ICIC Express Letters
Volume **15**, Number **6**, June **2021**

ICIC International ©2021 ISSN 1881-803X
pp. **567–573**

COURSE PERFORMANCE PREDICTION AND EVOLUTIONARY OPTIMIZATION FOR UNDERGRADUATE ENGINEERING PROGRAM TOWARDS ADMISSION STRATEGIC PLANNING

Sorawee Yanta, Sotarat Thammaboosadee*, Pornchai Chanyagorn and Rojjalak Chuckpaiwong

Faculty of Engineering
Mahidol University
25/25 Salaya, Phuttamonthon, Nakhon Pathom 73170, Thailand
*Corresponding author: sotarat.tha@mahidol.ac.th
sorawee.yat@student.mahidol.ac.th; { pornchai.cha; rojjalak.chu }@mahidol.ac.th

Received October 2020; accepted January 2021

Abstract. *Admission systems around the world are different in characteristics and processes. In Thailand, five different admission rounds affect admission strategic planning, students, and universities. This research proposes the course performance prediction and optimization to predict student performance data and find the optimum criteria for recruiting students in each engineering major respected for the undergraduate engineering program's admission round. The research uses data from undergraduate students in Engineering Faculty in Thailand during 2018-2020. The data preparation methods, such as missing value handling, feature generation, and correlation analysis for each course, are used. Predictive analytics aims to predict three engineering courses' average course grades using the generalized linear model, deep learning, and gradient booted tree. The model is evaluated by using relative error, root mean square error, and absolute error. Gradient boosted tree outperforms the other algorithms, which are 0-0.4% relative error. Prescriptive analytics is consequently used to optimize factors to get the optimum students to the faculty and major by using evolutionary optimization algorithms. This model is used to optimize decision-making in admission strategic planning of Engineering Faculty by optimizing students' number in each major and admission round.*
Keywords: Admission strategic planning, Data science, Course performance, Predictive analytics, Prescriptive analytics

1. **Introduction.** Higher education is crucial to the development of countries and the world [1]. Therefore, universities worldwide always need to improve the educational process for the emerging change in this disruptive world, and one of the essential educational processes is university admission system. In Thailand, the current admission system, called the Thai University Central Admission System (TCAS), has been used since 2018 to ensure every student's seat in the admission system. TCAS has a total of 5 registration rounds consisting of 1) portfolio round, 2) quota round, 3) direct admission round, 4) central admission round, and 5) direct admission by each faculty.

Due to some TCAS rounds' time conflict, students may have trouble in examination preparation for university admission. Moreover, a student in each engineering major has different characteristics and competencies. Thus, course performance prediction (average course grade) and optimization in each course and majors are challenging.

This research proposes the application of data science [2] methodology to create predictive and prescriptive analytics approaching the appropriate amount of recruiting students in each TCAS round to reduce the inappropriate qualification and low-performance students and achieve the highest course performance. The data is collected from Faculty of

DOI: 10.24507/icicel.15.06.567

Engineering in Thailand during 2018-2020. The research outcome will help the universities support decision-making in education process improvement and support strategic planning to lead the universities to world-class universities.

Several data analytics and data science approach to educational process improvement and student performance was conducted to help the institute decide on the educational process [3]. Xu et al. [4] proposed the research to predict students' performance. The method had been constructed into two stages. The first stage was implemented using different machine learning algorithms consisting of linear regression, logistic regression, random forrest, and K-nearest neighbors. The second layer was the ensemble predictor. The result stated that this method gave a superior performance of prediction than the standard approach. The Mean Square Error (MSE) on the prediction results of ensemble prediction in two subjects is 2.22 and 0.16, which are the best performance compared to using single based predictors.

Similarly, Chang et al. [5] used a 2-tuple fuzzy linguistic model and fuzzy analytical process in the research to improve the student selection process by comparing with the arithmetic average method. The result is that these methods gave more accuracy without losing any details.

