BINARY LOGISTIC REGRESSION MODELLING WITH APPROPRIATE SAMPLE SIZE IN DETERMINING GRADUATE EMPLOYABILITY FACTORS FOR PUBLIC UNIVERSITIES IN MALAYSIA

TENGKU SALBIAH BINTI TENGKU MOHAMED

UNIVERSITI TEKNOLOGI MALAYSIA

BINARY LOGISTIC REGRESSION MODELLING WITH APPROPRIATE SAMPLE SIZE IN DETERMINING GRADUATE EMPLOYABILITY FACTORS FOR PUBLIC UNIVERSITIES IN MALAYSIA

TENGKU SALBIAH BINTI TENGKU MOHAMED

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Philosophy

Faculty of Science
Universiti Teknologi Malaysia

AUGUST 2020

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, Professor Dr. Muhammad Hisyam Lee for all the guidance, support and advice throughout my journey in taking Master study in Universiti Teknologi Malaysia. He has contributed a lot towards my understanding and thoughts, not only on the academic field but also on my life. He had inspired me a lot through his passion for research, respectful personality and thinking skill. Throughout my journey in completing the study, I really enjoyed working with him and inspired me to achieve my dream.

I would like to express my wholehearted gratitude to my parents for their endless love, support, and guidance throughout my life. They are the main source of my strength and courage to be a successful person in this life. Not forgetting to my brothers and sisters which always gives words of enthusiasm and always encourage me to pursuit my life goal.

I would also like to thank Universiti Teknologi Malaysia (UTM) Zamalah for funding my Master and always encourage me to complete this study. And last but not least, my sincere appreciation also extends to all my friends especially to Airen Suhailah, Yusrina Andu and Nur Farzana for their friendship.

ABSTRACT

The performance of variable selection is essential to build an effective logistic regression model. Generally, p-values are used to identify significant variables or factors in the model. However, when dealing with real tracer study data for a country, the size of the data is typically large of which causes the p-values to be deflated and affect the variable selection performance. Therefore, it is crucial to have an appropriate sample size and sampling ratio for this purpose. In this study, the appropriate sample size has been proposed based on simulated correlation tests and significant variables in order to improve the accuracy of variable selection. In addition, the sampling ratio in the response variable shows its best when it reflects the population ratio. Based on the proposed samples, the logistic regression model for graduate employability factor is subsequently proposed. It has been found that age, Cumulative Grade Point Average (CGPA), discipline of study, gender, state, and type of universities are the factors that significantly affect graduate employability among public universities in Malaysia. The results show that the proposed model has successfully improved the variable selection, model fitting, and classification accuracy as compared to the full model. Thus, by using a smaller sample size, the proposed model is able to maintain its statistical power in real data scenario by accurately selecting the significant factors.

ABSTRAK

Prestasi pemilihan pembolehubah adalah penting untuk membina model regresi logistik yang berkesan. Umumnya, nilai-p digunakan untuk mengenal pasti pemboleh ubah atau faktor yang signifikan dalam model. Namun, ketika berhadapan dengan data kajian pengesanan sebenar untuk sesebuah negara, saiz data biasanya besar dimana akan menyebabkan nilai-p mengecil dan akan memberi kesan pada prestasi pemilihan pembolehubah. Oleh itu, sangat penting untuk mempunyai ukuran sampel dan nisbah persampelan yang sesuai untuk tujuan ini. Dalam kajian ini, saiz sampel yang sesuai telah dicadangkan berdasarkan ujian korelasi simulasi dan pembolehubah yang signifikan untuk meningkatkan ketepatan pemilihan pembolehubah. Di samping itu, nisbah persampelan dalam pembolehubah tindak balas menunjukkan yang terbaik apabila ia mencerminkan nisbah populasi. Berdasarkan sampel yang dicadangkan, model regresi logistik untuk faktor kebolehpasaran siswazah dalam kalangan graduan kemudiannya dicadangkan. Telah didapati bahawa umur, Purata Nilai Gred Kumulatif (PNGK), disiplin pengajian, jantina, negeri, dan jenis universiti adalah faktor yang sangat mempengaruhi kebolehpasaran siswazah antara universiti awam di Malaysia. Hasil kajian menunjukkan bahawa model yang dicadangkan telah berjaya meningkatkan prestasi pemilihan pembolehubah, pemasangan model dan ketepatan klasifikasi berbanding dengan model penuh. Oleh itu, dengan penggunaan saiz sampel yang lebih kecil, model yang dicadangkan dapat mengekalkan kekuatan statistiknya dalam senario data sebenar dengan mengesan faktor-faktor yang signifikan dengan tepat.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	-	Akaike Information Criteria
CA	-	Classification Accuracy
CGPA	-	Cumulative Grade Point Average
FP	-	False Positive
ME	-	Misclassification Error
MoE	-	Ministry of Education
MTUN	-	Malaysia Technical University Network
ROC	-	Receiver Operating Characteristics
TP	-	True Positive
TN	-	True Negative
WPKL	-	Wilayah Persekutuan Kuala Lumpur
WPP	-	Wilayah Persekutuan Putrajaya
WPS	-	Wilayah Persekutuan Labuan Sabah

