Identifying Patterns of Students Academic Performance from Tracer Evaluation using Descriptive Data Mining

Fadzilah Siraj, Nur Azzah Abu Bakar

Abstract: ***The Ministry of Higher Education Malaysia has collected data through tracer study since 2007. The aim is to gather feedbacks from graduates as a basis improve to basis in improving. The availability of tracer study data in digital format offers various advantages to decision makers as many tools are available to extract and discover the hidden knowledge within the large databases. This paper presents the applicability of descriptive data mining and logistic regression to discover the hidden knowledge within the tracer study data with respect to measuring academic performance of Arts and Sciences graduates of Malaysia public universities. The impact of independent variables, i.e. Bahasa Melayu, English Language and Malaysian University English Test on the academic performance is investigated. The empirical results suggest that the academic performance between male and female graduates from Arts and Science fields is significantly different. Variables such as Bahasa Melayu, English Language and Malaysian University English Test showed a significant correlation with academic performance. The results also exhibit that the impact on academic performance of Arts graduates is different from the Science graduates. Guided by these empirical findings, this study suggests an academic performance model for Arts and Science graduates of Malaysia public universities.***

Keywords: Academic Performance, Descriptive Data Mining, Logistic Regression, Tracer Study

I. INTRODUCTION

Student academic performance has long been regarded as an essential research topic in many academic disciplines for a number of reasons. The analysis of academic performance can help the educators to get some insight about student academic performance and the management to performance and the information obtained through the analysis could be undertaken by the management (Veenstra, 2009) could undertake the information obtained through the analysis. With the understanding about the impact of variables on academic performance, the university can identify intervention programs to accelerate the students' academic performance for the high and low achievers (Lowis & Castley, 2008). In addition, an instructor can utilize the findings to modify existing course curriculum.

As a means to improve the standard of higher education, the Ministry of Higher Education Malaysia (MOHE) has collected data from series of tracer study since 2007, involving all public and private universities. The survey captures feedbacks from graduates with respect to the program of study they have gone to undertaken; In MOHE databases to MOHE servers; the type of their current job and how they fare in their working world. The data were collected via online questionnaire and kept in the MOHE databases.

As higher learning becomes highly competitive, and as students' data become more accessible, it is an advantage to the higher learning institution (HIL) if such data can be used to provide information and thus knowledge in facilitating decision making. This study explores the impact of attributes in a tracer study on the academic performance of Arts and Science graduates. In essence, the impact of Independent Variables (IV), i.e. Bahasa Melayu, English Language and Malaysian University English Test (MUET) on the academic performance is investigated.

II. RELATED WORKS

To date, the academic performance is still one of the active research area (Siraj & Haris, 2011; Corengia, Pita, Measured, 2013; Siraj, 2015; Asif, Merceron & Pathan, 2015; Al-Ansari & El-Tentawi, 2015) and Cumulated Grade Point Average (CGPA) becomes one of the most important indicators for individual success (Nordaliela, Zaidah & Roziah, 2008). Numerous studies have been conducted to investigate factors that affect academic achievement (Siraj, 2015; Yadav & Sing, 2012). Factors such as gender shows different academic performance; female students show better performance in all examined academic subjects compared to male students (Shipley, Jackson & Segrest, 2008). Various other factors also have an impact on academic performance such as students socioeconomic background (Katsikan & Panagiotidis, 2010), attendance of lectures, knowledge of English, income of parents, perceptions of learning, attitudes of students and lecturers towards education, teaching aids and methods as well as environmental factors (Weerakkody & Ediriweera, 2008). In this study, descriptive data mining was explored in providing more insight in conjunction with academic achievement.

Revised Manuscript Received on June 22, 2019.

Fadzilah Siraj, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia

Nur Azzah Abu Bakar, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia

III. METHODOLOGY

Data mining (DM) is defined as computer automated exploratory data analysis of large complex data sets that can be used to discover patterns and relationships in data with an emphasis on large observational databases (Spangler et al, 2011). The tasks of DM can be modeled as either Predictive or Descriptive in nature (Luan, 2002). A Predictive model makes a prediction about values of data using known results found from different data. while the Descriptive DM is used to discover interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data (Delavari & Phon-amnuaisuk, 2008). In education, Descriptive DM is used to determine the demographic influence on factors of academic performance (Ervina & Md Nor, 2005).