Moreover, Chakraborty et al. [6] stated that business school faced difficulty selecting the appropriate students to achieve the placement goal after graduating. A novel hybridization of classification trees and artificial neural networks for selecting was proposed to ensure that the institute selected the program's right student. This research predicted the placement of the student after the curriculums according to the student's characteristics. The researchers stated that using this hybridization model gave an accuracy of 91.67%, which is the highest accuracy compared to single machine learning models.

Data science can also give an optimization action to support the business by prescriptive analytics. Khamis et al. [7] had proposed prescriptive analytics to analyze student performance based on Internet usage. The model predicted that student performance and the prescriptive analytics recommended a course of action to achieve the best student performance. This approach aimed to support institutes for decision making and reduce guesswork on admission strategic planning.

The researches mentioned above motivate this research's main objective since many researches deal with educational process improvements and admission strategic planning. Most of the researches are not focused on the Thai University Central Admission System or the engineering school context in Thailand. Interestingly, some research is just the conceptual model and has not been implemented in the entire system, which leads to the initiation of this paper.

2. Literature Review.

2.1. Data analytics.

Data analytics in data science has three main types: descriptive analytics, predictive analytics, and prescriptive analytics. This paper emphasizes only two analytics types.

Predictive analytics [8] analyzes the future outcome of the future or discovers possible input data outcomes. The example algorithms for the prediction are artificial neural network [3], deep learning [10], decision tree [4], regression [10], and gradient boosted tree [11].

Prescriptive analytics [9] delivers the most optimizing action plan to support the organization or business in solving the problem or improving the business to achieve the strategic goals.

2.2. Data science and data science process.

Data science [2] is a science of extracting value, deriving insight, predicting future events, or suggesting the best optimization from data, including analytics solutions and business intelligence. One of the data science

standard processes is the Cross-Industry Standard Process for Data Mining (CRISP-DM) [10]. This standard process template is applied to the methodology of this research.

2.3. **Machine learning algorithm.** In this research, three regression algorithms and one evolutionary optimization algorithm are chosen as follows.

Generalized Linear Model (GLM) [11] is the statistical method that finds the relationship between the dependent and independent variables, which has no error distribution limitations. It is concerned with the extension of regression model series such as linear regression, which becomes the effective data analysis model.

Deep Learning (DL) [12] is a submodel of artificial intelligence to create a large neural network model with multiple layers and is trained by the backpropagation method. Deep learning can be used for classification and forecasting.

Gradient Boosted Tree (GBT) [13] is enhanced from the decision tree [4] to improve the model's efficiency. The gradient boosted tree is processed by randomly creating many decision trees and evaluating the model until achieving the acceptable performance by gradient descent algorithm.

Evolutionary optimization algorithm [14] is the artificial intelligence algorithm to optimize the variables to achieve the best approach based on the genetic algorithm such as natural selection, species migration, or human culture. This research used the evolutionary optimization algorithm in parameter tuning and prescriptive analytics.

3. **Research Methodology.** The overall research methodology is shown in Figure 1.

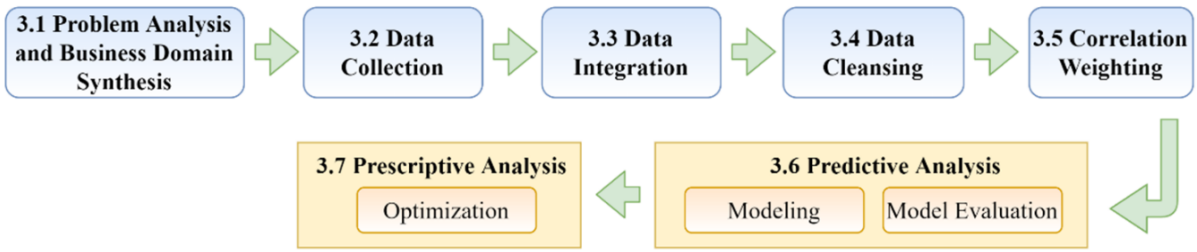

FIGURE 1. Overall research methodology

3.1. **Problem analysis and business domain synthesis.** The business requirement is to analyze the appropriate amount of recruiting students in each TCAS round respected to student majors to increase studying performance and find the relationship between TCAS rounds and student performance concerning first-year GPA.