LIST OF SYMBOLS

β	-	Parameter estimates for graduate’s factors
D	-	Cook Distance
n	-	Observations number
p	-	Probability of graduates to get employed
ρ	-	Linear relationship strength between variables
$\widehat{se}$	-	Standard Error
s^2	-	Standardized residual
τ	-	Kendall Tau Correlation Coefficient
x	-	Explanatory Variable for the graduate’s factors
y	-	Response Variable for the graduate’s status

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Statistical modelling has been widely applied to explain existing phenomenon in a mathematical-formalised way. Logistic regression is one of the statistical methods used to build models when the categorical data is a subject of interest in the response variable, whilst binary logistic regression is a method used when the response variable has two outcomes to be considered. This method belongs to the family of generalised linear models and has been extensively used until now, especially in the Biostatistics field (Beitia-Antero et al., 2018).

To build a good statistical model, a large sample is required in order to have efficient, representative, reliable and flexible results. Over the past decades, generated data collection systems have become common places for the production of immediate data on a large scale in various fields such as science, management, social, and environment. In the education field, for instance, the rising number of institutions has caused the number of graduates produced per year to escalate rapidly. Hence, recorded information is produced at a large scale. Despite that, in statistics, whole data collected may not necessarily have veracity and value (Hsu et al., 2019).

Data that has a large sample size is highly related to increased model statistical power whereby when the power is high, it will increase the likelihood of detecting the effect of independent variables on the response variable when there is an effect to be detected (Lorca-puls et al., 2018). However, in some cases, it does not always hold true; as data become larger or too large, the model to be built will be over in both parameterisation and estimation (Heckmann et al., 2014). This is due to the issue of a deflated *p*-value that has been critiqued as it becomes smaller when the sample size becomes larger. Thus, there is the tendency of increased probability of more variables falling into false positives as they falsely indicate that there is a significant effect when there is none. Hence, the model built is less reliable as the method is inaccurate in selecting the significant variables (Kirby & Sonderegger, 2018).

A good statistical model should only select the important variables that can give high variance in explaining the response variable and give the most accurate estimation to reflect the population scenario. However, the issue of p-value in large data can cause built models such as logistic regression models to have a high number of important variables but are poor in discriminating the groups in the response variable. Since the p-value is the main approach taken to measure variable significance, several suggestions have been made to address this matter (Park et al., 2018) such as lowering the α value than the traditional threshold at $\alpha = 0.005$, using intervals for null hypothesis rather than single value, and using small sample sizes to obtain a standardised p-value. Thus, the probability of rejecting the null hypothesis will be more stringent and the probability of making false positive can be reduced.

