Improving University Programme Recommender System Based on MBTI by Using Gradient Boosted Trees and Firefly Algorithm

Phuwadol Viroonluecha[1]
and Thongchai Kaewkiriya[2]
[1]Thai Nichi Institute of Technology, Thailand
[2]Panyapiwat Institute of Management, Thailand
[1]vi.phuwadol_st@tni.ac.th
[2]tkaewkiriya@gmail.com

Abstract - **Choosing a university major after school graduation is a tough question for an undecided student. These students could be qualified students or gifted students. Furthermore, the lack of guidance specialists is one of the various reasons. The framework in this paper aims to encourage students and guidance system with a hybrid machine learning approach such as gradient boosted trees and metaheuristics feature selection with the Myers-Briggs type indicator (MBTI) personality assessment. The main objective of the hybrid recommender system is to identify the pattern of student background, education capability, diverse influences to student, student personality and preferences. Moreover, this paper represents the implementation of a classifier such as gradient boosted trees which are performed well in particular dataset and feature engineering – firefly algorithm was used to improve accuracy and runtime. As a result, the system recommends the appropriate major for each individual need and can use for online guidance in a remote area where at least has a computer with internet access.**

Keywords - **University Programme Recommendation, MBTI, Gradient Boosted Trees, Firefly Algorithm, Machine Learning**

I. INTRODUCTION

In order to enter the university, school graduates were constrained to choose their major once they fill up an application. This is an ordinary process for university admission system in Thailand. But there are some problems with the system which is some students uncertain their preference and unaware of available programmes in universities. Furthermore, inadequacy of qualified guidance teachers in Thailand [1] is part of the problem. In this paper, the researcher demonstrates the concept of the university programme recommendation based on the student personality assessment, Myers-Briggs type indicator (MBTI) which developed from the personality theory of Carl Jung and the ensemble data mining technique called the Gradient boosted trees with metaheuristics feature selection - firefly algorithm.

Undecided student is a student who unwilling, unable or unready to make educational choices or who enter college with a tentative decision that changes. This kind of student could be a high ability student, student athlete, adult student or any student from multiple background [2]. There are several reasons for example, lack of self-information, lack information about majors or programs, lack information about careers. The assessment, student profile or background and environmental influence are encouraging these students for

decision making.

The Myers-Briggs type indicator is a popular behavioral assessment [3] for understanding a learning style which might include elements of extroversion, sensation, feeling, and perception as personality dimensions [4]. In Thailand, the MBTI assessment widely uses in Educational section and research, such as the relationships between the MBTI and academic achievement [5-6], the personalised course by MBTI [7].

In recent years, we clearly find the patterns of the dataset which is individual profile in research by data science. The previous researches about personalised learning is almost on course selection, such as Lalita Na Nongkhai and Thongchai Kaewkiriya who conducted in e-learning recommendation based on the index of the learning styles model (ILS) [8]. Researchers used ILS and decision tree technique to find the rule bases with 76.92% of accuracy as a measurement of the result and high appropriate evaluation from experts.

Furthermore, researchers found metaheuristics combines with data mining algorithm can create the hybrid method for enhancing the performance of model. In optimising SMOTE by metaheuristics with neural network and decision tree, Jinyan Li et al. [9] can improve their classifiers with a Bat-inspired algorithm (BAT) and particle swarm optimization algorithm (PSO). They conducted experiments with 30 datasets and the result of Kappa values was significant improvement.

In this paper, we would like to demonstrate improving university programme recommendation based on student personality assessment and gradient boosted trees with the firefly algorithm as a hybrid technique. The experiment was implemented with nature inspired metaheuristic approaches to optimise either accuracy or processing time. The paper was divided into six sections. The first section is an introduction. The background and related works located in section two and three. The fourth part is the process and algorithm. The noticeably different results were compared in the fifth section in this paper. The last section researchers have discussed the results and conclusion.

II. BENEFIT OF GREEN UNIVERSITY

A. Personality Theory of Carl Jung and MBTI

Carl Jung's theory of psychological types was based on clinical observation. Jung postulated a series of four cognitive functions which are thinking, feeling, sensation, and intuition, each having one or two poles, the total is eight dominant functions.