3.2. **Data collection.** The data used in this research consists of the following.

Student data. This data includes high school GPA, major, admission system round, and student ID, which are encrypted by asymmetric encryption [15] at the source system due to the privacy issue.

Course data. This data contains course ID, course name, instructor name, academic year, and semester. Three courses were analyzed: 1) EGCO111 – Computer Programming, 2) EGIE103- Engineering Materials, and 3) SCPY120 – Physics Laboratory II. The course selection criteria are the number of students and the variety of majors enrolled in each course.

Major. The list of majors of student consists of 1) Computer Engineering (CO), 2) Chemistry Engineering (CH), 3) Industrial Engineering (IE), 4) Electrical Engineering (EE), 5) Mechanical Engineering (ME), 6) Biomedical Engineering (BME), and 7) Civil Engineering (CE).

Student grade and GPA data. This data contains each student's cumulative GPA and grades in each course. Before the aggregation, the total amount of data consists of 1,193 students from seven majors in 3 courses from 2018 to 2020.

3.3. **Data integration.** Once the data is collected, the next process is data integration based on the Extract-Transform-Load process (ETL) [10] to integrate data into the same source and transform it into structured data.

3.4. **Data cleansing and preparation.** This process detects and corrects (or removes) corrupted or inaccurate records from the data.

Duplication eliminating. The research data may be duplicated due to the human input or redundancy of data collection. Student data and course data are needed to be checked by their id, the primary key, or the identifiable attributes.

Manage unmatched values. The values must be mapped into the same format. Since the major name from collected data is various, it must be mapped to one distinct major name.

Feature generation. Three new features had been extracted to be used in the research: 1) the average GPA per section, 2) the variance of GPA per section, and 3) the number of students per section.

3.5. **Correlation weighting.** This research uses correlation weighting [16] to analyze each parameter's relevance and importance to the average course grade's target parameter. Using correlation weighting makes it easier to identify the factor to adjust and improve admission strategic planning in the universities. Equation (1) is the weighted correlation computation, where S_x and S_y are weighted variances of x and y, and S_{xy} is weighted covariance.

$$\rho_{xy} = \frac{S_{xy}}{S_x S_y} \tag{1}$$

3.6. **Predictive analytics.** In predictive analytics, the objective is to predict the average course GPA based on TCAS round, a number of recruitments, major, and high school GPA. The techniques used in the analytics are three algorithms of machine learning consisting of 1) Generalized Linear Model (GLM), 2) Deep Learning (DL), and 3) Gradient Boosted Tree (GBT), which have been widely used in related researches. This predictive analytic delivers the predictive model of the average course GPA. The models are evaluated by three measurements, which are Relative Error (RE), Root Mean Square Error (RMSE), and Absolute Error (AE). The formula of each measurement is shown below:

$$RE = \frac{\sum_{i=1}^{n} |y_i - a_i|}{\sum a_i} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (a_i - y_i)^2}{n}} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{n} |a_i - y_i|}{n} \tag{4}$$

where a_i is the actual output for the ith instance, y_i is the predicted output for the ith instance, and n is the number of instances.

Those performance measurements are evaluated by 10-fold cross-validation [17]. It separated the dataset into ten folds and used one-fold as testing data. Those algorithms consist of some tunable parameters in gradient boosted tree, which consists of depth, number of trees, and learning rate. The parameters will be searched for their optimum values to achieve maximum performance using the evolutionary optimization algorithm [14].

The overall conceptual process of predictive analytics is shown in Figure 2.

