CTU Journal of Innovation and Sustainable Development

ISSN 2615-9422 | e-ISSN 2815-5602

DOI:10.22144/ctujoisd.2023.038

Predicting graduation grades using Machine Learning: A case study of Can Tho University students

Nguyen Minh Khiem[1*], Huynh Văn Tu[2], and Nguyen Hung Dung[3]

[1]*College of Information and Communication Technology, Can Tho University, Viet Nam*

[2]*Department of Academic Affairs, Can Tho University, Viet Nam*

[3]*FPT University, Viet Nam*

**Corresponding author (nmkhiem@cit.ctu.edu.vn)*

Article info.

Received 28 Jul 2023
Revised 17 Sep 2023
Accepted 30 Sep 2023

Keywords

Academic influenced factor, graduation grade, machine learning

ABSTRACT

A number of factors influence a student's attainment of graduation. Besides scholastic performance within the academic curriculum, other variables such as living circumstances, gender, and choice of major significantly contribute to the probability of achieving graduation. The capacity to forecast academic performance at the time of graduation holds profound importance for universities, especially in discerning the influential factors that contribute to a student's successful completion of their educational pursuits. This study employs multiple machine learning algorithms, including K-nearest neighbor, Neural network, Decision tree, Random forest, and Gradient boosting, to prognosticate the graduation outcomes of 7,837 undergraduate students from Can Tho University during the academic year 2022. These selected students were enrolled in 16 colleges and institutes affiliated with Can Tho University. The efficacy of the employed algorithms was assessed through performance evaluation metrics encompassing accuracy, precision, recall, and F-measure. Furthermore, a 15-fold cross-validation technique was employed for validation. The findings revealed that the Random forest model yielded the most reliable predictions. The factors that significantly impact graduation grades comprise GPA, training point, residential address, college, major, and gender. Based on the experimental findings, these factors were ranked to ascertain their effects on student graduation.

1. INTRODUCTION

The timely graduation of students is not only a concern for the students themselves but also for the university. A high graduation rate and good-quality training will attract numerous potential students to enroll in the future. This contributes to affirming the important role and high ranking of the university (Nick, 2016). There are several factors that impact the learning process and graduation of students, such as institutional support, supervisory practices, and self-management and research skills on postgraduate students' motivation (Priyadarshini et al., 2022). However, this evaluation needs to be conducted on a sufficiently large and long dataset to achieve high reliability. The choice of major also significantly influences graduation outcomes (Alsayed et al., 2021). Roksa and Kinsley (2019) showed that the role of family is also very important in academic achievement of a student at university level (Roksa & Kinsley, 2019).

Machine learning is employed as a novel and powerful technique to forecast the academic outcomes of students (Sekeroglu et al., 2021) by handpicking appropriate parameters and analyzing data. Machine learning algorithms, such as logistic regression, neural networks, and others, are used to compare and identify the optimal algorithm for predicting student performance in higher education institutions (Yakubu & Abubakar, 2022). Importantly, machine learning requires multiple parameters in order to achieve more accurate predictions.

In relation to economic and family factors, students living in off-campus accommodations outside university dormitories face more challenges, as indicated by variables related to demographics (AlHarthi & Kadhim, 2011). The difficulties experienced by students living off-campus include rental fees, transportation costs, and the ability to connect with other students.

Moreover, full participation in classes contributes to timely graduation. A study by (Abdul-Wahab et al., 2019) revealed the challenges faced by students who do not attend sufficient class hours because of reasons such as scheduling conflicts with other teachers, busyness with personal schedules, and seeking additional income through part-time jobs.

Can Tho University (CTU) holds a prominent position as the cultural, scientific, and technical of the Mekong Delta and Viet Nam. Since its establishment in 1966, CTU has been continuously enhancing and expanding its scope. Currently, it enrolls approximately 54,000 undergraduate students. Starting with only a few fields of study, the university has transformed into a comprehensive institution offering a wide range of disciplines. It comprises numerous colleges specializing in areas such as agriculture, aquaculture, technology, education, economics, natural sciences, social sciences. The on-time graduation rate is a prime concern for the university as it reflects the quality and reputation of its educational programs in nurturing the human resource development of the Mekong Delta region. Student information, academic majors, GPA, and disciplinary records are considered as factors that influence students' ability to graduate within the designated timeframe and determine their grades upon graduation.

