1.69	0.62
S.D.	0.2502	0.3002	0.2025	0.2785

Table 11 shows the students' achievement classified by academic achievement from regular education program. It was found that most of the students were in the highest achievement group or the excellent cluster. There were 95 students, representing approximately 45.02% of the total. The second was the high-achieving group or the good cluster. The number of members is 86 students, approximately 40.76%. While the last group was the failed cluster, there were 5 students in this group: approximately 0.62%. The results in this section were used to predict the academic achievement of junior students.

Table 12. Students' Achievement From Specialized Education Program

	Level of Students' Achievement			
	Excellent	**Good**	**Normal**	**Failed**
Quantity	19	1	0	0
Percentage	95.00%	5.00%	0	0
Mean of GPA	3.72	2.33	0	0
S.D.	0.2208	-	0	0

Table 12 shows the students' achievement classified by academic achievement from the specialized education program. It was found that most of the students were in the highest achievement group or the excellent cluster. There were 19 students, representing approximately 95.00% of the total. The second was the high-achieving group or the good cluster. It has only 1 member, approximately 5.00%.

Students' Academic Achievement Prediction Model

This part is to develop a model for predicting academic achievement of high school students using grade data from the lower secondary school level, as concluded with the conceptualization in Figure 4. The results of the analysis are as follows.

Regular Education Program

The students' academic achievement prediction model from the regular education program is summarized in Table 13. It shows analytical results from all three cross-validation models: 5-fold cross-validation, 10-fold cross-validation, and leave-one-out cross-validation. Ultimately, the selection result of the most effective model is summed up by selecting from the models with the highest accuracy. It can be summarized as detailed in Table 14.

Table 13. Classification Model Analysis

Depth	**k-Fold**	**Criterion of Model Performance**			
		IG	**GR**	**GI**	**Accuracy**
2	5	61.14%	53.12%	63.06%	60.23%
2	10	63.92%	53.12%	65.45%	59.70%
2	leave-one-out	65.40%	54.50%	65.40%	65.40%
3	5	56.35%	56.41%	62.12%	58.74%
3	10	61.62%	52.60%	63.98%	61.13%
3**	leave-one-out	64.93%	54.03%	72.51%**	70.62%
4	5	56.82%	55.91%	57.39%	55.51%

Table 13. Continued

Depth	k-Fold	Criterion of Model Performance			
		IG	GR	GI	Accuracy
4	10	55.89%	54.91%	64.94%	54.89%
4	leave-one-out	62.09%	54.50%	67.77%	65.40%
5	5	51.72%	55.91%	54.52%	58.76%
5	10	49.22%	52.62%	59.31%	57.40%
5	leave-one-out	58.29%	54.03%	61.14%	63.03%
6	5	56.89%	60.66%	60.66%	50.24%
6	10	53.14%	57.32%	59.24%	58.35%
6	leave-one-out	59.72%	54.50%	61.61%	62.56%

* IG = information_gain, GR = gain_ratio, GI = gini_index ** Highest

Table 13 illustrates the classification model analysis. It was found that the model at the 3rd depth of the model with leave-one-out cross-validation from gini_index criterion is the highest accuracy. It is equal to 72.51% of accuracy. The description of the analysis according to the confusion matrix process is presented in Table 14.

Table 14. Model Performance

Accuracy: 72.51%		**Actual Class**				**Class Precision**
		True Excellent	***True Good***	***True Normal***	***True Failed***	
Prediction Class	Pred. Excellent	76	16	3	0	80.00%
	Pred. Good	17	66	11	5	66.67%
	Pred. Normal	2	4	11	0	64.71%
	Pred. Failed	0	0	0	0	0.00%
Class Recall		80.00%	76.74%	44.00%	0.00%	

Table 14 details the most effective model. In addition, Table 13 presents a model with 3rd depth and tested with Leave-one-out cross-validation, it is most efficient. It has an accuracy rate of 72.51%. Moreover, it was found that there were important factors affecting the prediction model in six factors consisting of Mathematics 23201, Social Studies 20231, Social Studies 23104, Thai Language 21101, Art 21101, and Art 23102.