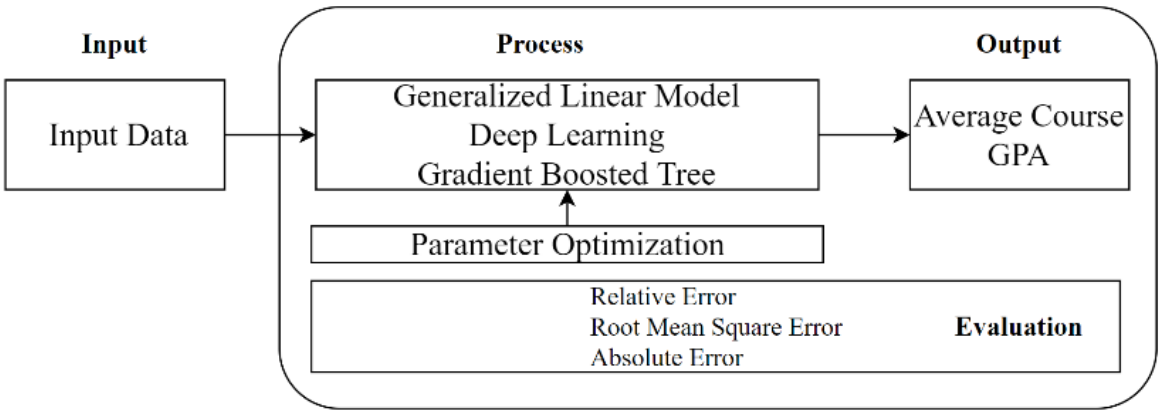

FIGURE 2. Predictive analysis process

3.7. **Prescriptive analytics.** Its objective is to optimize the criteria for achieving the best student performance based on TCAS. The technique is the evolutionary optimization algorithms [14]. In this research, the objective function is to prescribe the appropriate number of students the faculty needs to recruit in each TCAS round to get the maximum course performance. The constants for the objective function consist of TCAS round and student majors.

4. Experimental Results.

4.1. **Correlation weighting.** Figure 3 shows the correlation weighing results for three selected courses. There are six majors left due to the data cleaning process.

	EGCO111	EGIE103	SCPY120		EGCO111	EGIE103	SCPY120
Major = ME	0.49	0.58	0.36	AVG_HighSCH_GPA	0.42	0.05	0.24
Major = IE	0.68	0.42	0.15	TCASRound	0.15	0.30	0.25
AVG_Cum_GPA	0.50	0.14	0.33	TotalStudent	0.26		0.24
VAR_HighSCH_GPA	0.48	0.10	0.38	Major = EE		0.34	0.10
IsTCAS	0.18	0.04	0.64	Major = CII	0.19		0.19
Major = CE	0.41	0.36	0.03	VAR_Cum_GPA	0.16	0.11	0.02
Semester		0.10	0.64	Major = CO			0.07

FIGURE 3. Weight by correlation of EGCO111, EGIE103, and SCPY120

For EGCO111, the most correlated factor is the Industrial Engineering major. The next most correlated factor is the student's average cumulative GPA. For EGIE103, the most correlated factor is the Mechanical Engineering major. The next most correlated factor is the Industrial Engineering major. For SCPY120, the most correlated factor is the enrolled semester. The next most correlated factor is the admission system, which classifies whether the system is TCAS or not.

4.2. **Prediction model.** In the predictive analysis, the machine learning algorithms used to predict average course grade consist of Generalized Linear Model (GLM), Deep Learning (DL), and Gradient Boosted Tree (GBT) with each optimum parameter set. The overall results are shown in Figure 4.

The GBT gives the lowest relative error on every course. Therefore, the best machine learning algorithm to predict the average course grade in each course is GBT. The value of relative error result GBT for EGCO111, EGIE103, and SCPY120 is 0.4%, 0.1%, and 0.0%, which are the lowest error for the analysis GLM gives the highest relative error on every course. Since the GBT gives the lowest prediction error on almost every course, it is selected for optimization.

