DOI: https://doi.org/10.55164/ecbajournal.v17i1.273896

Benchmarking of Machine Learning for Predictive Model for Faculty Selection

Received : 30/04/2024
Revised : 13/07/2024
Accepted : 29/08/2024

Prajak Chertchom[1*]

Abstract

This study employed the Gradient Boosted Trees Machines Algorithm and conducted benchmarking of machine learning techniques for predictive modeling in faculty selection among students in Southern Thailand. The dataset included 12,125 students with variables such as High School GPA, blood group, district, province, and parent background. Key factors influencing model performance encompassed academic history, province of residence, and parental attributes. The Gradient Boosted Trees model achieved an impressive accuracy of 85% and precision of 87%, effectively identifying chosen faculties. Precision and recall metrics were 0. 594 and 0. 460 respectively, with an F1 Score of 0.518, underscoring the model's robustness in predicting student choices. Analysis of the SVM model revealed significant coefficients for features such as " BEFOREGPA" and " BLOODGROUP" , influencing predictions positively or negatively. The SVM model achieved an F1 score of 0. 33, indicating moderate performance in predicting student choices. The outcomes of the Gradient Boosting model demonstrate its effectiveness in predictive tasks, leveraging an iterative tree-building approach to correct errors systematically. However, careful monitoring of model performance is crucial, particularly when significant errors occur, to mitigate potential issues such as overfitting. In addition, from our analysis, it's evident that students' decisions regarding faculty selection are influenced by a complex interplay of various factors. Among these, province of origin and Grade Point Average (GPA) stand out as pivotal determinants shaping students' educational journeys.

Keywords: Faculty Selection, Predictive Modelling, Gradient Boosting, Benchmarking

Introduction

Quantifying faculty selection and predicting future choices are pivotal challenges in educational research. Understanding the factors influencing students' decisions and accurately predicting their choices can significantly enhance universities' recruitment strategies, optimize resource allocation, and tailor offerings to better meet student needs. Despite ongoing efforts, there remains a critical need for thorough and systematic benchmark comparisons of machine learning models designed for faculty selection tasks. Therefore, this study presents a comprehensive set of benchmarking results by applying

[1] Assistant Professor Dr., Faculty of Economics and Business Administration, Thaksin University
* Corresponding author: E-mail address: prajak.c@tsu.ac.th

various deep learning models to our educational dataset and comparing them with state-of-the-art machine learning approaches. Specifically, we benchmark the proposed Gradient Boosted Decision Trees model against other machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest to validate prediction results rigorously. Our aim is to offer a thorough evaluation of these methods, providing actionable insights for enhancing university recruitment and retention strategies. The availability of large educational datasets has significantly accelerated research in this domain, evident from numerous recent publications. Thailand has faced a pronounced decline in its birth rate, leading to a scarcity of university-aged students. This demographic shift has intensified competition among the country's universities, as they vie for a diminishing pool of prospective applicants. According to the Thailand Development Research Institute, the country recorded only 502,000 births in 2022, marking a shortfall of approximately 30% from the targeted 700,000 births (Nation Thailand, 2023).

The data for this study is obtained solely from the registration system of a prominent government university located in Songkhla Province, Thailand. The dataset used in this research consists of records for a total of 12,126 students who registered between 2020 to 2022, as presented in Table 1 and further elaborated in the methodology section. This dataset provides insights into the demographics, academic performance, and family background of students who enrolled in various faculties throughout Southern Thailand.

Table 1 Showing the Distribution of Students across Different Faculties based on their Enrollment Figures by Academic Year.

	Students' Figure by Faculty		
Academic Year/Faculty of	**2562**	**2563**	**2564**
Fine Arts	163.00	185.00	157.00
Economics and Business Administration	802.00	679.00	756.00
Education	866.00	739.00	894.00
Engineering	30.00	39.00	32.00
Humanities and Social Sciences	668.00	627.00	694.00
Agro-and Bio-Industry	29.00	65.00	37.00
INTER	15.00	8.00	13.00
LAW	599.00	668.00	716.00
Nursing	50.00	57.00	45.00
Science	154.00	202.00	189.00
Health and Sports Science	185.00	224.00	224.00
Technology and Community Development	104.00	102.00	80.00
Development College, Thaksin University (TSU-MDC)	227.00	330.00	471.00
	3,892.00	**3,925.00**	**4,308.00**
Total			**12,125.00**

This research project is structured into five primary sections. The first section provides an overview of the research needs and global issues. The second section outlines the scientific research process, along with the materials and methods utilized in the study. The third section presents the research findings obtained from the designed research framework. In addition, this study proposes benchmarking a predictive model for faculty selection among students in Southern Thailand using the Gradient Boosted Decision Trees model alongside other machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest. The proposed model aims to predict students' faculty choices based on their academic records, personal preferences, and demographic information. The model will be trained on a dataset of students, and its performance will be assessed using metrics such as accuracy, precision, and recall. The fourth section provides an in-depth analysis of the research results, including an evaluation of the predictive model's performance. Finally, the last section summarizes the research outcomes and outlines guidelines for future studies. The researcher is optimistic that this study will yield significant benefits to the field

By identifying the factors that are most important to students, universities can develop more effective marketing campaigns and outreach efforts that target these factors. This can lead to better recruitment efforts and higher quality applicants, resulting in higher graduation rates and better outcomes for both the students and the university. Furthermore, this study can help universities better allocate their resources. By knowing which faculties are in high demand and why, universities can allocate resources more effectively to ensure that they are meeting the needs of their students. This can include hiring more faculty members, investing in new facilities and technologies, or developing new programs and courses to meet emerging demand. Finally, this study can contribute to the development of machine learning in education. By exploring the effectiveness of the Gradient Boosted Trees Machines Algorithm in faculty selection, this study can pave the way for future research in educational data mining. It can also lead to the development of new tools and techniques that can improve the quality of education not only in Southern Thailand but also in other regions. (Çakıt & Dağdeviren, 2022; Kamal et al., 2020; Singh & Kaur, 2016).