Specialized Education Program

The students' academic achievement prediction model from the specialized education program is summarized in Table 15. It shows analytical results from all three cross-validation models: 5-fold cross-validation, 10-fold cross-validation, and leave-one-out cross-validation. Ultimately, the selection result of the most effective model is summed up by selecting from the models with the highest accuracy. It can be summarized as demonstrated in Table 16.

Table 15. Classification Model Analysis

Depth	k-Fold	Criterion of Model Performance			
		IG	GR	GI	Accuracy
2	5	95.00%	95.00%	95.00%	95.00%
2	10	95.00%	95.00%	95.00%	95.00%
2	leave-one-out	95.00%	95.00%	95.00%	95.00%
3	5	95.00%	95.00%	95.00%	95.00%
3	10	95.00%	95.00%	95.00%	95.00%
3	leave-one-out	95.00%	95.00%	95.00%	95.00%
4	5	95.00%	95.00%	95.00%	95.00%
4	10	95.00%	95.00%	95.00%	95.00%
4	leave-one-out	95.00%	95.00%	95.00%	95.00%

* IG = information_gain, GR = gain_ratio, GI = gini_index ** Highest

Table 15 illustrates the classification model analysis. It was found that the results of the model analysis were not different. Although the researcher showed the model efficacy analysis results in Table 16. It further reinforces that the sample data collected cannot analyze the achievement of the learners in this group.

Table 16. Model Performance

Accuracy: 95.00%		Actual Class				Class Precision
		True Excellent	True Good	True Normal	True Failed	
Prediction Class	Pred. Excellent	19	1	0	0	95.00%
	Pred. Good	0	0	0	0	0.00%
	Pred. Normal	0	0	0	0	0.00%
	Pred. Failed	0	0	0	0	0.00%
Class Recall		100.00%	0.00%	0.00%	0.00%	

From the analysis results of Table 15 and Table 16, it can be said that the researcher should study additional aspects of the learner's behavior in order to develop a model that can distinguish learners and predict their future achievement.

Discussion

The findings in the research were consistent with other studies (Chen & Guestrin, 2016; Zhang et al., 2018). The efficient machine learning method, random forests (RFs) in combination with extreme gradient boosting (XGBoost) can be used to improve the development of predictive models with increased accuracy, which is useful for analyzing student learning behavior. The random forests (RFs) is used to rank the features by importance, which many variables are generated from the knowledge and models of the original decision tree. Therefore, the discussion of the findings is summarized in accordance with the research objectives for maximum benefit. There are three key components: to educate the holistic context of rural education management, to study the relationship of continuing education at the upper secondary level, and to construct an appropriate education program prediction model for high school students in a rural school.

Holistic Context of Rural Education Management

The context of rural society in Thailand is that the majority of the population is agriculturist and engaged in agriculture. It is characterized as a large family and uses family toil as the main labor. Most of the students' parents had relatively low incomes. In addition, most of the parents' education promotion aims to provide their children with a basic level of knowledge. As a result, the promotion of education from parents to students is low. It has been observed that the number of students entering the upper high school level has decreased as demonstrated in Table 1 to Table 4. The number of students enrolled declined every year from 2015 to 2020 as detailed in Table 2. Although graduating students had the highest average of academic achievement as detailed in Table 11, and Table 12.

For this reason, the researcher developed a model to predict the interest in entering high school as presented in the analysis results in Tables 7 to Table 10, and Tables 13 to Table 16. It is discussed further in the next section. However, the researcher firmly believes that if schools can accurately predict the needs and know the characteristics of their learners, they can effectively promote and support education that is consistent with their potential. This is part of the byproduct of this research.

Relationship of Continuing Education

With regards to the studies on the relationship of further education among students in rural schools in Thailand, the researcher applied machine learning technology to create a model for predicting interest in entering high school. The results of the study were summarized as follows and are presented in Table 7 through Table 10.

From Table 7 and Table 9, it must be acknowledged that the findings from both tables can be widely used. Table 7 and Table 9 show a detailed analysis of the model in all dimensions. It consists of an in-depth analysis of the model, the use of cross-validation tools to display various analytical results, and a performance analysis display that enumerates the answers. The results of the analysis are in Table 7 and Table 9. The researcher selected the models with the highest accuracy, and more details are presented in Table 8 and Table 10.