2. RELATED WORKS

The utilization of Educational Data Mining (EDM) approach for data analysis in education has been addressed in numerous studies (Wook et al., 2017; Kumar & Sharma, 2017) . This technique supports the prediction of students' academic achievements, guidance, and decision-making based on collected data. Mustafa (2022) employed EDM to predict academic performance of students using machine learning algorithms. Using Educational Data Mining for prediction also sheds light on aspects of student retention, meaning students' progression from enrollment to graduation. This reflects the quality, reputation, and ranking of the university in terms of education. Students' inability to graduate can be attributed to various factors such as inappropriate choice of major, marriage, economic obstacles, and language barriers (Al-Mahrouqia et al., 2016). Issues related to educational background, social interaction abilities with peers, and family matters such as parental divorce also influence the psychological well-being and academic progress of students, leading to changes in graduation timelines (Zhu et al., 2022).

Regarding the prediction of student performance, research studies using machine learning have been conducted to support decision-making for management levels. A study by Ibrahim & Al-Barwani (1993), used data from 1511 students to predict the grade point average of first-year students. Research by AlGhanboosi and Kadhim (2004), highlighted the crucial role of academic supervision in student advising, particularly in social sciences departments facing more challenges in advising compared to natural sciences departments. Research by Al-Alawi et al. (2023) employed machine learning to identify factors influencing academic probation, such as gender, estimated graduation year, and cohort. E-learning is also one of the techniques that can support students in reviewing lectures and improving their understanding of course content through self-study (Nguyen et al., 2023).

The contribution of this study can be classified into two parts. First, it utilizes machine learning to predict the graduation grade of students based on information such as gender, demographics, academic majors, program types, GPA scores, courses taken, and extracurricular achievements in the dataset consisted of 7,837 students from 16 faculties and institutes affiliated with Can Tho University. Second, it evaluates the factors associated with student graduation in order to explain their impact on the prediction outcomes. In accordance with the prediction outputs generated by the model, the affiliated educational stakeholders possess the capacity to make the requisite decisions in a manner that aligns with the information provided.

3. MATERIALS AND METHODS

3.1. Dataset

The dataset used in this study was collected from the education department in the year 2022, encompassing data from 16 colleges and institutions within Can Tho University, referred to as education units. The total number of students meeting the graduation requirements is 7,837. Table 1 presents the distribution of students across each education unit, highlighting the School of Economics, College of Engineering Technology, and College of Agriculture as the three largest units with 1,609, 1,313, and 944 graduated students, respectively.

Table 1. Number of graduated students in 2022

No	College/Institution	Number of students
1	Mekong Delta Development Research Institute	22
2	Institute Of Food And Biotechnology	223
3	College of ICT	664
4	School of Foreign Languages	422
5	College of Rural Development	459
6	College of Natural Sciences	285
7	College of Economics	1609
8	School of Law	380
9	School of Political Science (Vietnamese)	133
10	College of Environment & Natural Resources	430
11	College of Agriculture	944
12	College of Engineering Technology	1313
13	College of Aquaculture & Fisheries	354
14	School of Social Sciences and Humanities	370
15	School of Education	200
16	College of Physical Education (Vietnamese)	29

Considering the progression of studies, students may graduate either later or earlier than the standard duration. The dataset includes enrollments from different years ranging from 2014 to 2019. There are 22 hypothesized independent variables, detailed in Table 2, presumed to influence the outcomes – “Graduation grade”. These variables are categorized into three groups: student information, major information, and graduation information.

The grading system comprises four categories: Excellence, Very Good, Good, and Ordinary, determined by the overall academic score and discipline score achieved by students during their training period. Upon graduation, students receive one of two main titles, namely engineer or bachelor, based on the specific training program they have pursued. A small number of veterinary students after graduation will be conferred the title of veterinarian.

3.2. Data cleansing

In order to achieve optimal prediction results, a series of data processing steps were implemented.