The task of quantifying the variables that affect academic performance may be constrained by the nature of the graduates' data under consideration. Although, several statistical techniques are available only a few conform to the outcome of non-continuous DV. Thus, logistic regression is utilized in the study since logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable (DV) and one or more categorical or continuous predictor variables (IV) (Chao-Ying, 2002). In fact, logistic regression has been used in educational research as reported by Peng & So (2002) and Peng, So, Stage and St. John (2002). Correlation and cross-tabulation analysis are mainly used to support descriptive DM.

Following Siraj (2015), the general approach of this study involves 6 steps as shown in Figure 1.

Figure 1: The approach of the study

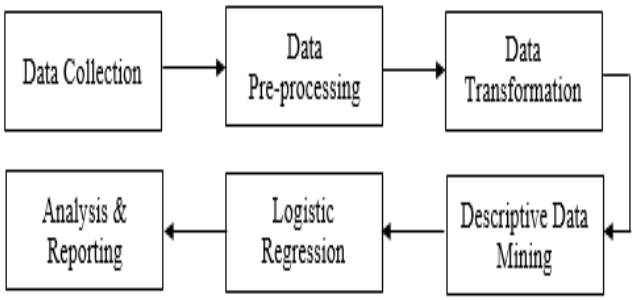

Fig. 1 The approach of the study

This study utilizes population data from the tracer study comprising of 22 public universities graduates who have completed their study in three consecutive years (2007 to 2009). Once the data were retrieved, they were cleaned, missing values were checked, and suitable attributes were selected as independent variables during **Data Pre-processing** stage. In this paper, only significant factors are reported.

For the analysis purposes, a total of 40,937 cleaned data comprising of Arts (n=26,613) and Science (n=16,324) graduates. Some of the attributes or DVs in the data required recoding and **transformation**. For example, transformation of DV was carried out prior to performing exploratory data analysis such that the CGPA was transformed into class 1 to 4 as listed in Table 1.

Table. 1 Target Class for the CGPA

CGPA Range	Class
Below-2.49	1
2.50-2.99	2
3.00-3.49	3
3.50-4.00	4

Descriptive Data Mining

Descriptive DM was performed to investigate the nature of the dataset and the distribution of each attribute. In this study, the attributes were selected based on the statistical results and the academic achievement was measured by CGPA. The comparison among graduates was conducted in accordance to the type of domains undertaken at the public universities.

Descriptive DM approach has been widely used in past studies to explore the data in as many ways as possible until a plausible story of the data emerges. The results can be used to improve the understanding of a data set mainly in the following three main aspects: (1) The results from cross tabulation analysis can be used to visualize the distribution of the data that represent the attribute of the graduates with respect to their CGPA by using SPSS version 22 and also Microsoft Excel 2013 (Ratner, 2009). (2) The correlation analysis can be used to determine the relationship between Dependent Variable (DV) and IV (Ratner, 2009). (3) The bar charts can be used to unveil the connect ion between the logistic regression, correlation and cross tabulation analysis.

Logistic Regression

Logistic Regression (LR) analysis model is also known as one of the most useful tools in quantitative analysis phase of the decision-making process (O'Connor et al., 2002). The use of LR and Descriptive DM could lead to some insight with regard to correlation and association of the attributes with the academic achievement.

Analysis and Reporting

Frequency tables were generated, and the correlation analysis has been conducted to determine the relationship between the attributes, including Cross Tabulation Analysis (contingency tables). Some related works has have been reported in Siraj (2015). In addition, logistic regression model is also presented in the Results and Discussion section.

IV. RESULTS AND DISCUSSION

The findings of this study reveal that factors such as grades of BM and English Language upon entering the university as well as MUET have significant impact on the academic performance. This study also investigates how the correlation values and the logistic regression variables signify the impact on the academic performance. The distribution of graduates with respect to the field of study and their CGPA indicates that there is not much difference between the two major fields except that more graduates of

Arts and Science obtain a CGPA between 2.50 – 2.99 (with the difference of 6.8%) compared with Science graduates. However, for the CGPA 3.50 – 4.00, the graduates from the Science study domain show better academic performance than the Arts and Science graduates (difference by 6.9%).