Objective

1. To develop and benchmark a predictive model using the Gradient Boosted Decision Trees algorithm for faculty selection among students in Southern Thailand, comparing its performance with other machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM).

2. To identify the significant factors that influence the selection of faculty among students in Southern Thailand.

Literature review

Gradient Boosting, also known as the Gradient Boosted Trees Machines Algorithm, is a highly effective machine learning technique that has become increasingly popular in recent times. (Ben-Assuli & Vest, 2022) By combining numerous weak models, primarily decision trees, this algorithm generates a robust prediction model that can accurately predict both regression and classification tasks. It employs a sequential regression structure that uses a low- precision classifier aggregation approach to generate subsequent trees based on the previous tree computation errors, thereby creating an improved classifier. The Gradient Boosted Trees Machines Algorithm has found numerous applications in education, including predicting student performance, identifying at- risk students, and recommending customized learning materials. In recent years, its use in the education field has gained traction. (Patcharacharoenwong et al., 2020)

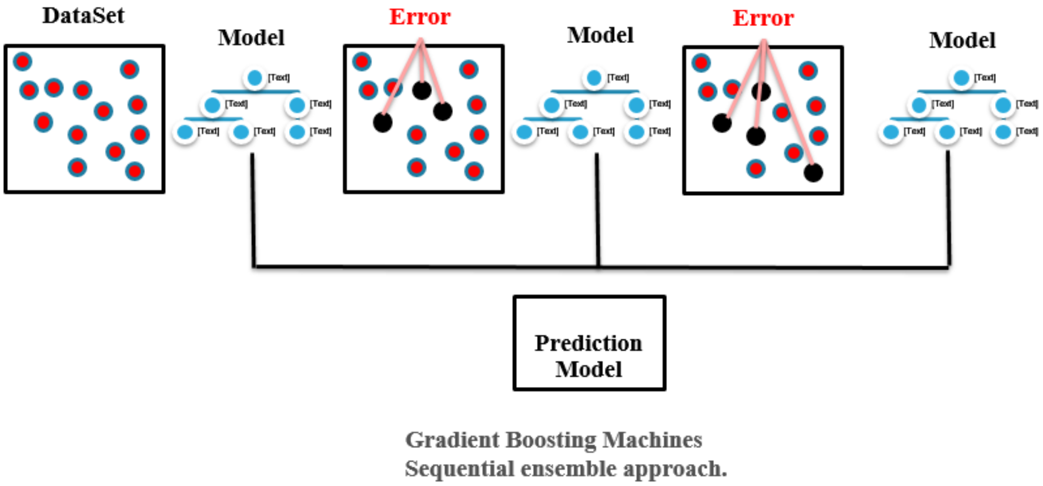


Figure 1 showing Gradient Boosting Machines' Sequential ensemble approach. (Natekin & Knoll, 2013)

We first provide a brief review of machine learning and deep learning models for educational data, and then discuss existing works on benchmarking education datasets. Studies have shown that this algorithm can be used to predict various outcomes in education, such as college graduation, student performance, and student retention. Researchers have developed efficient and powerful machine learning-based frameworks for predicting outcomes in education using Gradient Boosted Trees Machines Algorithm. (Oztekin, 2016; Hutt et al., 2018; Bilquise et al., 2019; Aiken et al., 2020; Hew et al., 2020) The accuracy of the predictive models developed using this algorithm has been verified by comparing their performance with that obtained from other machine learning methods. (Lee et al., 2022) The potential of Gradient Boosted Trees Machines Algorithm in education lies in its ability to provide more accurate and effective predictive models that can be used to improve student outcomes and inform decision- making in educational settings. By analysing data from student demographics, past performance, and other relevant factors, this algorithm can provide insights into student behaviour and learning outcomes, which can inform the development of more effective educational strategies. One notable application of Gradient Boosted Trees Machines Algorithm in education is faculty selection. By analysing factors that influence a student's decision to choose a particular faculty, this algorithm can help universities tailor their programs and offerings to meet the needs and interests of their students. By designing programs and courses that are more relevant and attractive to students, universities can increase enrolment rates and improve retention rates. (Nagy & Molontay, 2018; Ketui, 2019)

Benchmarking experiments are a crucial component of machine learning research, serving to evaluate and compare the performance of different algorithms across standardized datasets. Such experiments help in understanding the strengths and weaknesses of various methods and guide the selection of appropriate algorithms for specific tasks. Benchmarking provides a systematic way to assess the efficacy of machine learning models. It involves running multiple algorithms on the same dataset and comparing their performance using predefined metrics. This process helps in identifying the best-performing models and understanding the scenarios in which certain algorithms excel. According to Fernandez-Delgado et al. (2014), a comprehensive benchmarking study compared 179 classifiers over 121 datasets, revealing insights into algorithm performance variability across different data types. Benchmarking has been extensively used to evaluate various machine learning algorithms. In a study by Kamal & Talbert (2024), multiple machine learning algorithms were benchmarked on classification tasks, showing that ensemble methods, such as boosting and bagging, often outperform single algorithms like decision trees and support vector machines. The study also highlighted the importance of hyperparameter tuning in achieving optimal performance.