In contrast, Lazic (2017) said that lowering the significance threshold could reduce the statistical power of the model. Meanwhile, Solmi et al., (2019) said that setting $\alpha = 0.005$ is not stringent enough. There is also a lack of consensus regarding the best α value for this subject matter. Fisher and Neyman-Pearson said that such probability,p is used to test either the effect to be detected is due to random effects or there is indeed a significant effect in real life. Based on this statement, it can be concluded that the p-value does not indicate that the variable being investigated is important or otherwise. In fact, the p-value is an indicator used to go against the null hypothesis when there is inevitable sampling variability (Domenech, 2018). It can also be said as an effect that occurs towards the response variable when there is a given magnitude based on the sample used (Kirby & Sonderegger, 2018).

Furthermore, in the logistic regression method, the proportion of two groups in the response variable can also affect the statistical power of a model. Based on a study by Nad & Ka, (2018), for their hard-to-detect study, it has been found that different weights of sampling ratio will significantly produce different coefficient estimates and probability for each observation to fall into an interest group. Even so, the underlying proportion is rarely emphasised on by researchers when using this method, especially in experimenting with data sample techniques that can affect the sampling ratio. In Nad & Ka, (2018) study, a unique census dataset that contained all the used roosting cavities of the tree-dwelling bat *Nyctalus leisleri* and all cavities where the species was absent was used. Several logistic models were constructed with varying ratios of occupied and unoccupied cavities to investigate the effect of using different sample ratios.

Based on their hard-to-detect study, it has been found that the sample ratio in the response variable should reflect the actual population ratio; otherwise, it will cause low predictive power to the logistic regression model and the conclusions to be made become less reliable. In addition, Nad & Ka, (2018) found that setting the proportion at 1:1 is the most improper method as it will produce a model that is far from what it should reflect (existing phenomenon) and inaccurate in variable selection.

On the other hand, issues on large data size can be such as the growth of institutions over the years that has resulted in a large number of graduates. This situation has led to an increase in competition among graduates in securing their first job. The issue of graduate employability has been disputed from time to time either in terms of their academic performance, the ability of tertiary schools in producing more employable graduates, and the skills needed by the industry. Employability among graduates has been highly associated with tertiary education’s ability to provide graduates who fulfil the basic prerequisites made by employers or the demands of the labor market. Although the issue of unemployment among graduates has long been discussed, this issue still holds great concern in many countries because they form the backbone of professional human resources for every country.

Unemployment among graduates can be caused by individual factors such as academic background, skills, experience, demographics, attitude and aptitude. Based on previous studies, CGPA and gender have repeatedly been found to be significant factors of graduate employability (Pinto & Ramalheira, 2017; Piad & Ballera, 2016; Hashim et al., 2015; Sapaat et al., 2011). In addition, it has been found that low CGPA is a factor depriving graduates of attaining career goals, other than work experience

(Yusof & Jamaluddin, 2015). However, as information technology (IT) has a pervasive influence in current global market, it has been found that technical skills and higher-order thinking skills are among the most significant factors for IT graduate employability (Kumar & Khurana, 2017).

Meanwhile, in terms of the method used, it cannot be denied that through its benefits in terms of classification as well as simultaneous prediction, logistic regression has been frequently applied in determining the factors affecting graduate employability. For example, logistic regression models have been applied in a study on predicting the factors affecting employability among IT graduates by Piad & Ballera (2016), job attainment for Bachelor holders in Australia (Jackson, 2014) and factor affecting the employability of people with epilepsy (Wo et al., 2016). Moreover, in India, logistic modelling was applied to give more understanding on the matter from the psychological point of view (Pandit at al., 2015). In Slovenia, the effectiveness of higher education, democracy change and economic crisis in affecting employability among political sciences graduates (Dezˇelan & Hafner, 2014) was studied using logistic regression.

In Malaysia, the issue of graduate employability is actively discussed as the level of unemployment among youths is globally on the rise since 2010 (Machart, 2017). The percentage of jobless youths had risen from 10.78% up to 11.18% between 2016 and 2018 (World Bank, 2019). Besides, statistics show that the rate of unemployment in Malaysia has kept increasing from the year 2014 up to 2016 at 2.9%, 3.1% and 3.5% respectively. Moreover, Malaysia's Minister of Human Resources, Datuk Seri Richard Riot said that with the 3.5% unemployment level in Malaysia for 2016, over 200,000 out of 500,000 graduates are considered unemployed (Carvalho et al., 2017). This can be supported by statistics from Malaysia's Ministry of Education (MoE) website which shows that 238,187 graduates were unemployed in the year 2016.