The cross-tabulation analysis between gender, field of study and CGPA is obtained to further understand whether there is a difference in performance between male and female graduates. The analysis results are shown in Figure 2.

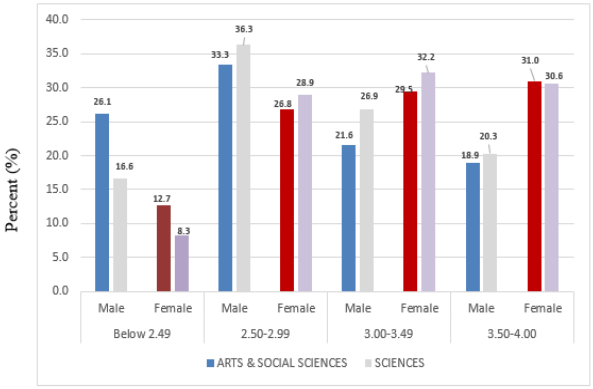

Fig. 2 The Distribution of Graduates with Respect to CGPA and gender

Among male graduates of Arts and also Science, the highest percentage gained by the graduates is the CGPA between 2.50-2.9. However, the male graduates that score Below 2.49 from the Arts show higher percentages than Science graduates (26.1% versus 16.6%). In fact, the overall performance of male graduates from Science show better academic performance than the Arts graduates. Similar observation is shown by the female graduates of Science with the lowest percentage of graduates score below 2.49. Comparing the performance of female graduates from Arts and Science, the percentage of the first field of study is higher than the latter (12.7% versus 8.3%). Comparing the performance of the female with respect to male graduates, the highest percentage scored by the female graduates is the CGPA between 3.50 and 4.00. As for the male graduates, the highest percentage shown by them is the CGPA between 2.50 – 2.99. This is true for both Arts and Science fields. Hence, the comparison analysis between the male and female graduates for both fields of study show the female graduates outperform the male graduates such that the highest percentage of female graduates score CGPA between 3.00 – 3.49. The logistic regression and correlation analysis results are depicted in Table 2. Three factors that are commonly significant to academic performance for both study fields are BM, English Language and MUET.

To get more insight how BM affects the academic performance, the cross-tabulation analysis is conducted and the results are shown in Figure 4. The correlation value (r = --. 245) for the Arts graduates is the second largest value of being among the variables listed in Table 2.

Table. 2 List of Variable Included in Students' Achievement dataset for Arts & Science for three consecutive years.

Effect	Model Fitting Criteria	Likelihood Ratio Tests				Correlation	
	-2 Log Likelihood of Reduced Model	Chi-Square	Sig.	B	Sig.	r.	Sig.
	ARTS AND SOCIAL SCIENCES						
Intercept	17004.884	1340.993	000	-3.775	000		
Gender	16175.017	511.127	000	-.157	000	.187**	.000
Education Level	16784.211	1120.321	000	.975	000	-.189**	.000
Sponsor	15684.193	20.302	000	-.035	000	-.157**	.000
Bahasa Melayu	15898.952	235.062	000	.040	000	-.245**	.000
English Language	16405.904	742.014	000	.151	000	-.284**	.000
MUET	15883.008	219.117	000	-.187	000	.210**	.000
	SCIENCES						
Intercept	17343.764	31.233	.000	.863	.000		
Gender	17657.491	344.960	.000	-.297	.000	.160**	.000
Race	17369.213	56.682	.000	-.187	.000	.238**	.000
Entry Qualification	17641.549	329.019	.000	-.097	.000	.198**	.000
Bahasa Melayu	17387.800	75.269	.000	.050	.000	-.118**	.000
English Language	17536.891	224.361	.000	.120	.000	-.297**	.000
MUET	17713.544	401.014	.000	-.307	.000	.346**	.000
Bahasa Melayu Skill	17508.634	196.103	.000	.238	.000	-.193**	.000

Figure 4 indicates that those who obtained A in Bahasa Malaysia upon entry, their academic performance is proportionally related to their CGPA. However, those who obtained a grade D for BM, a higher percentage of them achieved a CGPA between 2.50-2.99.