– The birthday of each student serves as their distinct identification, and it was converted into a different variable called "age". Since multiple students may have the same age, they were categorized into groups accordingly. These age groups can impact the accuracy of predictions depending on the algorithm being used.

– The registration of residential addresses was segmented into two administrative units: district/ward and province/city. These units will be linked to their respective local unit codes specified by the Ministry of Education and Training.

Several variables that showed negligible impact on the prediction of the graduation grade were removed. These variables include:

– Graduation Decision Number: This variable does not offer any relevant information for determining the graduation grade.

– Student ID: It serves as a unique identifier for each student and does not contribute to predicting the graduation outcome.

– Place of birth of the student: This information, being part of the student's personal profile, does not influence the graduation result.

– ClassID: Each student is assigned to a specific class within their field of study and course, but it does not significantly affect the graduation grade.

– Email: This variable represents the contact information of the student, which is unrelated to the prediction of the graduation grade.

Following the data cleansing process, the resulting dataset contains18 variables and a total number of samples (N) = 7837.

Table 2. List of independent variables

No	Group	Name of variable	Description
1	Student's Information	Ethnicity	1: Kinh ; 0: Others
2		Gender	1: Male; 0: Female
3		Birthday	Birthday of student
4		Student ID	Student identification is issued from enrollment until graduation
5		Birthplace	The birthplace of the student is recorded on the birth certificate.
6		Email	Email of student
7		Place of residence	
8	Major's information	Major	Major of student
9		Educational program	Educational program 1: Full-time training 2: Advanced training 3: High quality training
10		ClassID	Student class identification when attending classes.
11		Type of admission	Type of admission 1: First major 2: Secondary major 3: transferred major 4: Direct admission 5: other
12		AUN assessment	The training program has been assessed according to the AUN standards 0: Not assessed 1: AUN assessment by Vietnamese Ministry of Education and Training 2: AUN assessment by ASEAN University Network
13		Year of enrollment	From 2014 to 2019, depending on the students, they may graduate late or graduate early.
14		Enrollment semester	Semester 1 or Semester 2. Students may either enter regular programs or transfer from other majors
15		Course	From course 38 to course 46, depending on the students, they may graduate late or graduate early.
16		Department	There are 16 department or Institution, numbered as Table 1
17	Graduation's Information	Graduation semester	Semester 1, 2 and 3 in the year 2022
18		Graduation Decision Number	Graduation decision number issued by the Rector of Can Tho University
19		Title	There are 3 titles of graduation student: 1: Engineer, 2: Bachelor and 3: veterinarian
20		GPA	The overall academic score of student, from 2.0 to 4.0
21		Training point	The overall training point of student, from 0 to 100
22		Total credit	The number of credits
23		Graduation grade	This is the outcome variable in prediction. There are 4 types of grade: 1: Excellence, 2: Very Good, 3: Good, 4: Ordinary

3.3. Validation

There are two methods commonly used to partition the dataset for machine learning prediction: simple hold-out, where the data is divided into separate training and testing sets, and k-fold cross-validation. In this study, the hold-out method was employed, with a ratio of 3:1 between the training set and the testing set (equivalent to 5,878 and 1,959 samples, respectively). Additionally, the k-fold cross-validation method was utilized, with a value of k = 15 folds.

In the case of cross-validation tests, the values of k were algorithm-dependent and iterative-fold values. Each fold comprised 7837/k samples. It should be noted that each sample needed to appear once in the testing subset and might be present in both the training and testing subsets during different iterations of the tests.

3.4. Evaluation Accuracy

Assessing the precision and dependability of machine learning algorithms holds significant importance in prediction tasks. However, because of the varying number of students across different grade levels and the substantial magnitude of this disparity (Figure 1), the classes can be considered imbalanced.