The Myers–Briggs Type Indicator (MBTI) is a popular assessment in education which was proposed for the purposed of indicating different human psychological preferences underlie interests, needs, values, and motivation. There are four opposite pairs of preferences: introversion (I) and extraversion (E), intuition (N) and sensing (S), feeling (F) and thinking (T), and perception (P) and judging (J). These abbreviations are applied to all sixteen types including their proportion from collecting dataset as shown in fig. 1.

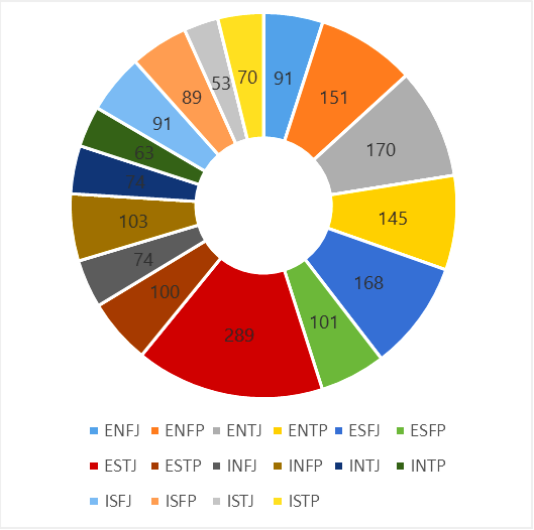

Fig. 1 The Proportion of MBTI Personality Type in Dataset.

B. Firefly Algorithm as Feature Engineering

Nature inspired metaheuristic approach is a method for finding the optimal answer or output with natural mimicry. One of them is the firefly algorithm proposed by Xin-She Yang [10] by assuming three firefly's behaviours:

first, all of them are unisexual, so that any single firefly will be attracted to all others; second, attractiveness is corresponding to their brightness for any two fireflies, the brighter one will move towards to the less bright one; in contrast, the apparent brightness decrease as their mutual distance increases; and the last, if there are no fireflies brighter than a given firefly, it will move randomly.

The attractiveness of a firefly, β, is a monotonically diminishing function regarding distance, r:

$$\beta(r) = \beta_0 e^{-\gamma r^2}$$

The range between two fireflies r corresponds to their Euclidian distance. The terms $\beta 0$ and γ signify the attractiveness at r = 0 and light absorption coefficient respectively. The movement of a firefly $\mathbf{x}i$ because of an attraction to firefly $\mathbf{x}j$ is presented as follows:

$$x_i = x_i + \beta_0 e^{-\gamma r^2}(x_j - x_i) + \alpha\epsilon_i$$

Correspondingly, α denotes the randomization parameter, while the term ϵi is a vector of random numbers drawn from the uniform distribution U (-0.5, 0.5).

C. Gradient Boosted Trees

Gradient Boosted Trees (GBT) is a supervised machine learning method for classification problem, which builds an ensemble model from small decision trees. Each tree attempts to correct errors from the previous stage. GBT works, including the loss function, weak learners and the additive model. The algorithm designed for distributed computing, so its execution speed may slow down on single processing.

To increase the model performance and speed time, in this paper, the researcher uses the Extreme Gradient Boosting (XGBoost) which is a novel classifier based on an ensemble of classification and regression trees (CART). XGBoost proposed by Tianqi Chen and Carlos Guestrin [11], aims to provide a scalable, portable and distributed gradient boosted library.

Let the output define as:

$$f(x) = w_q(x_i)$$

where

x is the input vector.

w_q is the score of the corresponding leaf q.

The output of an ensemble of N trees will be:

$$y_i = \sum_{n=1}^{N} f_n(x_i)$$

The XGBoost algorithm attempts to minimize the following objective function J at step t:

$$J(t) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(f_i)$$

where

The first part contains the train loss function L (e.g. mean squared error) between real class y and output $\hat{y}$ for the n samples.

The second part is the regularization step, which controls the complexity of the model and avoid overfitting.

In XGBoost, the complexity is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where

T is the number of leaves.

γ is the pseudo-regularization hyperparameter, depending on each dataset.

λ is the L2 norm for leaf weights.

Using gradients for second order approximation of the loss function and finding the optimal weights w, the optimal value of objective function is:

$$J(t) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\epsilon 1} g_i)^2}{\sum_{i\epsilon 1} h_i + \lambda} + \gamma T$$

where

$gi = \partial \hat{y}\, t{-}1L(y, \hat{y}\, t{-}1)$

$hi = \partial\ 2\ \hat{y}\, t{-}1L(y, \hat{y}\, t{-}1)$ are the gradient statistics on the loss function, and I is the set of leaves.