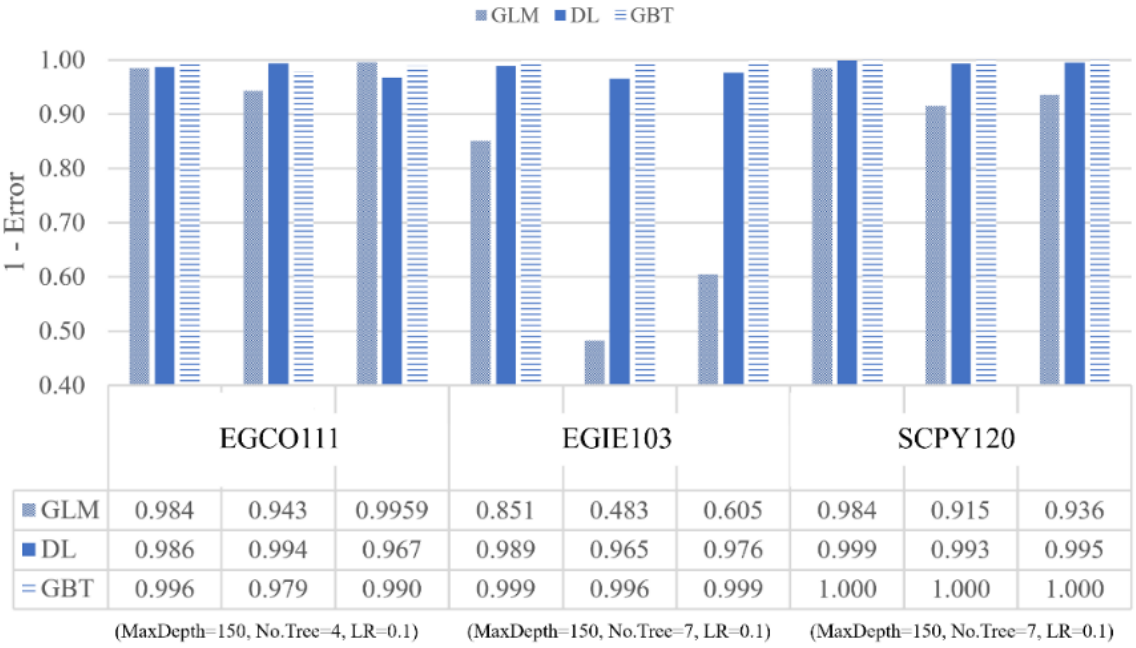

FIGURE 4. Experimental result of predictive models

4.3. **Prescriptive analytics result.** After generating a model from predictive analytics, gradient boosted tree is used to optimize the parameter to get the maximum average course grade in each course. Table 1 shows the average amount of students and the maximum GPA of each course. It stated an appropriate number of students to recruit to achieve the best performance to support admission strategic planning. For example, the highest Chemistry Engineering major's performance is from TCAS 2, and the number of recruits as 13 leads to maximum course performance. However, the result states that students from a Computer Engineering major from every TCAS round can perform well on the physics course. For other majors, the propagation of students to be recruited in each TCAS round is similar for Industrial Engineering and Mechanical Engineering major, which are around 10.

TABLE 1. Average result of the appropriate amount of recruit students in each major

Major	TCAS 1		TCAS 2		TCAS 3		TCAS 4		TCAS 5	
	Amt.	GPA	Amt.	GPA	Amt.	GPA	Amt.	GPA	Amt.	GPA
CE	6	3.79	6	3.79	7	3.71	6	3.77	6	3.78
CH	8	3.73	13	3.85	9	3.84	11	3.75	11	3.75
CO	22	4.00	22	4.00	14	4.00	16	4.00	16	4.00
EE	17	3.41	26	3.41	22	3.31	23	3.34	23	3.34
IE	11	3.84	11	3.85	8	3.78	9	3.79	9	3.80
ME	10	2.97	10	2.98	11	2.99	10	2.99	10	2.99

5. **Conclusion.** This research proposed course performance prediction and optimization for undergraduate engineering programs towards admission strategy. The three courses of the first-year student with the most variety of admission systems and students major enrolled in the course have been selected for the research. The three machine learning algorithms were used in the predictive analysis. The model that gives the highest accuracy is gradient boosted tree. It is selected to use prescriptive analytics to optimize the appropriate amount to recruit students in each admission system and majors and

generate the action plan to support admission strategic planning. However, there is some limitation on this research. Since TCAS has been announced to use since 2018, the data records cover only two years for an analysis that might not effectively collect each TCAS's characteristics each year.