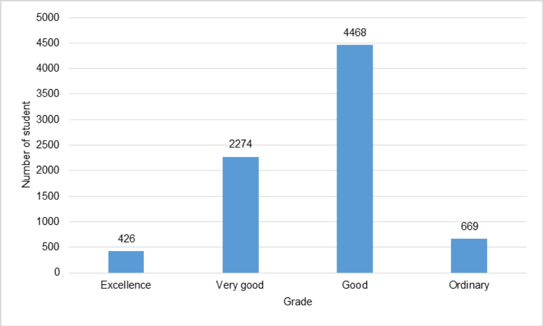

Figure 1. The number of students by grade

In order to ensure fair evaluation for these classes, there are four measurements commonly used to evaluate prediction models, namely accuracy, precision, recall and F-score

– Accuracy: is the fraction of predictions the model got right (Vidiyala, 2023). This value can be understood as the ration between the accurately classified values and the whole dataset. The formula (1) is calculated in terms of positives and negatives of prediction.

$$Accuracy = \frac{TN+TP}{TP+FP+TN+FN} \qquad (1)$$

– Precision which is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (Powers, 2020).

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

– Recall which is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples (Zhang, 2016).

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

– F-score is a measure of a model's accuracy on a dataset which is employed when Precision and Recall are equally important in model (Vidiyala, 2023. The formula of F_1-score as

$$F_1 = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (4)$$

In that, TP = True Positive ; FN = False Negative; TN = True Negative; FP = False Positive

3.5. Machine learning

3.5.1. K- nearest neighbor:

The KNN (K-Nearest Neighbors) algorithm is a versatile technique used for both classification and regression tasks. It operates on the assumption that similar data points tend to be closer to each other (Pedregosa et al., 2011). This algorithm relies on the concept of similarity, which is determined by calculating the distances between data points on a graph, typically using the Euclidean distance measure. When K is set to 1, the object is assigned to its single nearest neighbor. This approach is referred to as "lazy learning" as it makes local approximations without considering global patterns. The impact of neighboring data points on classification is more significant when they are closer in proximity rather than being distant. The runtime of the algorithm depends on the chosen value of K; larger values of K can lead to faster execution.

In our study, we set K to 5 and used the KNN procedure available in the scikit-learn Python package (Pedregosa et al., 2011).

3.5.2. Neural network

The concept of simulating the human brain led to the development of the neural network method (Zou & So, 2008). A neural network comprises multiple nodes or neurons organized in layers. These layers include the input, hidden, and output layers. The input layer receives information from the external environment, while the hidden layer processes internal patterns. The output layer communicates the results back to the outside world. Data is processed within each node using mathematical operations to produce outcomes.

In the hidden layer, every node receives input from the preceding nodes and combines it with weights or coefficients to learn and calculate results for subsequent nodes. These weights can be positive or negative, influencing the input and output data. The interconnectedness and weights of the nodes contribute to the intelligence of the algorithm.

In our research, we utilized the "lbfgs" weight solver, which belongs to the quasi-Newton method family and is implemented in the scikit Python package (Pedregosa et al., 2011). Here, the hidden_layer_size parameter was configured with 3 layers, each comprising 30 neurons. The activation function was set to "tanh," while the solver was set to "lbfgs." For stochastic optimizers, the batch_size was established as 10, determining the size of minibatches. The learning_rate, governing the schedule of weight updates, was set to "constant." The remaining parameters were left at their default values.

3.5.3. Decision Tree

A decision tree employs a tree-like structure to forecast target values based on input variables. It consists of a root node and several internal nodes inputs, with each leaf representing an output. The dataset is categorized into distinct classes. The algorithm's effectiveness relies on its depth, with a deeper tree indicating better training and increased accuracy. To accomplish this, we employed the decision tree from the scikit Python package (Pedregosa et al., 2011). Here, The parameter weights was set to the value “distance”, leaf_size was set to 20, the distance metric was set to “minkowski,” and algorithm was set to “kd_tree.” Other parameters were set to default values.

3.5.4. Random forest

Random forest is a machine learning technique that builds upon the principles of decision trees. In a decision tree, the root and internal nodes serve as inputs, while the leaf nodes represent the outputs or predictions. Random forest constructs a prediction model by randomly selecting samples and utilizing different features to create multiple decision trees. For each tree, a random vector value is determined (Breiman, 2001). The final result is obtained through majority voting among the decision trees. This characteristic makes random forest more favorable and robust compared to using a single decision tree for prediction. The Random forest method falls under the category of bagging techniques, which involve training numerous individual models in parallel. To accomplish this, we employed the Random forest from the scikit Python package (Pedregosa et al., 2011). Here, the min_samples_split parameter, determining the minimum number of samples needed to split an internal node, was configured as 5. The max_depth was also set to 5. As for the other parameters, including min_samples_leaf, min_weight_fraction _leaf, max_features, and min_impurity_split, they were all left at their default values.