Table 8 presents the model analysis from regular education program. It contains the academic results of students from Mathayom 1 to Mathayom 3 and the records of those who continued upper secondary school (Mathayom 4). The analysis results from Table 8 showed that the random tree model had a high level of accuracy where the selected model had the 10th depth of model using 20-fold cross-validation. It has an accuracy rate of 80.70%, a precision rate of 90.91%, and a recall rate of 34.09%. Please note that the precision rate and recall rate is focused on ignorance of continuing education at the upper high school level.

On the other side, the results of the model analysis from the specialized education program, it had a low level of efficacy, as detailed in Table 9 and Table 10. It has an accuracy rate of 66.67%, a precision rate of 66.67%, and a recall rate of 30.00%. It should therefore be concluded that the model obtained should not be applied to small and special groups of students.

Technology and Tools to Support Education

The final objective of the research is to develop innovations to support the development of students' knowledge. The innovation discovered in this research was the acquisition of a model for predicting the interests and achievement of high school students. The researcher believes that if a pattern or key factor is discovered, those involved will be able to plan and control problems in the future.

The technologies and tools to support the education system of Manchasuksa School discovered in this research were designed and presented in Figure 4. The result was a two-part predictive model that predicted further students' interest and a model predicted the high school academic achievement. While the model obtained from the regular educational program that predicted learning achievement had a high level of accuracy as concluded in Table 14, it had an accuracy rate of 72.51%. In addition, there are also six important factors from this model: Mathematics 23201, Social Studies 20231, Social Studies 23104, Thai Language 21101, Art 21101, and Art 23102.

In contrast to the models obtained from specialized education programs, Although the model has a high accuracy (accuracy rate equal to 95%), in practice it cannot be properly analyzed for learning achievement, as shown in Table 15 and Table 16. However, it is the next effort that the researcher needs to commit on solving this problem. Overall, this research has been successful, and it has achieved all the objectives set forth, with the researcher looking forward on implementing these findings and knowledge into the future. It is consistent with Halperin's work (Halperin, 2020), which draws its conclusions directly to the study's subject matter.

Conclusion

In conclusion, it can be said that this research has achieved all the objectives set. It composed of three important goals: to educate the holistic context of rural education management, to study the relationship of continuing education at the upper secondary level, and to construct an appropriate education program prediction model for high school students in a rural school. The data for research is the academic achievement data of 1,859 students from Manchasuksa School at Mancha Khiri District, Khon Kaen Province, Thailand, during the academic year 2015-2020. Research instruments are divided into two phases. The first phase is a basic statistical analysis step, it consists of frequency analysis, percentage analysis, mean analysis, and standard deviation analysis. Another phase is the data mining analysis phase, which consists of discretization technique, XGBoost classification technique (Decision Tree, Gradient Boosted Trees, and Random Forest), confusion matrix performance analysis, and cross-validation performance analysis.

At the end, the research results found that the reasonable distribution level of student achievement consisted of four clusters classified by academic achievement. All four clusters were modeled on predicting academic achievement for the next generation of students. In addition, there are four success models in this research.

The first model is a model for predicting students' interest in entering the upper high school level from a regular education program. The results of the cross-validation performance test showed that the model had a high level of efficiency, in which the model's accuracy is equal to 80.70%, the model's precision in positive class of discontinued is equal to 90.91%, and the model's recall in positive class of discontinued is equal to 34.09% as concluded in Table 8. While the second model is a model for predicting students' academic achievement in the upper high school level from a regular education program. The results showed that the model had a high level of efficiency, in which the model's accuracy is equal to 72.51% as shown in Table 14. It identified six factors that were significant for predicting academic achievement: Mathematics 23201, Social Studies 20231, Social Studies 23104, Thai Language 21101, Art 21101, and Art 23102.

On the other side, the third model is a model for predicting students' interest in entering the upper high school level from a specialized education program. The results of the cross-validation performance test showed that the model had a low level of efficiency, in which the model's accuracy is equal to 66.67%, the model's precision in positive class of discontinued is equal to 66.67%, and the model's recall in positive class of discontinued is equal to 30.00% as shown in Table 10. Lastly, the fourth model is a model for predicting students' academic achievement in the upper high school level from a specialized education program. It was concluded that the resulting model should not be applied to small, special groups of students. However, it can also be said that it is necessary to improve the model in this section even though it has a high level of accuracy.