The correlation value for Science graduate is -1.18. Note that the negative value of r depends on how the variable under correlation analysis is represented. In this case, the value of CGPA is set from poor to excellent (from 1 to 4). However, the grade that represents Bahasa Melayu is set from A to represents poor.

Referring to Figure 4, the bar charts show that Science graduates who obtained A in Bahasa Melayu are likely to obtain a CGPA between 2.50 to 4.00. One interesting observation to note is that the graduates who obtained a grade D in Bahasa Melayu upon entering the university shows no effect on the academic performance of the Science graduates. However, this is not true for the Arts graduates.

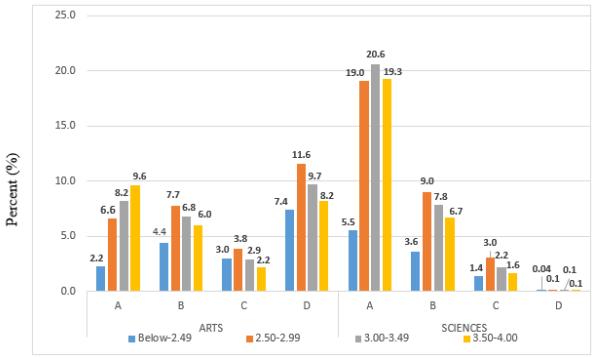

Fig. 4 The distribution of graduates with respect to Bahasa Melayu and CGPA

The cross tabulation results between English Language grade upon entering the university and the final CGPA of the graduates is depicted in Figure 5. The correlation values are-. 245** and -. 297** or the Arts, and Science graduates respectively (Table 2).

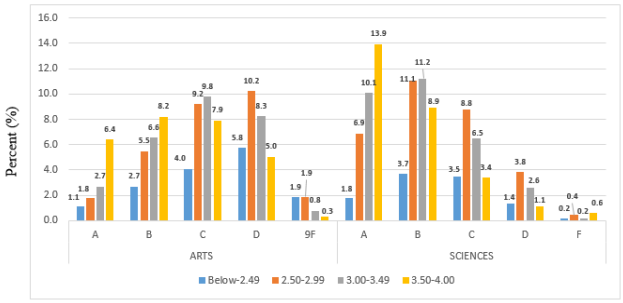

Fig. 5 The distribution of graduates with respect to English Language and CGPA

Similar observation is shown for graduates with grade A and B for Arts, and grade A in Science. However, the differences in the percentage shown by the bar charts between both study fields are less compared with Figure 5. Hence, the better English Language grades for the Arts graduates upon entering the university, the better CGPA the graduates obtained upon graduation. The analysis also includes the graduates who obtained a grade F for the English Language upon entering the universities. The results indicate that these students were able to obtain mostly CGPA between 2 – 2.99; they are mainly from Arts.

The correlation values for the Arts is 0.210 while for the Science is a 0.346 (Table 2) and the cross tabulation for MUET with respect to CGPA and study domain is shown in Figure 6.

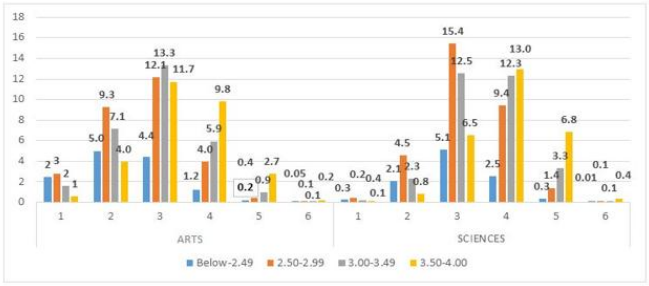

Fig. 6 The distribution of graduates with respect to MUET and CGPA

The overall MUET's band obtained by Arts graduates is 2, 3 and 4. The graphical representation also shows that less percentage of Arts graduates has Band 5 MUET compared with the Science graduates. In fact, higher percentage of Arts graduate, with Band 1 MUET than Science graduates. Although, the MUET's band is very low, some of the graduates manage to get CGPA between 3.50 –4.00. However, most of them from this cohort obtained CGPA between 2.00 – 2.99.