3.5.5. Gradient boosting

Gradient boosting, like random forest, is a highly effective machine learning technique that builds upon the principles of decision trees. It falls under the category of boosting techniques, which involve training a sequence of individual models sequentially. Each model learns from the errors or mistakes made by its predecessors.

The objective of this algorithm is to enhance a weak learner and transform it into a strong learner. Over time, through multiple applications and iterations, gradient boosting has been developed and refined (Natekin & Knoll, 2013). In our study, we used gradient boosting and random forest implementations available in the scikit-learn Python package (Pedregosa et al., 2011). Here, the parameter max_depth was set to 5. The loss function used was least squares regression. The parameter min_samples_split was also set to 5. The parameter sub-sample used to control variance and bias was set equal to 1. Other parameters, such as alpha, max_features, and min_impurity_split, were set to their default values.

4. EXPERIMENTAL RESULTS

The accuracy of KNN result is shown in Table 3. The predictive accuracy rates of the hold-out test were 84.4%. Here, the total numbers of true-positive and true-negative cases for each class are 233 for excellence, 322 for very good, 556 for good, and 488 for ordinary.

The cross-validation test was iteratively utilized for 15 folds. The highest accuracy for this test was 82.0% at 6 folds, less than hold-out approach.

Table 3. The result of KNN

Class	Precision	Recall	F1-score
Excellence	0.801	0.815	0.808
Very Good	0.832	0.840	0.836
Good	0.825	0.791	0.808
Ordinary	0.804	0.869	0.837

Neural Network:

The predictive accuracy of the hold-out test of Neural Network algorithm was 80.2%. The precision, recall and F1-score were shown in Table 4. Here, the total numbers of true-positive and true-negative cases for each class are 231 for excellence, 304 for very good, 540 for good, and 455 for ordinary.

Table 4. The result of Neural Network

Class	Precision	Recall	F1-score
Excellence	0.794	0.813	0.803
Very Good	0.786	0.806	0.796
Good	0.801	0.785	0.793
Ordinary	0.750	0.772	0.761

Similar to KNN, the ANN was iteratively tested with 15 folds for cross validation test. The highest accuracy at 5 folds, obtained 77.8%.

Decision Tree

The predictive accuracy of the hold-out test of Decision tree algorithm was 96.8%. The precision, recall and F1-score were shown as in the Table 5.Here, the total numbers of true-positive and true-negative cases for each class are 280 for excellence, 382 for very good, 635 for good, and 577 for ordinary.

Table 5. The result of Decision tree

Class	Precision	Recall	F1-score
Excellence	0.962	0.941	0.952
Very Good	0.987	0.956	0.972
Good	0.942	0.965	0.954
Ordinary	0.951	0.977	0.964

The cross validation test was iteratively employed with 15 folds; the highest accuracy at 8 folds, obtained 95.0%.

Random Forest

Similar to Decision tree, the Random forest algorithm was utilized that obtained an accuracy of 98.2%. The precision, recall and F1-score were shown in Table 6. Here, the total numbers of true-positive and true-negative cases for each class are 280 for excellence, 382 for very good, 635 for good, and 577 for ordinary.

Table 6. The result of Random forest

Class	Precision	Recall	F1-score
Excellence	0.986	0.981	0.984
Very Good	0.972	0.996	0.984
Good	0.990	0.975	0.982
Ordinary	0.957	0.996	0.977

The cross validation test was obtained the highest accuracy at 9 folds, obtained 96.2%.

Gradient Boosting

The Gradient boosting algorithm obtained an accuracy of 94.6%. The precision, recall and F1-score were shown as in the Table 7. Here, the total numbers of true-positive and true- negative cases for each class are 269 for excellence, 366 for very good, 655 for good, and 566 for ordinary.