This paper also reviews the use of the Gradient Boosted Decision Trees (GBDT) algorithm for faculty selection, comparing its performance with other machine learning algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM).

GBDT is an ensemble learning method that builds multiple decision trees in a sequential manner, where each tree attempts to correct the errors of its predecessor. This method is known for its high predictive accuracy and ability to handle various types of data, making it suitable for complex prediction tasks (Friedman, 2001). In educational data mining, GBDT has been effectively utilized for predicting student performance and enrollment behaviors (Kabra & Bichkar, 2011). K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm that classifies data points based on their proximity to other points in the dataset. Despite its simplicity, KNN can be powerful, particularly for smaller datasets or problems where the relationships between data points are highly local. In the context of educational data, KNN has been applied to classify students based on their academic performance and demographic attributes (Issah et al., 2023). For Support Vector Machines (SVM), it is a robust supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVM is particularly effective in high-dimensional spaces and cases where the number of dimensions exceeds the number of samples. In educational settings, SVM has been employed to predict student success and identify at-risk students (Pallathadka et al., 2023).

Several studies have benchmarked GBDT against KNN and SVM, demonstrating its superior performance in various predictive tasks. For instance, Lv et al., (2023) conducted extensive experiments comparing several algorithms, including GBDT, KNN, and SVM, on different datasets. Their results indicated that GBDT often outperformed KNN and SVM in terms of accuracy and robustness, particularly in complex and noisy datasets. In the specific context of faculty selection, predictive models must account for various factors such as academic records, personal preferences, and demographic information. GBDT's ability to handle diverse and complex features makes it particularly suitable for this task. Studies have shown that ensemble methods like GBDT can capture intricate patterns and interactions between features that simpler models like KNN might miss.

Understanding the factors that influence a student's decision to pursue higher education is crucial for universities and policymakers. Several factors can influence a student's decision, including their profile and family background. Kamal et al., (2020) described in their research that making a decision about academic program selection is crucial as it can significantly impact career opportunities. Recent research using psychometrics has highlighted a clear correlation between academic fields and students' personality traits. To assist students in making informed academic decisions, in their study they find out that the Hierarchical classification approach using Random Forest Classifier outperformed the One-level Random Forest classifier approach with an accuracy of 96.1% for the 1st-level and 92.86%, 89.29%, and 94.74% respectively for 2nd-level. This study's findings demonstrate that the proposed framework has enormous potential in assisting prospective students in making informed decisions about suitable higher study options, thereby unlocking human potential. In a study by Dalcı et al. (2013), the researchers investigated the factors influencing Iranian students' decision to choose accounting as a major. The results showed that financial and job-market factors were significantly more important to Iranian students who intended to major in accounting compared to those who chose a non-accounting major. The discriminant analysis indicated that these factors had the highest discriminatory power, which suggests that the more importance Iranian students placed on financial and job-market factors, the more likely they were to pursue an accounting major. The study speculated that this trend could be due to several factors such as war, high inflation rates, foreign embargoes, and high unemployment rates in Iran, which may influence students to prioritize financial stability and career opportunities when selecting their career paths. In addition, Yin et al. (2015) explored the factors that influence the decision-making process of Chinese students when considering Thai universities. The research identified several key factors that significantly influence the college choice decisions of Chinese students. These included agency, referral from friends, joint programs and academic cooperation, cooperating with high schools, online information, and the Chinese National Matriculation Examination (NME) score. The study suggests that these factors play a crucial role in the recruitment process and have a significant impact on Chinese students' college choice decisions. Moreover, Ghansah et al. (2016) examined the factors that influence students' college choice decisions in Ghana. The researchers surveyed 2,534 admitted students at a major university in Ghana, receiving a response rate of 39%. Through factor analysis, they ranked 26 university attributes that students consider when making their decisions. Based on the findings, the study identified various marketing strategies that higher education institutions can use to increase student enrolment, ranked in order of perceived effectiveness. The study revealed that factors such as word of mouth (from family and friends), proximity/ nearness, accreditation/ affiliation, affordability, and lecturing style were the most influential factors that affect students' college choice decisions.

Predictive modelling and machine learning are becoming increasingly popular in higher education. These technologies can be used to analyse large amounts of data and predict student outcomes, such as grades, retention rates, and graduation rates. This information can be used to develop personalized interventions for students to increase their chances of success. One area where predictive modelling and machine learning are being used is in enrolment management. (Al Ka'bi, 2023) By analyzing past enrolment data, institutions can develop models to predict the likelihood of a student accepting an offer of admission. This information can be used to target marketing efforts to students who are most likely to enroll. Predictive modelling and machine learning are also being used to identify students who are at risk of dropping out. By analyzing data such as grades, attendance, and engagement with course materials, institutions can develop models to predict which students are most likely to drop out. This information can be used to develop personalized interventions for these students, such as academic advising or tutoring. (Niyogisubizo, 2022; Lottering & Lall, 2020) In addition Itani & Garlatti (2018) developed a drop-out prediction system that utilizes supervised machine learning and employs both aggregated and explicative classifiers. Their study revealed that the aggregated classifiers can accurately detect students who are at risk of dropping out, thereby enabling automated motivational feedback to be sent to the learners. Conversely, the explicative classifiers enable personalized intervention by teachers. The findings of the study were grouped into three primary tested axes. Firstly, the explicative models' readability, such as Decision Trees and Logistic Regression, allow for a detailed inspection of the classification process and the effect of features on the classification. These models may be difficult for non-experts to interpret, but they can provide valuable information for teachers to analyze and make personalized interventions.