In fig. 2, illustrated some branches and leaves in the business and law tree 1 from the GBT model with start from student's GPAX then if student meet criteria C+ or B+ the model determines the influence from career market if not in medium influence then we can see student's age if under 23 years old then the model will return the 0.7 to compare with another trees.

Fig. 2 Some Branches and Leaves from the GBT Model.

III. RELATED WORK

Advanced forecasting of career choices for college students based on campus big data was proposed by Min Nie et al [12]. Their framework assists college graduates with psychological questionnaires, student behaviour in ten million data over four thousand students. Random forest was chosen due to its high performance above other comparing algorithms then they applied feature analysis and reading interest analysis to improve the career prediction model.

Course Navigator is a personalized recommendation system proposed by Prateek Basavaraj and Ivan Garibay [13]. The guidance for IT student based on their goal orientation aims to provide more flexibility and improve student advising. From 510 survey data, the result of data analysis was divided into two course recommendations: Mastery-Goal-Oriented and Performance-Goal-Oriented.

Advising academic major and university selection with machine learning based on ontology was proposed by Charbel Obeid et al [14]. Their approach uses a user profile to recommend a university programme from the data mining model with alumni students' survey results and profile with ontology.

On the other hand, [11] is focused on downstream process. From previous research, 50% - 70% of college students will change their major at least once during their college study [12]. The course navigator [13] intends to recommend only course structure for IT students and the framework from [14] is not based on students' grades, interests, behaviour and preferences and it is only framework not implement yet. In this paper, researchers try to apply the hybrid framework which is psychological assessment, machine learning algorithm and feature engineering to predict appropriate university major for school graduates. The main purposes are encouraging students and guidance system in school and avoiding waste of time in the wrong specialisation of freshman students.

IV. PROCESS AND ALGORITHM

A. Data Exploration and Data Preprocessing

A plenty of factors which influence on choosing the major decision but from the literature review, we found that three significant groups of factors are common in several pieces of research that are student personal profile, influences for surrounding environment and personality from the assessment. Therefore, we created an online questionnaire to collect the data from the distance method so the data came from every province in Thailand.

TABLE I
FEATURES FROM QUESTIONNAIRE

No.	Question	Attribute
1	**Gender**	Male, female, LGBT
2	**Age**	19 – 25 and above
3	**Year of study**	1, 2, 3, 4 and completed
4	**GPAX**	Four-scale grading system
5	**Birthplace**	76 provinces with Bangkok
6	**Previous Education**	High school (art, maths), non-formal, vocational, high vocational and bachelor
7	**Participation in prospective university activity**	Yes or no
8	**Influence form the surrounding people of student**	High, medium, low impact
9	**Student interests with choosing programme**	High, medium, low impact
10	**Reputation of university**	High, medium, low impact
11	**Instructor and programme, quality of teaching**	High, medium, low impact
12	**Future career market**	High, medium, low impact
13	**MBTI type**	16 types: INTP, INFP, INTJ, INFJ, ENFP, ENTJ, ENTP, ENFJ, ISFJ, ISFP, ISTJ, ISTP, ESFJ, ESFP, ESTJ, ESTP
14	**University major**	social sciences and journalism, sciences and mathematics, information and communication technology, medical and health sciences, engineering and industrial, education, business and law, arts and humanities.

From table I, data gathering from an online questionnaire includes personal information such as age, gender, education year, GPAX, birthplace, previous education, activity with prospective university; influences from various factors such as influence from the surrounding people of student, student interests with choosing programme, reputation of the university, instructor and programme, the quality of teaching, equipment and classroom's atmosphere, and future career market; the last section is the MBTI assessment to find out one of sixteen types of the interviewee.

In the cleansing data process, we worked with some outlier data since respondents free to type some answers; removed missing and uncommon data rows; duplicate cases and Cases in the data who met exclusion criteria and shouldn't be in the study. Finally, there are 1832 questionnaires with 14 attributes, then labelled them by classifying programmes of education into eight groups referred to the international standard classification of education by UNESCO Institute for Statistics [15]: social sciences and journalism, sciences and mathematics, information and communication technology, medical and health sciences, engineering, education, business and law, arts and humanities.