Finally, it marks the beginning of efforts to develop technology and offer tools to support the education system for secondary schools in rural Thailand. Therefore, the researcher's future work aims to develop an application to facilitate instruction for learners by integrating prediction models into the mobile application to promote the utilization of modern technology.

Recommendations

Based on the findings of the research, the researchers found that it is of great importance to the study of learners' learning styles. A learning model that uses scientific and analytical technology processes known as "Machine Learning: ML". It has a wide positive effect on the Thai education industry. Therefore, the recommendation for future research in Thai academia is to promote the use of machine learning technology to create new knowledge and to create learning developments that are consistent with each individual for sustainable, stable and lifelong learning.

Limitations

The limitation of this research is that it is during the severe outbreak of the coronavirus disease 2019. As a result, the goal of bringing the found knowledge to schools for the development of learners in an appropriate way has not yet been fulfilled. However, in the future when schools in Thailand are able to resume normal teaching and learning. The research team will proceed according to the resolutions that have been set in every respect.

Acknowledgements

This research project was supported by the Thailand Science Research and Innovation Fund and the University of Phayao (Grant No. FF65-UoE006). In addition, this research was supported by many advisors, academicians, researchers, students, staff, and agencies from two organizations: the School of Information and Communication Technology at the University of Phayao, and the Manchasuksa School at Khon Kaen province, Thailand. The author would like to thank all of them for their support and collaboration in making this research possible.

Authorship Contribution Statement

P. Nuankaew: Conceptualization, design, analysis, and writing. W. S. Nuankaew: Editing/reviewing, supervision, and final approval.

References

Almetov, N., Zhorabekova, A., Sagdullayev, I., Abilhairova, Z., & Tulenova, K. (2020). Engineering education: Problems of modernization in the context of a competence approach. *International Journal of Engineering Pedagogy*, *10*(6), 7–20. https://doi.org/10.3991/ijep.v10i6.14043

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D.Shen & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Halperin, K. (2020). 'Race', parental occupation and academic performance in a public school population. *European Journal of Mathematics and Science Education*, *1*(1), 25–30. https://doi.org/10.12973/ejmse.1.1.25

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Kersanszki, T. L., & Nadai, L. (2020). The position of STEM higher education courses in the labor market. *International Journal of Engineering Pedagogy*, *10*(5), 62–76. https://doi.org/10.3991/ijep.v10i5.13905

Li, X., Zhou, P., Wu, J., Shanthini, A., & Vadivel, T. (2020). Research on artificial intelligence learning system based on psychological knowledge to adjust anxiety and depression. *Behaviour & Information Technology*. Advance online publication. https://doi.org/10.1080/0144929X.2020.1846077

May, D., Wold, K. A., & Moore, S. L. (2014). Developing cultural competencies through transnational learning experiences in active online learning environments. *International Journal of Engineering Pedagogy*, *4*(5), 12–19. https://doi.org/10.3991/ijep.v4i5.3534

Nuankaew, W., & Nuankaew, P. (2019). The study of the factors and development of educational model: The relationship between the learner context and the curriculum context in higher education. *International Journal of Emerging Technologies in Learning*, *14*(21), 205–226. https://doi.org/10.3991/ijet.v14i21.11034

Onah, D. F. O., & Sinclair, J. E. (2017). Assessing self-regulation of learning dimensions in a stand-alone MOOC platform. *International Journal of Engineering Pedagogy*, *7*(2), 4–21. https://doi.org/10.3991/ijep.v7i2.6511

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, *208*, 109520. https://doi.org/10.1016/j.petrol.2021.109520

Pappas, M., & Drigas, A. (2016). Incorporation of artificial intelligence tutoring techniques in mathematics. *International Journal of Engineering Pedagogy*, *6*(4), 12–16. https://doi.org/10.3991/ijep.v6i4.6063

Prudnikov, V. A. (2020). Modern development prospects of programs in engineering education inside national university that use information and communication technologies. *International Journal of Engineering Pedagogy*, *10*(3), 64–81. https://doi.org/10.3991/ijep.v10i3.12619

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Skarpathiotaki, C. G., & Psannis, K. E. (2022). Cross-industry process standardization for text analytics. *Big Data Research*, *27*, 100274. https://doi.org/10.1016/j.bdr.2021.100274

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*, *6*, 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678