Methodology

The study received data from the Academic Office of Southern University in Thailand, which was stored in their educational administration system. The dataset consists of information on 12,126 undergraduate students who were enrolled between 2020 and 2022, including their profile and family background. To protect the students' privacy, the data was anonymized before being provided to researcher. The Cross-Industry Standard Process for Data Mining (CRISP-DM) (Huber et al., 2019) was adopted for this study as the standard process for analyzing data mining data, as illustrated in Figure 2.

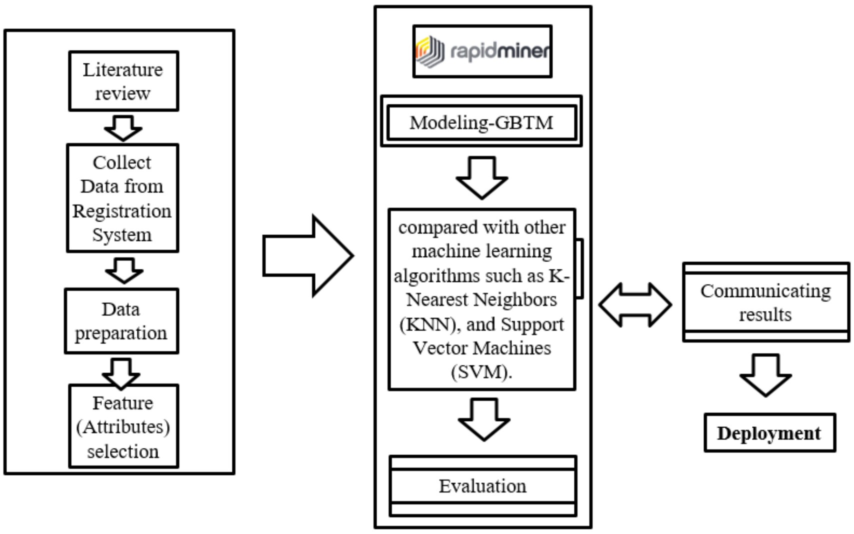

Literature review
Collect Data from Registration System
Data preparation
Feature (Attributes) selection
rapidminer
Modeling-GBTM
compared with other machine learning algorithms such as K-Nearest Neighbors (KNN), and Support Vector Machines (SVM).
Evaluation
Communicating results
Deployment

Figure 2 Research Framework

Rapid Miner Studio 10.1 software was employed to streamline the processes of data import, preparation, and analysis. After receiving the student data sheets, they were merged into a unified table and filtered to retain only those students who had either graduated or dropped out. The key steps for data preprocessing and cleaning included:

A. Handling missing data.

B. Transforming and creating attributes

C. Reducing dimensionality and eliminating redundancy

E. Attributes: The study utilized various attributes, categorized into ten type according to their reference: Admission Year, High school GPA, District, Province, Parent Status, Parent Relative, Parent Revenue, and Parent Occupation. Furthermore, there is a binary target variable that determines the faculty of the student.

The present study implemented the Gradient Boosted Trees Machines (GBTM) algorithm, trained on a preprocessed dataset. To optimize the performance of the model, grid search and k-fold cross-validation techniques were employed for hyper parameter tuning. These methods aimed to identify the most suitable hyper parameters for the GBTM algorithm to achieve optimal performance.

The GBTM algorithm was evaluated using various performance metrics, including accuracy, precision, recall, F1 score, and area under the curve (AUC). Additionally, the performance of the GBTM algorithm was compared with other machine learning algorithms such as K-Nearest Neighbors (KNN), and Support Vector Machines (SVM).