B. Feature Engineering

A researcher aims to avoid garbage in, garbage out situation and reduces all noisy and irrelevant features of the model. In this stage, the target feature is university's programme then researcher chose four metaheuristics which is ant algorithm, bat algorithm, cuckoo algorithm, and the firefly algorithm for the purpose of selecting relevant attributes which exact impact on the prediction.

On one hand, ant algorithm chose only four features from 14 features which are gender, birthplace, previous education and MBTI while bat algorithm added one more feature from ant algorithm – age. On the other hand, cuckoo algorithm and firefly algorithm chose seven cognate features which is age, birthplace, previous education, influence from the surrounding people of student, future career market, MBTI but the last feature cuckoo selected GPAX while firefly picked gender.

TABLE II
ATTRIBUTES AFTER FEATURE ENGINEERING

No.	Algorithm	Selected Features
1	**Ant**	Gender, birthplace, previous education and MBTI
2	**Bee**	Age, gender, birthplace, previous education and MBTI
3	**Cuckoo**	Age, birthplace, GPAX, previous education, influence from the surrounding people of student, future career market and MBTI
4	**Firefly**	Age, birthplace, gender, previous education, influence from the surrounding people of student, future career market and MBTI

C. Gradient Boosted Trees Modelling and Performance

The proposed model in this paper is a gradient boosted trees (GBT) performing classification which is a supervised forward-learning ensemble method that obtains predictive results through gradually improved estimations. Accuracy was used to evaluate the performance of this model.

$$\%\text{Accuracy} = 100 - \%\text{Error}$$

The percentage of error can find by finding a relative error.

$$\text{Relative error} = \left|\frac{X_{mea} - X_t}{X_t}\right|$$

then

$$\%\text{Error} = \text{Relative error} \times 100$$

where

Xmea is measure value.

Xt is true value.

Researcher separated data into two parts, one for training the model and other for evaluating the model by the ratio 8:2.

To compare the results, researchers use a decision tree algorithm (DT) which is only one tree with leaves and a random forest algorithm (RF) which is bootstrapped or bagging of decision trees. Both are in the same group for classification problem solving. Subsequently, one of them will be chosen to compare with various feature selection techniques.

V. EVALUATION AND RESULTS

The result of a gradient boosted trees was significant from another two algorithms – decision tree and random forest. Accuracy of GBT was notable from the two remaining by 82.88 % of accuracy that significant from 78.53% and 59.24% of the RF and DT respectively. On the contrary, GBT used to spend most processing 1 minute and 41 seconds, which we can improve it by hybridising with metaheuristics.

TABLE III
PERFORMANCES OF DT, RF, AND GBT APPROACHES

Algorithm	Accuracy %	Precision %	Runtime (sec)
Decision Tree	59.24	58.59	**0.46**
Random Forest	78.53	77.99	14.33
Gradient Boosted Trees	**82.88**	**82.43**	84.89

Afterwards, researchers used GBT by applying four feature selection techniques. The result shows in table IV. Ant and bat algorithms contain a minimal number of features so their accuracy could not improve while cuckoo well performed in terms of reducing the process time for 11.24 seconds. At any rate, cuckoo could not gain its accuracy whilst firefly demonstrated its performance by increasing an accuracy 83.97% different from the original model 1.09 % and the hybrid method GBT with firefly algorithm decreased runtime as well.

The cross-validation was used for finding the optimal parameters of the GBT model which are the number of trees and maximal depth. The particular step delivers the average of the performances over the number of folds iterations. We set 3 folds iteration and the result as shown in fig. 3. The optimal parameters of the GBT model were 73% accuracy from the training data when the maximal depth is 7 and 140 trees.

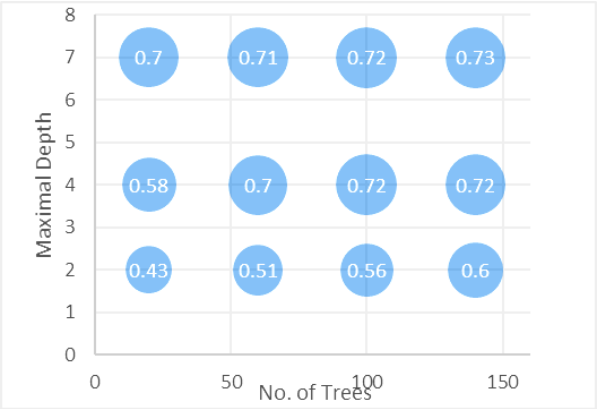

Fig. 3 The Optimal Parameters of GBT Model.