Comparing the Science with Arts graduates, higher percentage of Science graduates has Band 5 MUET. Those who have MUET's band 4 and 5, the graduates have a better tendency to obtain a CGPA between 3.50 and 4.00.

The occurrence of a relationship between the CGPA and combination of independent variables is based on the statistical significance of the final model chi-square. In this analysis, the probability of the model chi-square (4771.09) is 0.000, less than or equal to the level of significance of 0.05. The null hypothesis that there is no difference between the model without the IV and the model with DV was rejected. The existence of a relationship between IV (Gender, Education Level, Sponsor, BM, BI, MUET) and DV (CGPA) is supported. The independent variables are linearly related to the log odds of the dependent variable (for each independent variable)/Based on the information exhibited in Table 2, therefore the logistic regression model for the Arts graduates can be written as

*P(arts) = Exp (-3.775 - 0.157 * Gender + 0.975 * Edn Level – 0.035 * Sponsor + 0.040 * BM + 0.151 * BI - 0.187 * MUET)*

*(1 - 3.775 - 0.157 * Gender + 0.975 * Education Level – 0.035 * Sponsor + 0.040 * BM + 0.151 * BI - 0.187 * MUET)* (6)

The probability of the chi-square model for Science graduates (3129.148) is less than or equal to the level of significance of 0.05. Therefore, there is a relationship between IV (Gender, Race, Entry Qualification, BM, BI, MUET, BM Skills) and DV (CGPA). Based on the information provide in Table 2, the logistic regression equation for the science graduates is written as

*P (science) = Ex ((0.863 – 0.297 * Gender – 0.187 * Race – 0.097 * Entry Qual + 0.050 * BM + 0.120* BI- 0.307 * MUET + 0.238 * BM Skill)*

*(1 + 0.863 – 0.297 * Gender – 0.187 * Race – 0.097 * Entry + 0.050 * BM + 0.120* BI - 0.307 * MUET + 0.238 * BM Skill)* (7)

Based on the results exhibited in Table 2 and the analysis conducted for Arts and Science graduates, the summary of variable that are common to both Arts and Science graduates can be summarized as shown in Figure 7.

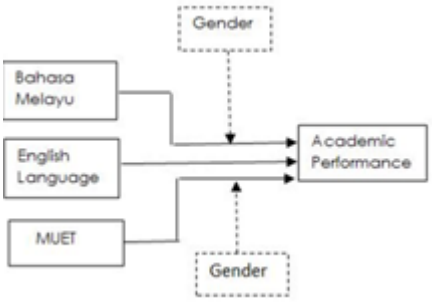

Fig. 7 The academic performance model for Arts and Science graduates

Figure 7 indicates that although Bahasa Melayu, English Language and MUET are significant IV for academic performance, the empirical results indicate that the influence of gender on the model cannot be ignored. Thus, further work should focus on the impact of gender on the IV. It is recommended that the analysis will focus on the effect for each cohort of gender, so that more information can be uncovered.

V. CONCLUSION

Empirical investigation results reveal that three IV has a significant correlation with academic performance.

However, the overall performance of the Arts and Science graduates based on gender is significantly different. In general, the female graduates obtained higher CGPA when compared to female graduates for both Arts and Science study fields.

Comparing graduates of Arts and Science with respect to the entry requirement upon entering the university, Science graduates acquired better grades in Bahasa Melayu, English and MUET. These students, in fact obtain a better CGPA than the Arts students. The findings suggest that more exploratory study should focus on gender, and determine why there is a difference in academic performance between gender, entry requirements and field of study. Perhaps with a more detail analysis, special intervention program could be formulated to reduce the gaps between gender and field of study.