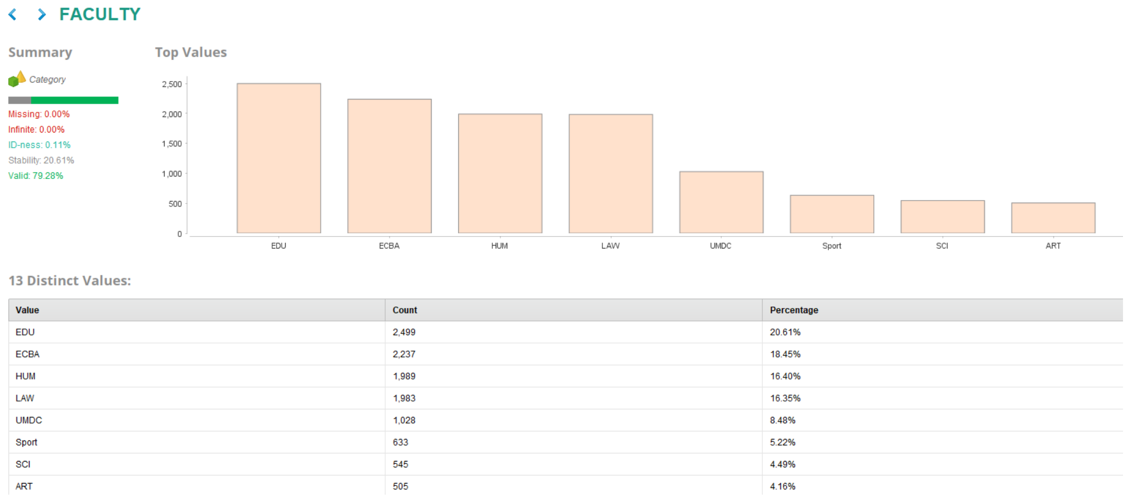
Results

Data Analysis and Interpretation

Interpretation of Table 2: the table above presents data on student enrollments at a university across different faculties and admission years (2020, 2021, and 2022). It includes the total number of students enrolled in each faculty for each of the three admission years, as well as the percentage of the total student population that each faculty represents. Among the faculties, the Education faculty had the highest total enrollment over the three-year period, with 2,499 students enrolled, while the Engineering faculty had the lowest total enrollment, with 101 students enrolled. The Economics and Business Administration faculty had the second-highest total enrollment, with 2,237 students, and the Management for Development College had the highest percentage of the total student population, at 8.48%. The table provides similar information for the other faculties, including Humanities and Social Sciences, Agro-and Bio-Industry, International College, LAW, NURSE, Science, Health and Sports Science, and Technology and Community Development. The data can be used to compare enrollment trends across faculties and admission years and to identify any areas of growth or decline in student populations. The statistics in the table summarize key features of the data, such as the number of students enrolled and various percentiles, which can be used to gain insight into the enrollment patterns at Southern University during the given time frame.

Table 2 Showing Descriptive Statistics Pertain to the Enrollment of Students at Southern University, Covering the Years 2020 through 2022

Faculty	Admission Year			Total	%
	2020	2021	2022		
Fine Arts	163	185	157	505	4.16
Economics and Business Administration	802	679	756	2,237	18.45
Education	866	739	894	2,499	20.61
Engineering	30	39	32	101	0.83
Humanities and Social Sciences	668	627	694	1,989	16.40
Agro-and Bio-Industry	29	65	37	131	1.08
International College	15	8	13	36	0.30
LAW	599	668	716	1,983	16.35
NURSE	50	57	45	152	1.25
Science	154	202	189	545	4.49
Health and Sports Science	185	224	224	633	5.22
Technology and Community Development	104	102	80	286	2.36
Management for Development College	227	330	471	1,028	8.48
Total	3,892	3,925	4,308	12,125	100


FACULTY

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.11%
Stability: 20.61%
Valid: 79.28%

Top Values

13 Distinct Values:

Value	Count	Percentage
EDU	2,499	20.61%
ECBA	2,237	18.45%
HUM	1,989	16.40%
LAW	1,983	16.35%
UMDC	1,028	8.48%
Sport	633	5.22%
SCI	545	4.49%
ART	505	4.16%

Figure 3 Displaying the Distribution of Students across Faculties.

Figure 3 depicts the enrollment distribution of students among various faculties. The first column shows the abbreviated names of each faculty, such as EDU for Education, ECBA for Economics, Business, and Accounting, etc. The second column represents the total number of students enrolled in each faculty. The third column displays the percentage of students enrolled in each faculty out of the total number of students in the institution. For example, out of the total number of students, 2,499 students, which account for 20.61%, are enrolled in the Education faculty (abbreviated as EDU). Similarly, 2,237 students (18.45%) are enrolled in the Economics, Business, and Accounting faculty (abbreviated as ECBA), 1,989 students (16.40%) are enrolled in the Humanities faculty (abbreviated as HUM), 1,983 students (16.35%) are enrolled in the Law faculty (abbreviated as Law), 1028 students (8.48%) are enrolled in the Management for Development College faculty (abbreviated as UMDC), 633 students (5.22%) are enrolled in the Health and Sports Science faculty, 545 students (4.49%) are enrolled in the Science faculty (abbreviated as SCI), and 505 students (4.16%) are enrolled in the Fine Arts faculty (abbreviated as Fine ART

