AIP Conference Proceedings

A study of graduate on time (GOT) for Ph.D students using decision tree model

Cite as: AIP Conference Proceedings **2138**, 040006 (2019); https://doi.org/10.1063/1.5121085
Published Online: 21 August 2019

Wan Yung Chin, Chee Keong Ch'ng, and Jastini Mohd Jamil

A Study of Graduate on Time (GOT) for Ph.D Students using Decision Tree Model

Wan Yung Chin[1,a)], Chee Keong Ch'ng[1,b)] and Jastini Mohd Jamil[1,c)]

[1]*School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia*

[a]helen_cwy91@hotmail.com
[b]Corresponding author: chee@uum.edu.my
[c]jastini@uum.edu.my

Abstract. Over the years, there has been exponential growth in the number of Doctor of Philosophy (Ph.D) graduates in most of the universities all around the world. The increment of Ph.D students causes both university and government bodies concern about the capability of the Ph.D students to accomplish the mission of Graduate on Time (GOT) that is stipulated by the university. Therefore, this study aims to classify the Ph.D students into the group of "GOT achiever" and "non-GOT achiever" by using decision tree models. Historical data that related to all Ph.D students in a public university in Malaysia has been obtained directly from the database of Graduate Academic Information System (GAIS) in order to develop and compare the performance of decision tree models (Chi-square algorithm, Gini index algorithm, Entropy algorithm and an interactive decision tree). The result gained in four decision tree models illustrated that the attributes of English background, gender and the Ph.D students' entry Cumulative Grade Point Average (CGPA) result are the core in impacting the students' success. Among all models, decision tree model with Entropy algorithm perform the best by scoring the highest accuracy rate (72%) and sensitivity rate (95%). Therefore, it has been selected as the best model for predicting the ability of the Ph.D students in achieving GOT. The outcome can certainly ease the burden of universities in handling and controlling the GOT issue. Also, the model can be used by the university to uncover the restriction in this issue so that better plans can be carried out to boost the number of GOT achiever in future.

INTRODUCTION

Education is vital to prepare individuals with various learning skills and knowledge in a particular profession to face the demands and challenges of the era of globalization. Therefore, it successfully attracted individuals' interest and attention to further study in higher education, especially in doctorate degree so that they could obtain a better career opportunity. As a result, the number of Doctor of Philosophy (Ph.D) students increased dramatically from nearly 4,000 in year 2002 to approximately 40,000 in year 2012 [1]. The ability of Ph.D students to complete their studies within four years from registration date has been a concern as the rapid growth of Ph.D students in the past decade. They also found out that the doctoral students with thesis tend to complete their studies within 4.84 years averagely which mean that they are unable to achieve the mission of Graduate on Time (GOT) [2].

One of the northern public university in Malaysia is heading toward GOT mission. The GOT achievement is measured by accessing the students' completion time within 48 months from the registration date. UUM begins Ph.D program since year 1992 and there were 6 candidates in the first enrolment. In year 2014, doctoral students has increased to 506 candidates and the ability of the Ph.D students in attaining GOT has been a concern of students, lecturers, supervisors, faculty, school and university [3]. Although Ph.D students who are "GOT achiever" had increased to 449 in year 2017, there are still 130 postgraduates are "non-GOT achiever". This scenario is indeed worrying as the Ph.D students had lengthened the study period. Therefore, decision tree models have been developed in this study to classify the students into "GOT achiever" and "non-GOT achiever" based on their attributes so that effective and drastic solutions can be implemented by the three graduate schools of the university, namely Othman

The 4th Innovation and Analytics Conference & Exhibition (IACE 2019)
AIP Conf. Proc. 2138, 040006-1–040006-6; https://doi.org/10.1063/1.5121085
Published by AIP Publishing. 978-0-7354-1881-3/$30.00

Yeop Abdullah Graduate School of Business (OYAGSB), Awang Had Salleh Graduate School of Arts and Sciences (AHSGS) and Ghazali Shafie Graduate School of Government (GSGSG) to boost their number of GOT achiever.