Table 7. The result of Gradient boosting

Class	Precision	Recall	F1-score
Excellence	0.924	0.932	0.928
Very Good	0.946	0.921	0.933
Good	0.972	0.943	0.957
Ordinary	0.932	0.954	0.943

The cross validation test obtained the highest accuracy at 8 folds, with 92.1%.

Feature importance

To assess the significant contribution of each independent variable to the prediction outcome, these variables are analyzed to determine which ones have a high impact and which ones have a low impact on the result. The Random forest algorithm yields the highest prediction accuracy. Therefore, this algorithm is used to calculate the importance of features. Random forest provides feature importance scores based on reducing the criterion used to select split points. Usually, these scores are based on entropy impurity measurements (Kursa & Rudnicki, 2011). The results show the order of importance of the selected features for prediction (Figure 2).

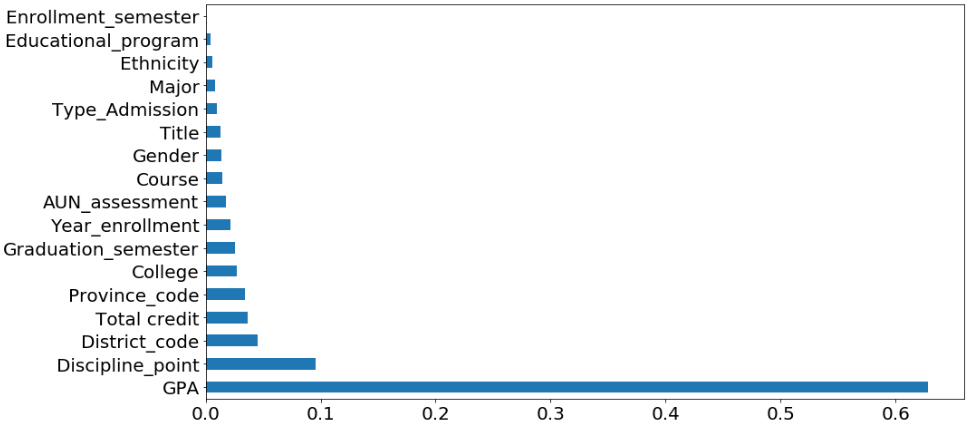

Figure 2. The importance of features

5. DISCUSSION AND CONCLUSION

In this study, machine learning techniques were applied to predict the graduation grade of 7,837 students from 16 faculties and institutes affiliated with Can Tho University. The algorithms employed included K-Nearest Neighbors (KNN), Neural Network, Decision Tree, Random Forest, and Gradient Boosting, yielding prediction results with an accuracy rate above 80%, specifically 84.4%, 80.2%, 96.8%, 98.2%, and 94.6% respectively. Notably, the tree-based algorithms, namely Decision tree, Random forest, and Gradient boosting, outperformed others with a high accuracy rate (>90%). Among them, the Random forest demonstrated the best performance, achieving a 98.2% accuracy rate for the hold-out test and 96.2% for the cross-validation test across 9 folds. The discrepancy between the accuracy rates of the two methods, hold-out and cross-validation, was insignificant. Therefore, the predictive models generated by these algorithms are considered highly reliable.

The Random Forest algorithm is highly accurate and possesses advantageous features. It constructs a tree by randomly selecting a set of K features at each node from a pool of possible trees. This approach allows for efficient generation and the combination of numerous random trees typically results in precise models. In the context of the given dataset, Random Forest demonstrated its predictive power by successfully making predictions for 7,837 students.

This study contributes to predicting the graduation grade of students based on meaningful criteria. Besides GPA, other related information also positively contributes to the prediction, including student information and academic program information. Student information includes residence address, gender, and ethnicity, while academic program includes educational program, AUN-assessment programs, courses, and department. Furthermore, other useful information such as title (engineer and bachelor) and disciplinary scores also significantly influence the graduation classification.