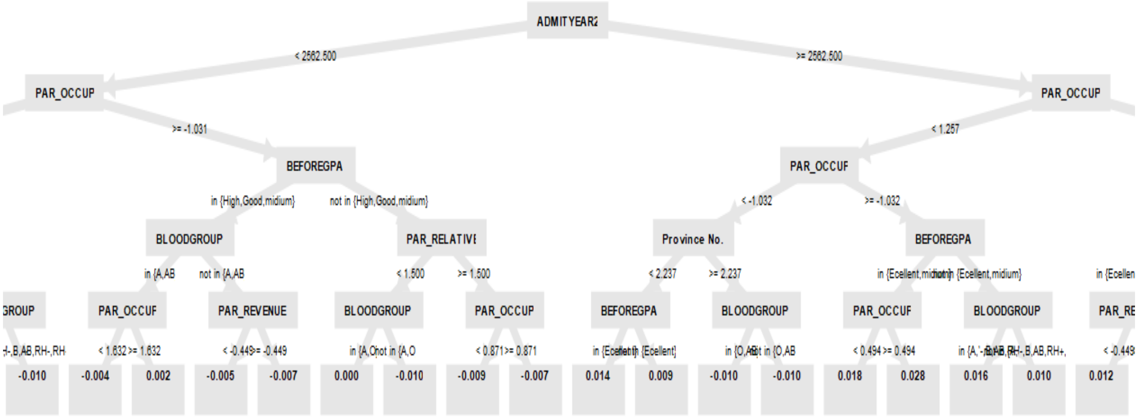


Figure 4 A Tree in the Gradient Boosted Trees Machines Algorithm

Firstly, figure 4 displays a decision tree that predicts a student's selection based on input factor, namely, "ADMITYEAR", "PaAR_OCCUP" and "BEFORGPA" respectively. The Gradient Boosted Trees (GBT) model achieved an impressive accuracy of 85% and a precision of 87% in predicting student faculty selections, indicating its strong performance in correctly identifying the chosen faculties based on various input variables. The evaluation of our multinomial classifier model reveals several key metrics that highlight its performance and effectiveness. First, the Mean Squared Error (MSE) stands at 0.754178. MSE measures the average squared difference between predicted and actual values, providing insight into the model's accuracy. Lower MSE values indicate a better fit of the model to the data, suggesting that the predictions are closer to the actual outcomes. This metric is crucial as it quantifies the error in the model, allowing us to assess how well it is performing overall. Next, the Root Mean Squared Error (RMSE), which is the square root of the MSE, is recorded at 0.86843425. The RMSE offers a direct interpretation of the error in the same units as the target variable, making it more intuitive to understand. A lower RMSE indicates better performance, signifying that the model's predictions are more accurate and have less deviation from the actual values. The R^2 (R-squared) value for our model is 0.91582996. This metric represents the proportion of variance in the target variable that is explained by the model. An R^2 value close to 1 indicates a high explanatory power, meaning the model accounts for a large portion of the variability in the data. This high R^2 value suggests that our model effectively captures the underlying patterns and relationships within the dataset. In terms of classification performance, the model achieves a Precision of 0.594. Precision measures the accuracy of the positive predictions, indicating the proportion of true positive results among all positive predictions made by the model. A higher precision value means that the model has a higher accuracy rate in predicting the relevant cases, reducing the number of false positives. The Recall metric for the model is 0.460. Recall, also known as sensitivity, measures the ability of the model to identify all relevant instances within the dataset. It indicates the proportion of true positive cases that were correctly identified by the model out of all actual positive cases. A higher recall value implies that the model is effective in detecting the true positive cases, reducing the number of false negatives. Finally, the F1 Score for the model is 0.518. The F1 score is the harmonic mean of precision and recall, providing a balanced measure that considers both metrics. It is particularly useful when dealing with imbalanced datasets, as it ensures that neither precision nor recall is favored disproportionately. The F1 score provides a comprehensive evaluation of the model's performance, highlighting its ability to make accurate and relevant predictions.

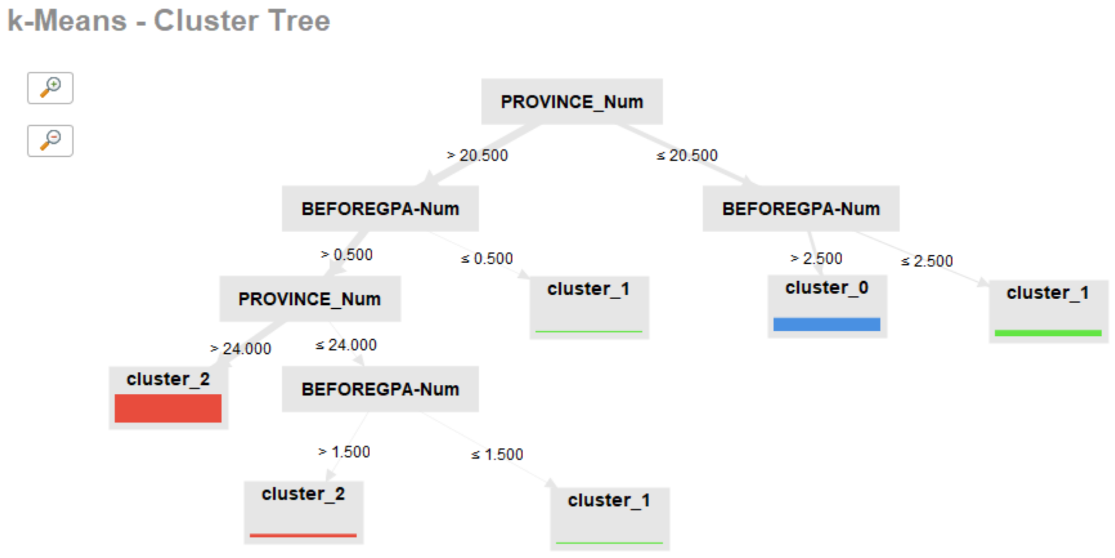


Figure 5 Modeling K-means Cluster Tree (Chertchom, 2023)

Figure 5 illustrates a K-means Cluster Tree that is utilized to predict a student's faculty selection based on specific input factors. The two primary factors taken into consideration in this analysis are "PROVINCE_Num" and "BEFORGPA". This figure provides a visual representation of how these variables influence student choices and allows for a deeper understanding of the clustering patterns within the dataset. To measure accuracy, we assess how similar the data within the same cluster is, using a metric called the Davies-Bouldin (DB) Index. The DB Index measures the clustering algorithm's error in grouping by calculating the relationship between the distance among the centroids of each cluster and the size of the clusters. It evaluates how distinct each cluster is. In the DB Index, a lower value indicates less difference between clusters, and a value of 0 indicates highly efficient clustering. In our model, the DB Index is -0.680. In this case, being close to 0 suggests that the clustering is highly efficient.