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problem through data analysis. Data mining tools allow enterprises to predict future trends. Recently, the applications of data mining have been widely found in various fields such as sales forecasting and analysis, relationship marketing, customer profiling, outliers identification and fraud detection [4]. There are two types of data mining techniques which are supervised learning and unsupervised learning. Supervised learning is used to make a prediction of specific target or future values using known results obtained from historical dataset [5]. Unsupervised learning is used to identify hidden patterns or relationships that describe the data and draw the inferences consisting of input data without labelled responses in the data set [6]. This technique describes unknown value that has happened in the past and the new properties are not predicted. Basically, unsupervised learning applies in profiling or segmenting, which require assisting from a domain expert when exploring the properties of the examined data [7]. In this study, decision tree approach is carried out to classify the Ph.D students into "GOT" and "non-GOT" based on the data driven from database of Graduate Academic Information System (GAIS).

The decision tree is a flow chart like tree structure that uses a set of simple decision rules to divide up a large heterogeneous population into smaller and more homogeneous groups with respect to the target. Usually, the target variable is categorical and decision tree is used to classify and justify the probability of a record belongs to the most likely category. Decision tree consists of the root node, child node and leaf node. Root node is the top most node. It represents the entire population or sample where it will be further divided into two or more homogeneous groups. A child node is a descendent node with exactly one incoming node and two or more outgoing nodes. While a leaf node act as terminal node, it has exactly one incoming node and no outgoing node [8]. The data is split randomly into training set and test set where the training set is used to construct a tree and the latter is used to evaluate the constructed tree [9]. The use of training set and test set would avoid the construction of over-performed tree, hence provide a reliable tree for future classification. The tree is built in accordance with a splitting rule which divide the data into smaller part where the objects of the same class are assigned into the same nodes. This process is repeated on each derived subset by top-down induction of decision tree until each leaf consists of a single observation [10], and this scenario is referred as maximum homogeneity [11]. However, if a tree is deep or bushy, some branches of the tree may reflect inconsistencies due to noisy data or outliers. Therefore, tree pruning technique is needed to identify and eliminate the irrelevance branches in order to produce a simpler and informative tree [4].

METHODOLOGY

Secondary data that related to Ph.D students who are registered from year 1992 to year 2016 are obtained from the GAIS. The variables involved are Ph.D program, gender, date of birth, nationality, financial support status, registration date for Ph.D programs, date of proposal defense, date of viva , date of senate, previous academic background, entry cumulative grade point average (CGPA), name of supervisor, English language background. However, there are missing values existed which need to be handled before constructing the decision tree models. There are four stages along the process.

Stage 1: Data Selection Process

There are 544 data of Ph.D students selected from year 2011 to 2016 . Eight potential variables involved are Ph.D program, Gender, Nationality, Present of financial support, Age when register for Ph.D Program, Ph.D students' entry CGPA result, Number of supervisors and English background.

Stage 2: Pre-processing Process

Data cleaning are carried out to handle the problem in the data before constructing the tree models. For instance, for *Entry CGPA Result* variable, mean value is used to deal with missing values issue. Also, data reduction is applied to merge "Supervisor Name 1" and "Supervisor Name 2" variables to *Number of Supervisors*. Moreover, data discretization is used for *Age* variable where the values are discretized. Then, preamble analysis will be carried out to screen on the results.

Stage 3: Transformation Process

The variables that has been gone through transformation process by categorized the data into the forms that suitable for data mining process (see TABLE 1).

TABLE 1. Variables that undergo transformation process

Variable	Description	Category
Nationality	The variable is grouped as local and international	− Local − International
Present of Financial Support	This variable indicates whether a Ph.D student has any financial support during their study	− Yes − No
English Background	This variable indicates whether a Ph.D student has taken any English test before pursuing their Ph.D program	− Yes − No

Stage 4: Data Mining Process

Decision tree technique is chosen to classify the instances into “GOT achiever” or “non-GOT achiever” based on the selected variables in the historical data. At first, the sample is divided into 80% of training set and 20% of validation set. The classifier is developed by the training set and test set is used to provide the estimation of its performance. The assessment of the model will be done through the percentage of the error rate estimation.

In this study, the decision tree models are using Chi-square, Gini index, Entropy and interactive model. An interactive decision tree is conducted to determine whether a bushy tree brings more information. The process of model comparison has been done based five aspects, which are accuracy, sensitivity, specificity, Receiver Operating Characteristic (ROC) and average squared error (ASE) to select the best decision tree model where it will be used for predicting the ability of Ph.D students in achieving GOT. The formula used to calculate the accuracy, sensitivity and specificity are as below:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$Sensitivity = \ TP/((TP + FN) \quad (2)$$

$$Specificity = \ TN/(TN + FP) \quad (3)$$

RESULT

Preamble analysis has been carried out as in FIGURE 1 to screen on the demographics of students.