This research can be applied to other institutes like the graduate program and other admission systems to generate the action plan and support decision-making in admission strategic planning. The research can be further developed to probation status prediction and optimization for undergraduate engineering students toward admission strategic planning. It could further generate the entire stream of data analytics and fully support the educational process improvement and gain the highest benefit to students and institutes.

Acknowledgement. This research was supported by "The 60th Year Supreme Reign of His Majesty King Bhumibol Adulyadej Scholarship 2019" from the Faculty of Graduate Studies, Mahidol University.

REFERENCES

[1] J. Johnes, M. Portela and E. Thanassoulis, Efficiency in education, *J. Oper. Res. Soc.*, vol.68, no.4, pp.331-338, 2017.

[2] D. M. Blei and P. Smyth, Science and data science, *Proc. of the National Academy of Sciences of the United States of America*, vol.114, no.33, pp.8689-8692, 2017.

[3] S. Fong, Y. W. Si and R. P. Biuk-Aghai, Applying a hybrid model of neural network and decision tree classifier for predicting university admission, *The 7th International Conference on Information, Communications and Signal Processing (ICICS2009)*, pp.1-5, 2009.

[4] J. Xu, K. H. Moon and M. van der Schaar, A machine learning approach for tracking and predicting student performance in degree programs, *IEEE Journal of Selected Topics in Signal Processing*, vol.11, no.5, pp.742-753, 2017.

[5] K.-H. Chang, K. Chain and M.-T. Chou, Integrating the 2-tuple model and fuzzy analytical hierarchy method to solve higher education student selection problems, *International Journal of Innovative Computing, Information and Control*, vol.11, no.2, pp.733-742, 2015.

[6] T. Chakraborty, S. Chattopadhyay and A. K. Chakraborty, A novel hybridization of classification trees and artificial neural networks for selection of students in a business school, *OPSEARCH*, vol.55, no.2, pp.434-446, 2018.

[7] S. B. Khamis, A. B. H. Ahmad and I. D. Muraina, An overview of using analytics approach to predict Internet usage and student performance in education, *International Journal of Education, Psychology and Counseling*, pp.1-7, 2018.

[8] N. Mishra and S. Silakari, Predictive analytics: A survey, trends, applications, *Int. J. Comput. Sci. Inf. Technol.*, vol.3, no.3, pp.4434-4438, 2012.

[9] D. den Hertog and K. Postek, *Bridging the Gap between Predictive and Prescriptive Analytics – New Optimization Methodology Needed*, 2015.

[10] R. Wirth and J. Hipp, CRISP-DM: Towards a standard process model for data mining, *Proc. of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp.29-39, 2000.

[11] R. H. Myers and D. C. Montgomery, A tutorial on generalized linear models, *J. Qual. Technol.*, vol.29, no.3, pp.274-291, 1997.

[12] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2017.

[13] M. Ebrahimi, M. Mohammadi-Dehcheshmeh, E. Ebrahimie and K. R. Petrovski, Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep learning and gradient-boosted trees outperform other models, *Comput. Biol. Med.*, vol.114, no.5, 2019.

[14] M. Saadatseresht, A. Mansourian and M. Taleai, Evacuation planning using multiobjective evolutionary optimization approach, *Eur. J. Oper. Res.*, vol.198, no.1, pp.305-314, 2009.

[15] M. B. Yassein, S. Aljawarneh, E. Qawasmeh, W. Mardini and Y. Khamayseh, Comprehensive study of symmetric key and asymmetric key encryption algorithms, *Proc. of 2017 Int. Conf. Eng. Technol. (ICET2017)*, pp.1-7, 2018.

[16] J. Ye, Fuzzy decision-making method based on the weighted correlation coefficient under intuitionistic fuzzy environment, *Eur. J. Oper. Res.*, vol.205, no.1, pp.202-204, 2010.

[17] J. H. Kim, Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap, *Comput. Stat. Data Anal.*, vol.53, no.11, pp.3735-3745, 2009.