REFERENCES

1. Al-Ansari, A.A. & El Tentawi, M.M. (2015). Predicting Academic Performance of Dental Stu dents using Perception of Educational Environment. *Journal of Dental Education*, March, 337-344.
2. Asif, R., Merceron, A. & Pathan, M. K. (2015). Predicting Student Academic Performance at Degree Level: A Case Study. *I.J. Intelligent Systems and Applications*, 01, 49-61.
3. Chao-Ying, J. P. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 9(1), 3-14.
4. Corengia, A., Pita, M., Mesurado, Belen. & , A. (2013). A Predicting Academic Performance and Attrition in Undergraduate Students. Liberalist. Rebista de Psicologia, 19(1), pp. 101-112. Universidad de San Martin de Porres. Lima, Peru.
5. Debreceny, R. S., Gray, G. L.: Data mining journal entries for fraud detection: An exploratory study. Int. J. Account. Inf. Syst. (2010). Vol. 11 (3) 157–181.
6. Delavari, N., Phon-amnuaisuk, S.: Data Mining Application in Higher Learning Institutions. (2008). 7(1). 31–54.
7. Ervina, A., Md Nor, O.: Undergraduate students' performance: the case of University of Malaya. Quality Assurance in Education. (2005). 13(4). 329-343.
8. Katsikan, G. & Panagiotidis, L. (2010). Factors Affecting Students' Quality of Academic Performance: A Case of Secondary School Level. Journal of Quality and Technology Management.
9. 7(6). pp. 1 - 14
10. Lowis, M. & Castley, A. (2008) Factors Affecting Student Progression and Achievement: Prediction and Intervention. A two-year study. *Innovations in Education and Teaching International*, Volume 45 (Number 4). 333-343. ISSN 1470-3297.
11. Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Dir. Institutional Res. (2002). 2002(113). 17–36.
12. Nordaliela, M. R., Zaidah, I. & Roziah, M. J. (2008). Predicting Students' Academic Achievement: Comparison between Logistic Regression, Artificial Neural Network, and Neuro-Fuzzy. *In Proceedings of International Symposium on Information Technology (ITSim 2008)*. Vol. 1. 1 – 6
13. O'Connor, M., Marquez, I., Hill, T., Remus, W.: (2002). Neural network models for forecast a review, IEEE proceedings of the 25th Hawaii International Conference on System Sciences. Vol. 4. 494-498.
14. Peng, C. Y. & So, T. S. H. (2002). Modeling strategies in logistic regression. *Journal of Modern Applied Statistical Methods*, 14, 147-156.
15. Peng, C. Y., So, T. S. Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education*, 43, 259-293.
16. Ratner, G. (2009). *Journal of Targeting, Measurement and Analysis for Marketing* (2009) **17,** 139–142. doi:10.1057/jt.2009.5; published online 18 May 2009
17. Siraj, F. & Haris, M.F. (2011). Profiling of UUM Graduates Based on Academic Achievement and Colleges. *Proceeding of the International Soft Science Conference (ISSC 2011)*, Ho Chin Minh, Vietnam.
18. Siraj, F. (2015). Modeling Academic Achievement of UUM Graduate Using Descriptive and Predictive Data Mining, *Advanced Computer and Communication Engineering Technology*, Volume 362 of the series Lecture Notes in Electrical Engineering, Eds, Hamzah Asyrani Sulaiman, Mohd Azlishah Othman, Mohd Fairuz Iskandar Othman, Yahaya Abd Rahim, Naim Che Pee. Springer, pp 609-62.
19. Spangler, S.,Ying, C., Kreulen, J., Boyer, S., Griffin, T., Alba, A., Kato, L., Lelescu, A., S. Yan, S. Exploratory analytics on patent data sets using the SIMPLE platform. World Pat. Inf. (2011). Vol. 33. No. 4. 328–339.
20. Veenstra (2009). A Strategy for Improving Freshment Colleague Retention. *The Journal for Quality & Participation*, January, 19-29.
21. Weerakkody, W.A.S. & Ediriweera, A.N. (2008). Influence of gender on academic performance: an empirical study of Human Resource Management students (undergraduates) in University of Kelaniya, Sri Lanka. *Proceedings of the 5th International Conference on Business Management*, Vol. 5.
22. Yadav, R. S. & Sing, V., P. (2012). Modeling Academic Performance Evaluation Using Fuzzy C-Means Clustering Technique. *International Journal of Computer Applications*, 60(8), 15-23.