Based on the predictions and the evaluation of the importance of various features, The GPA result mainly determines the grade of graduation of students. Besides that, the training point is a significant factor that influences students' graduation rankings. According to the regulations of Can Tho University, the training point and GPA are criteria for awarding scholarships to students in each semester. This means that students with high GPAs will strive to take part in activities and be actively involved to get high training points and achieve scholarships. This explains why the training point has a positive impact on the prediction. The yearly grades also affect the final graduation outcome of students. When students achieve good results in the current academic year, it motivates them positively for the following years. Throughout the learning process, the results of students fluctuate in different stages. Freshmen usually have higher grades as they study general subjects, which becomes more challenging to achieve high grades when delving deeper into specialized subjects, which require comprehensive knowledge and demanding exams. Demographic information, such as district and province of residence, also influences the ranking outcome. Green & Celkan (2011) also found that demographic characteristics correlate with academic performance. A demographic inventory reveals students' educational backgrounds. Students with good backgrounds tend to have good skills in acquiring knowledge in the university environment, resulting in higher graduation outcomes. Household registration indirectly affects financial support from families for students' education (Solis & Durband, 2015). Students in major cities or economically developed regions may receive better support from their families, while those in rural areas struggle more to afford expenses related to studying, such as part-time jobs, cost-saving for study materials, practical experience costs, and less time dedicated to studying. Therefore, financial support impacts students' graduation outcomes. The graduation timing reflected in the feature “Graduation semester” is also significant. Students who graduate on schedule tended to have higher rankings, such as excellence or very good, compared to those who graduate later. Students who graduate late often have to retake failed courses or face other reasons that delay their learning process. Students who graduate late show less excellence or very good rankings.

Gender plays an important role in the impact on academic achievement. The findings of Zhu et al. (2022) also clearly elucidate the influence of gender on academic performance. Employing machine learning to highlight this factor, it is observed that male students have higher rates of achieving excellent and outstanding grades in engineering subjects, whereas female students excel in social sciences. However, overall, there is a significantly higher proportion of male students graduating with ordinary degrees compared to their female counterparts because of a distinct gender disparity.

According to the ranking results, the college is also an important factor that influences students' graduation grade. Students majoring in engineering and technology are likely to have a harder time achieving higher grades. In fact, the actual data shows that the percentage of students achieving excellence and very good grades in engineering programs is about 10%, while this percentage reaches 15% in social science fields such as economics, law, and journalism. The quality of the training program also plays a significant role in students' learning process. Accredited programs by the ASEAN University Network (AUN) will attract more students to enroll, and the program itself will be enhanced, improving the quality of faculty and lectures. As a result, enrolled students will have access to new knowledge, fostering interest and enhancing their learning abilities.

Additionally, factors such as the year of enrollment, total duration of the course, and so on will have less influence on the prediction results because when students enroll, they already know the study duration of their chosen field and they have the necessary psychological preparation and strategies throughout the learning process.

The achieved outcome of this research is a prediction model, with the best model being the Random forest model using the graduation data of students in 2022. This model will assist managers, especially departments and faculties, in developing strategies and advice for students to help them graduate on time and achieve high classifications. Identifying the factors that impact graduation outcomes will aid managers in planning interventions to improve students' academic performance.

In the future, to enhance prediction results and increase the reliability of the prediction model, we will expand the scope of research to include both the number of graduating students over the years and the number of features. There are many factors that influence student rankings, such as living conditions, whether students are studying their chosen majors, the influence of romantic relationships, and satisfaction levels with lectures. These factors will be collected to construct a more precise model.

REFERENCES

AlHarthi, H., & Kadhim, A. (2011). Predicting the difficulties faced by students living outside the university campus in light of some demographic variables. *Journal of Qualitative Educational Research*, *18*(3), 306–430.

Abdul-Wahab, S. A., Salem, N. M., Yetilmezsoy, K., & Fadlallah, S. O. (2019). Students’ reluctance to attend Office hours: Reasons and suggested solutions. *Journal of Educational and Psychological Studies*, *13*(4), 715–732.

Al-Mahrouqia, & R., Karadsheh, M. A. (2016). Sultan Qaboos University students reasons of being underObservation. *Humanities and Social Sciences*, *43*(3), 2343–2360.