In our examination of the SVM model, we analyzed the coefficients assigned to various features to understand their impact on the model's predictions. Each coefficient represents the weight of a feature, indicating how strongly it influences the predicted outcome. Positive coefficients suggest that an increase in the feature value positively affects the prediction, while negative coefficients indicate the opposite. Key features such as "BEFOREGPA" (Grade Point Average before admission), "BLOODGROUP" (Blood Group), and socioeconomic factors like " PAR_STATUS" (Parental Status) and "PAR_REVENUE" (Parental Revenue) displayed significant coefficients. These coefficients reveal whether each feature contributes positively or negatively to the predicted outcome. For instance, a coefficient of 0.051 for 'BEFOREGPA = Bad' indicates a strong positive influence on the prediction, whereas a coefficient of –0.009 for 'BLOODGROUP = O' suggests a negative influence. Moreover, the SVM model achieved an F1 score of 0.33, an F1 score of 0.33 typically indicates moderate performance. This performance demonstrates that the model effectively predicts outcomes related to student choices based on the analyzed features.

From our analysis, it's evident that students' decisions regarding faculty selection are influenced by a complex interplay of various factors. Among these, province of origin and Grade Point Average (GPA) stand out as pivotal determinants shaping students' educational journeys. The influence of province and GPA on student selection is profound and multifaceted. Provinces impact educational choices by influencing accessibility, fostering familiarity, and reflecting regional context. Concurrently, GPA serves as a crucial indicator of academic capability and preparedness for rigorous academic pursuits. It acts as a benchmark against which students measure their readiness for challenging coursework and their potential for success in their chosen fields of study.

Discussion

Based on the outcomes of the Gradient Boosting model, it is evident that this algorithm is highly effective for predictive modelling tasks. Its iterative tree-building approach systematically corrects errors from previous iterations, leading to robust predictive accuracy. However, it's essential to monitor the model's performance closely, particularly when significant errors occur, as this may indicate overfitting or other underlying issues that require careful investigation. When compared with other machine learning algorithms like Support Vector Machines (SVM) and K-Nearest Neighbours (KNN), the Gradient Boosting model demonstrates competitive performance. Specifically, it achieves a precision of 0.594, indicating that it accurately predicts a student's faculty selection 59.4% of the time. The recall score of 0.460 suggests that the model captures 46.0% of all positive instances correctly. Moreover, with an F1 score of 0.518, the model strikes a balance between precision and recall, providing a comprehensive measure of overall performance.

To further optimize the Gradient Boosting model's effectiveness:

1. Experiment with different values for hyper parameters such as learning rate, number of trees, and tree depth. This iterative process aims to identify the optimal combination that maximizes model performance and minimizes overfitting.

2. Feature Importance: Utilize feature importance metrics to identify and prioritize the most influential variables. By focusing on these key features, the model can be refined to enhance accuracy and relevance in predicting student faculty selections.

When comparing the GBT model with other algorithms like KNN and SVM, the F1 score can provide a more balanced evaluation than accuracy alone, especially in cases of class imbalance. K-Nearest Neighbours (KNN): Typically, KNN can perform well with balanced datasets but might struggle with class imbalance and high-dimensional data. SVMs can handle high-dimensional data and class imbalance well if appropriate kernel functions and regularization parameters are chosen. F1 scores can be high if the model is well-tuned. The GBT model shows strong performance overall, but detailed comparisons with other algorithms using F1 scores would give clearer insights into which model best suits the faculty selection task.

Suggestion

Suggestions for Applying Research Findings

The study's findings offer several recommendations for future research and management practices, focusing on benchmarking and practical applications. Firstly, expanding the study's scope to include additional faculties or universities across Thailand would not only enhance the accuracy of predicting student selection in Southern Thailand but also facilitate benchmarking against diverse educational contexts. Secondly, improving the model by integrating additional input variables such as student demographics, extracurricular activities, and personal interests could further enhance its predictive capabilities and benchmarking metrics. Thirdly, to strengthen the model's reliability and applicability, rigorous benchmarking against other machine learning algorithms and educational datasets from different regions should be undertaken. Furthermore, educational institutions in Southern Thailand could utilize benchmarking results to identify prospective students likely to choose specific faculties, thereby optimizing recruitment and retention strategies. Lastly, leveraging benchmarking insights to provide personalized guidance and counselling services to students who are undecided about their faculty selection based on their individual characteristics and preferences would enhance the model's practical application and benchmarking utility.

Suggestions for Future Research or the Future Study

For future research endeavors, investigating the influence of external factors such as economic conditions, social and cultural norms, and institutional policies on students' faculty selection would be beneficial. Additionally, comparing the predictive performance of different machine learning algorithms such as random forests or neural networks could provide insights into selecting the most effective models for educational applications. Overall, the Gradient Boosted Trees Machines Algorithm shows promise as a valuable tool for enhancing student recruitment, retention, and guidance services in Southern Thailand.