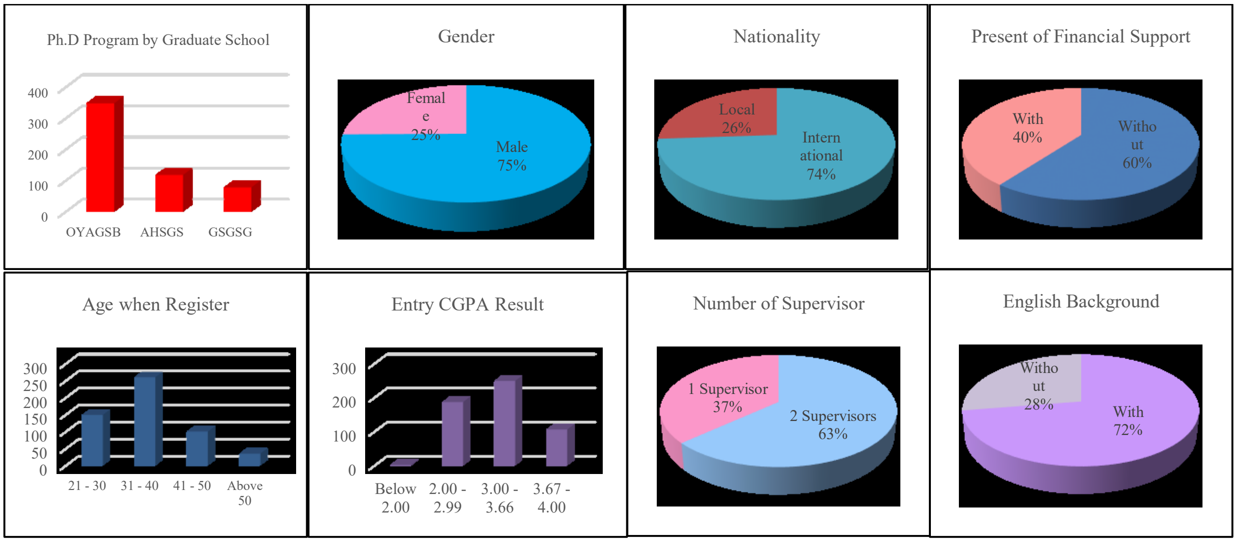

FIGURE 1. Demographic of the Ph.D students

From the data, we have constructed some preamble analysis to screen the information of each variable. We found that 75% of respondents are males and majority of them are not the local students. The preference age of further their study are 31 to 40. About 63% of the students are having second supervisor along their study in university. Besides, 60% of the students have no financial support which a barrier for the students could be to accomplish their GOT [12].

In this study, there are four decision tree models are tested by using different algorithms. The tree models' performance is assessed by using five aspects, which are accuracy, sensitivity, specificity, ROC Index and ASE. TABLE 2 and TABLE 3 show the evaluation results of the decision tree models from training data (80%) and validation data (20%).

TABLE 2. Evaluation results from training data for four decision tree models

	Training Data (80%)							
	Chi-square		Gini Index		Entropy		Interactive	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
True	273	39	279	37	275	44	253	59
False	96	25	98	19	91	23	76	45
Accuracy	72%		73%		74%		72%	
Sensitivity	92%		94%		92%		85%	
Specificity	29%		27%		33%		44%	
ROC Index	64%		65%		71%		68%	
ASE	20%		19%		18%		19%	

TABLE 3. Evaluation results from validation data for four decision tree models

	Validation Data (20%)							
	Chi-square		Gini Index		Entropy		Interactive	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
True	70	7	72	7	72	8	65	11
False	28	6	28	4	27	4	24	11
Accuracy	69%		71%		72%		68%	
Sensitivity	92%		95%		95%		86%	
Specificity	20%		20%		23%		31%	
ROC Index	61%		62%		69%		58%	
ASE	21%		20%		19%		20%	

The output shows that decision tree (Entropy) is the most potential tree model among all trees. For accuracy criteria, it scores the highest for both training data (74%) and validation data (72%). Besides, for sensitivity analysis, it scores second highest score with 92% for training data and highest score with 95% for validation data. While for analysis of specificity, it stands on the second place with the score of 33% for training data and 23% for validation data. Moreover, it has the highest ROC index for both training and validation data, with 71% and 69% respectively. Lastly, for ASE criteria, it scores 18% in training data and 19% in validation data, which are the lowest ASE values compared to the others. Therefore, decision tree (Entropy) model is selected as the predictive model for classifying those "GOT achiever" and "non-GOT achiever".