TABLE IV
PERFORMANCE OF FEATURE SELECTION METHODS WITH GBT APPROACH

Group	Accuracy %	Precision %	Runtime (sec)
All	82.88	82.43	74.46
Ant	71.74	70.73	65.43
Bat	78.80	78.24	63.65
Cuckoo	82.61	8196	**63.22**
Firefly	**83.97**	**83.38**	65.99

VI. CONCLUSION

In the present research, we have been used the gradient boosted trees technique with MBTI to construct a model which predicts the university programme solving a classification problem which performed excellent outcome. We used data from online questionnaire, treating 80% of data as learning, while the 20% rest was used for validation. As a result, Gradient boosted trees model has significant performance, either accuracy or process time by accuracy 82.88% and 74.46 seconds of runtime. Furthermore, comparing with cross-validation from the previous step we have got the accuracy 73% of training data, this guarantee the model does not contain the overfitting.

Furthermore, to avoid irrelevant parameters in some subsets, we applied feature selection to improve the accuracy and the runtime by deducting the noisy feature.

Thus, gradient boosted trees with the best feature selection for this experiment – firefly algorithm improved the accuracy to 84.78% and reduced processing time to 65.99 seconds. From the prototype, the result demonstrates good performance with acceptable accuracy. In the future work, we can use the framework to assist undecided students make an academic decision and to support guidance system in distance area by running on server which can access anywhere and anytime once user only has a computer with internet. To improve choices and gain more major, the model needs big data with a plenty of data to find a general pattern for single personality of the student.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Visessuvanapoom, P. and Wongwanit, S. (2013). "The Role of Teachers in Guidance Services for Education Reform in Thailand". HRD JOURNAL.

[2] Gordon, V. and Steele, G. (2015). "The Undecided College Student:An academic and career advising challenge". Springfield, IL: Thomas.

[3] Zamir, S. and et al. (2014). "Relationship between Personality and Occupational Stress among Academic Managers at Higher Education Level". Research on Humanities and Social Sciences.

[4] Picciano, G. (2017). "Theories and Frameworks for Online Education: Seeking an Integrated Model". Online Learning.

[5] Iaosanurak, C. and et al. (2017). "The Relationships Between Gender Majors' Subject Academic Achievement and Personality with Learning Style of Student Teachers". Journal of graduate studies Valaya Alongkorn Rajabhat University.

[6] Anutharun, N. and Romphoree, N. (2014). "Correlation Between MBTI Personalities and Academic Achievement of Nursing Students of Siam University". Siam University, Bangkok.

[7] Tananchai, A. (2017). "The Personality of Students Studying the Social Etiquette and Personality Development Course by Myers Briggs Type Indicators (MBTI) Theory". Swarm and Evolutionary computation.

[8] Nongkhai, L. and Kaewkiriya, T. (2015). "Framework for e-Leaming recommendation based on index of learning styles model". in Proceedings of the 7th Information Technology and Electrical Engineering (ICITEE).

[9] Li, J. and et al. (2015). "Optimizing SMOTE by metaheuristics with neural network and decision tree". in Proceedings of

the 3rd International Symposium on Computational and Business Intelligence (ISCBI), IEEE.

[10] Marichelvam, M. and et al. (2014). "A discrete firefly algorithm for the multi-objective hybrid flowshop scheduling problems". IEEE transactions on evolutionary computation.

[11] Chen, T. and Carlos, G. (2016). "Xgboost: A scalable tree boosting system". In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

[12] Nie, M. and et al. (2018). "Advanced forecasting of career choices for college students based on campus big data". Frontiers of Computer Science.

[13] Basavaraj, P. and Garibay, I. (2018). "A Personalized Course Navigator Based on Students' Goal Orientation". in Proceedings of the 2018 ACM Conference on Supporting Groupwork.

[14] Obeid, C. and et al. (2018). "Ontology-based Recommender System in Higher Education". in Companion of the Web Conference 2018 on the Web Conference.

[15] UNESCO and UNESCO Institute for Statistics. <http://uis.unesco.org/sites/default/files/documents/ international-standard-classification-of-education-fields-of-education-and-training-2013-detailed-field-descriptions-2015-en.pdf>.