Alsayed, A.O., Rahim, M.S.M., AlBidewi, I., Hussain, M., Jabeen, S.H., Alromema, N., Hussain,S., & Jibril, M. L. (2021). Selection of the right undergraduate major by students using supervised learning techniques. *Appl. Sci.*, *11*, 10639. DOI: 10.3390/app112210639

AlGhanboosi, S., & Kadhim, A. (2004). Problems of Academic Supervision at Sultan Qaboos University from Professors and students perspectives. *Journal of Education*, *10*(2), 39–75.

Al-Alawi, L., Al Shaqsi, J., & Tarhini, A. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Educ. Inf. Technol*. DOI: https://doi.org/10.1007/s10639-023-11700-0

Breiman, L. (2001). Random forests. *Mach. Learn*. DOI: https://doi.org/10.1023/A:1010933404324

Green, L., & Celkan, G. (2011). Student demographic characteristics and how they relate to student achievement. *Procedia - Social and Behavioral Sciences*, *15*, 341-345.

Ibrahim, A., & Al-Barwani, T. A. (1993). A study of Omani secondary school Certificate Examination as a predictor of academic performance of Sultan Qaboos University. *Research in College Teaching Practicum Research in Sultan Qaboos University*, *1*(1), 1–29.

Kursa, M., & Rudnicki, W. (2011). The All Relevant Feature Selection using Random Forest. *Computer science, Artificial Intelligence*. https://doi.org/10.48550/arXiv.1106.5112

Kumar, R., & Sharma, A. (2017). Data mining in education: A review. *International Journal of Mechanical Engineering and Information Technology*, *5*(1), 1843–1845.

Mustafa, Y. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ*. *9*(11). https://doi.org/10.1186/s40561-022-00192-z

Natekin, A., & Knoll, A. (2013). Gradient boosting machines. *Front Neuro-robot*, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Nick, D. (2016). The effect of university attended on graduates' labour market prospects: A field study of Great Britain. *Economics of Education Review*, *52*. DOI: https://doi.org/10.1016/j.econedurev.2016.03.001.

Nguyen, H. D., Duc, T., Sang, V., Diem, N., Hung, N., & Nha, T. T (2023). Knowledge Management for Information Querying System in Education via the Combination of Rela-Ops Model and Knowledge Graph. *Journal of Cases on Information Technology, 25*(1) pp.1-17. http://doi.org/10.4018/JCIT.324113

Priyadarshini, M., Gurnam, K. S., Hoon, T. S., Geethanjali, N., & Fook, C. Y. (2022). Key Factors Influencing Graduation on Time Among Postgraduate Students: A PLS-SEM Approach. *Asian Journal of University Education*, *18*(1), 51-64. DOI: https://doi.org/10.24191/ajue.v18i1.17169.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,… & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830

Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. ArXiv abs/2010.16061.

Roksa, J., & Kinsley, P. (2019). The Role of Family Support in Facilitating Academic Success of Low-Income Students. *Research in Higher Education*, *60*(4), 415–437. http://www.jstor.org/stable/45180388

Solis, O., & Durband, D. B. (2015). Financial support and its impact on undergraduate student financial satisfaction. *College Student Journal*, *49,* 93-105.

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. *Applied Sciences*, *11*(22), 10907

Vidiyala, R. (2023, Jul 18). *Performance metrics for classification machine learning problems*. https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007.

Wook, M., Yusof, Z. M., & Nazri, M. Z. A. (2017) Educational data mining acceptance among undergraduate students. *Educ. Inf. Technol*., *22,* 1195–1216.

Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students'performance in higher educational institutions. *Kybernetes*, *51*(2), 916–934. https://doi.org/10.1108/K-12-2020-0865.

Zhu, Y., Xu, S., Wang, W., Zhang, L., Liu, D., Liu, Z., & Xu, Y. (2022). The impact of online and offline learning motivation on learning performance: the mediating role of positive academic emotion. *Education and Information Technologies*, *27*(7), 8921-8938.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, *4*(11), 218. https ://doi.org/10.21037 /atm.2016.03.37

Zou J, Han Y, & So, S. S. (2008). Overview of artificial neural networks. *Methods Mol Biol*., *458*, 15–23.