FIGURE 2 below shows the structure of decision tree model with Entropy algorithm. The students with the age range from 31 to 50 are more likely to be "GOT achiever". Besides, those with Entry CGPA Result less than 3.395 has higher ability to be "GOT achiever" than cases with Entry CGPA Result greater than or equal to 3.395. Also, male students without English Background are more likely to be "GOT achiever". Moreover, female students without English Background and Entry CGPA Result less than 3.275 are "non-GOT achiever". Surprisingly, female Ph.D students without English Background and Entry CGPA Result greater than or equal to 3.41 are tend to be "non-GOT achiever". The result also points out that female students without English Background and Entry CGPA Result between 3.275 and 3.41 opt to be "GOT achiever".

Decision Tree Model (Entropy)

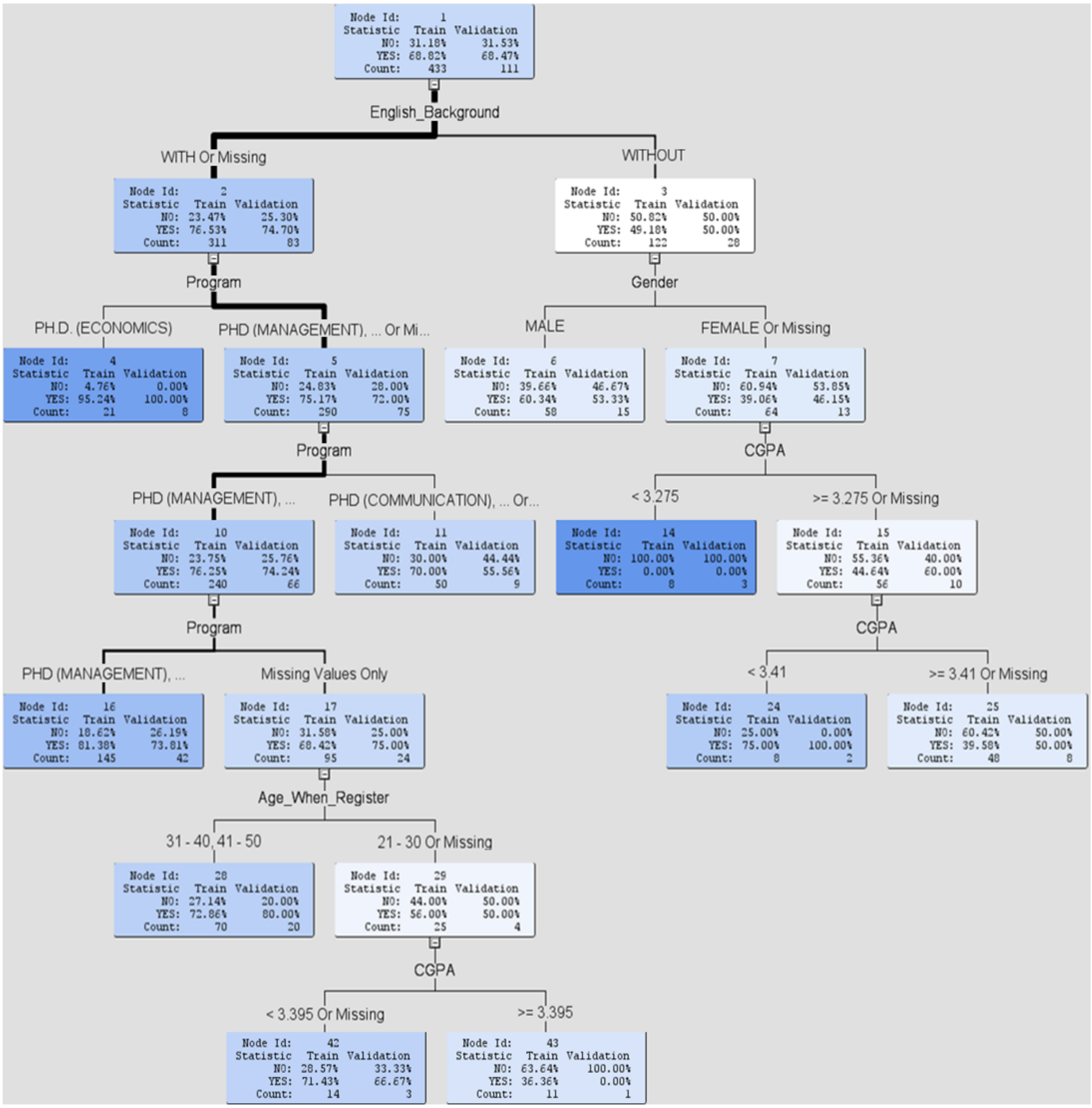

FIGURE 2. Decision tree structure with Entropy algorithm

DISCUSSION

The decision tree models illustrated that Ph.D students with English Background are mainly “GOT achiever”. This result matches the finding from the study of [13]. Besides, the result also shows that the gender are essential in affecting the completion time of their studies. The output of the decision trees are associated with the research conducted by [14] and prove that male student has higher possibility to complete Ph.D study on time than female student. Moreover, the Ph.D students’ previous academic performance which represented by *Entry CGPA Result* in this study are also influencing the GOT mission.

Decision tree model which is constructed by Entropy algorithm perform the best with the highest accuracy, sensitivity, and ROC index. Therefore, it has been selected as the best model for predicting the ability of the Ph.D students in achieving GOT. The influencing attributes according to the level of importance gained are English Background, Gender, Program, CGPA and Age (from the date of registration).

CONCLUSION

Good education is vital important to bring a country to move forward to economy advancement and flourish. In Malaysia, to achieve the mission of being high income status by year 2020, Ministry of Higher Education Malaysia has targeted to produce 60,000 Ph.D graduates by year 2020. Consequently, as the number of Ph.D students increased, the ability of postgraduates to complete their studies has become a constraint to the students, lecturers, supervisors, college, school and university. Therefore, the main objectives of this study is to classify the Ph.D students into "GOT achiever" and "non-GOT achiever" according to