References

Aiken, J. M., De Bin, R., Hjorth-Jensen, M., & Caballero, M. D. (2020). Predicting Time to Graduation at a Large Enrollment American University. *Plos One*, 15(11), e0242334.

Al Ka'bi, A. (2023). Proposed Artificial Intelligence Algorithm and Deep Learning Techniques for Development of Higher Education. *International Journal of Intelligent Networks*, *4*, 68-73.

Ben-Assuli, O., & Vest, J. R. (2022). Return Visits to the Emergency Department: An Analysis using Group based Curve Models. *Health Informatics Journal, 28*(2), 14604582221105444.

Bilquise, G., Abdallah, S., & Kobbaey, T. (2019). *Predicting Student Retention among a Homogeneous Population using Data Mining*. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019. Cham: Springer International Publishing.

Çakıt, E., & Dağdeviren, M. (2022). Predicting the Percentage of Student Placement: A Comparative Study of Machine Learning Algorithms. *Education and Information Technologies*, *27*(1), 997-1022.

Chertchom, P. (2023). Application of Data Mining in Studying Factors in the Selection of University Admission Through Clustering Modeling. *Journal of Administration and Social Science Review, 6*(4), 157-168.

Dalcı, I., Araslı, H., Tümer, M., & Baradarani, S. (2013). Factors that Influence Iranian Students' Decision to Choose Accounting Major. *Journal of Accounting in Emerging Economies*, *3*(2), 145-163.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *The Journal of Machine Learning Research, 15*(1), 3133-3181.

Ghansah, B., Benuwa, B. B., Ansah, E. K., Ghansah, N. E., Magama, C., & Ocquaye, E. N. N. (2016). Factors that Influence Students' Decision to Choose a Particular University: A Conjoint Analysis. *International Journal of Engineering Research in Africa, 27*, 147-157.

Itani, A., Brisson, L., & Garlatti, S. (2018). Understanding Learner's Drop-out in MOOCs. In *Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, November 21–23, 2018, Proceedings, Part I 19* (pp. 233-244). Springer International Publishing.

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What Predicts Student Satisfaction with MOOCs: A Gradient Boosting Trees Supervised Machine Learning and Sentiment Analysis Approach. *Computers & Education, 145*, 103724.

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.

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018). *Prospectively Predicting 4-year College Graduation from Student Applications*. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 280-289).

Kabra, R. R., & Bichkar, R. S. (2011). Performance Prediction of Engineering Students using Decision Trees. *International Journal of Computer Applications, 36*(11), 8-12.

Kamal, N., Sarker, F., & Mamun, K. A. (2020). *A Comparative Study of Machine Learning Approaches for Recommending University Faculty*. In 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) (pp. 1-6). IEEE.

Kamal, M., & Talbert, D. (2024). *Beyond Size and Accuracy: The Impact of Model Compression on Fairness*. In The International FLAIRS Conference Proceedings (Vol. 37).

Ketui, N., Wisomka, W., & Homjun, K. (2019). *Using Classification Data Mining Techniques for Students Performance Prediction*. In 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON) (pp. 359-363). IEEE.

Lee, S. B., Kim, Y. J., Hwang, S., Son, H., Lee, S. K., Park, K. I., & Kim, Y. G. (2022). Predicting Parkinson's Disease using Gradient Boosting Decision Tree Models with Electroencephalography Signals. *Parkinsonism & Related Disorders, 95*, 77-85.

Lottering, R., Hans, R., & Lall, M. (2020). A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study. *International Journal of Advanced Computer Science and Applications, 11*(10), 417-422.

Lv, S., Zhu, Y., Cheng, L., Zhang, J., Shen, W., & Li, X. (2024). Evaluation of the Prediction Effectiveness for Geochemical Mapping using Machine Learning Methods: A Case Study from Northern Guangdong Province in China. *Science of The Total Environment, 927*, 172223.

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A Systematic Review of the Literature on Machine Learning Application of Determining the Attributes Influencing Academic Performance. *Decision Analytics Journal, 7*, 100204.

Nagy, M., & Molontay, R. (2018). Predicting Dropout in Higher Education based on Secondary School Performance. In 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES) (pp. 000389-000394). IEEE.

Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, a Tutorial. *Frontiers in Neurorobotics*, *7*, 21.

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting Student's Dropout in University Classes using Two-Layer Ensemble Machine Learning Approach: A Novel Stacked Generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066.

Oztekin, A. (2016). *A Hybrid Data Analytic Approach to Predict College Graduation Status and its Determinative Factors*. Industrial Management & Data Systems.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). *Classification and Prediction of Student Performance Data using Various Machine Learning Algorithms*. Materials Today: Proceedings, 80, 3782-3785.

Patcharacharoenwong, C., Hernmek, K., & Kimpan, W. (2020). Arrival Time Prediction Model to a Pier for Public Transportation Boats. *Journal of Science Ladkrabang*, *29*(2), 31-44.

Singh, W., & Kaur, P. (2016). Comparative Analysis of Classification Techniques for Predicting Computer Engineering Students' Academic Performance. *International Journal of Advanced Research in Computer Science*, *7*(6).

Zhou, L., Fujita, H., Ding, H., & Ma, R. (2021). Credit Risk Modelling on Data with Two Timestamps in Peer-to-Peer Lending by Gradient Boosting. *Applied Soft Computing, 110*, 107672.