their students' profile from the historical data that retrieved from data base UUM (GAIS). The finding of this study shows that English background, gender and entry CGPA result are the main factors that influence the ability of the Ph.D students in achieving GOT. Also, the result indicates that decision tree model with Entropy algorithm perform the best. The model can be used by the university to uncover the restriction in this issue so that better plans can be carried out to boost the number of GOT achiever in future.

ACKNOWLEDGMENTS

We would like to acknowledge the work that led to this paper, which is fully funded by the Research Generation University Grant of Universiti Utara Malaysia.

REFERENCES

1. Education in Malaysia, see http://wenr.wes.org/2014/12/education-in-malaysia
2. N. Z. Abiddin and A. Ismail, International Review of Social Sciences and Humanities **1(1)**, pp. 15-29 (2011).
3. W. Y. Chin, C. K. Ch'ng, J. M. Jamil and I. N. M. Shaharanee, "*Analyzing the factors that influencing the success of post graduates in achieving graduate on time (GOT) using analytic hierarchy process (AHP)*", in Proceedings of the 13th IMT-GT International Conference of Mathematics, Statistics and Its Applications (2017) (ICMSA, 2017), (Universiti Utara Malaysia, Malaysia).
4. G. K. Gupta, *Introduction to data mining with case studies* (PHI Learning Pvt. Ltd., 2006).
5. K. Cawley, When to use supervised and unsupervised data mining, see https://www.predictiveanalyticsworld.com/patimes/use-supervised- unsupervised-data-mining/4046/
6. S. Kathait, Introduction to reinforcement learning, see https://valiancesolutions.com/introduction-to-reinforcement-learning/
7. J. Han and M. Kamber, *Data mining concepts and techniques* (Morgan Kaufmann Publishers, Los Angeles, 2001).
8. P. N. Tan, M. Steinbach, A. Karpatne and V. Kumar, *Introduction to data mining (Second Edition)* (Pearson Education India, 2018).
9. H. Du, *Data mining techniques and applications: an introduction* (Cengage Learning, Boston, 2010).
10. L. Rokach and O. Maimon, *Data Mining with decision trees theory and applications (Vols. 69)* (World Scientific, Singapore, 2008).
11. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and regression tree* (Wadsworth, Inc., Monterey, Calif., U.S.A., 1984).
12. W. Y. Chin, C. K. Ch'ng and J. M. Jamil, The Journal of Social Sciences Research, pp. 1186-1193 (2018).
13. J. Rodwell and R. Neumann, *Predicting timely doctoral completions: an institutional case study of 2000-2005 doctoral graduates* (2005).
14. V. Jiranek, International Journal of Doctoral Studies **5**, pp. 1-13 (2010).