In Malaysia, a tracer study will be conducted by MoE every year to trace the status of graduates from all institutions in Malaysia. Such data from MoE is genuine and the best available data to be analysed to investigate factors behind graduate employability. Moreover, MoE’s mission which is to produce competent graduates in order to fulfil national and international manpower needs with 75% graduates employed in their relevant fields within six months of graduation has yet to be achieved.

Motivated by the unemployment issue among graduates, it is important to investigate the factors affecting graduate employability based on their profiles. However, conducting the data tracer study which involved 112,547 total observations together with their multiple record profiles did not necessarily mean that the entire data useful in providing information in terms of graduate employability. In addition, the size of the data itself can produce unreliable results in terms of variable selection due to its deflated p-value. In Zhou and Li's (2016) study, the number of samples needed was determined using a supervised learning approach in order to build an effective logistic model. Furthermore, the effect of sampling ratio in logistic regression modelling for the data tracer study conducted for 2016 also needs to be verified to build an effective model.

1.2 Problem Statement

To build an effective logistic regression model, sample size and sampling ratio do play vital roles in accurately determining factors of graduate employability. However, as mentioned in the study background, when the data sample is too large, there will be an issue on the deflated p-value. As a result, a graduate employability model to be built may be inaccurate in terms of variable selection. Besides, directly applying the logistic regression method using the actual amount of data may produce unreliable results. In addition, the effect of different sampling ratios on response variables also needs to be discovered to improve the statistical power of the logistic regression model for tracer study data. Meanwhile, regarding the factors of graduate employability, the factors of CGPA and gender need to be identified further in terms of their effect towards graduate employability status since gender is a norm for human nature, and graduates with low CGPA can be a great disadvantage in reaching graduates’ preferred career path.

1.3 Objective of Study

The objectives of the research are:

(a) To improve variable selection in logistic regression model by proposing the appropriate sample size.

(b) To improve the power of the logistic regression model by determining the difference in sampling ratio for tracer study data.

(c) To determine the factors affecting graduate employability by evaluating the performance of the proposed approach.

(d) To identify the factors affecting graduate employability in terms of low CGPA and gender group.

1.4 Scope of Study

This study focuses on modelling the tracer study data to improve variable selection when dealing with large sample sizes. Data used in this study is tracer study data from MoE based on the questionnaire version 2016. The analysis done throughout this thesis used the values given by the MoE and the status of graduates was taken within six months of their graduation. In this study, the search for the appropriate sample size was done through several processes. R software is used in this study for data analysis. Meanwhile, logistic regression is the main method to be used and applied on Malaysia’s tracer study data to achieve the objectives of the study. Assessment of the model’s performance is based on the proposed sample in terms of variable selection, classification correctness and model fitting. This is done by comparing the proposed model with a model built from the actual sample data.

1.5 Limitation of Study

This study has limitations, especially on the data used. The study only uses tracer study data for the year 2016 to determine factors of graduate employability. Besides that, in this study, graduate employability is defined based on MoE’s definition, which is either they are employed or unemployed within six months of their graduation. Furthermore, in terms of determining the appropriate sample size to improve variable selection, the performance of the proposed method for this study is only tested using tracer study data for the year 2016. Thereupon, the models obtained represent a model of graduate employability among Malaysian public university graduates for 2016.

1.6 Significance of Study

The development of an effective statistical model is crucial to determine the factors affecting graduate employability as well as the accuracy of classifying the status of graduates. This study focuses on developing such a desired feature for good statistical modelling when dealing with large sample data sizes. The model built based on the proposed sample size and sampling ratio is expected to have improved performance in terms of significant factor selection, classification accuracy, and model fit compared to models based on the actual sample. In addition, it is believed that the procedure of building an effective model can be applied to other real data scenarios to handle the issue of the deflated p-value.

1.7 Structure of Thesis

There are five chapters included in this study. In Chapter 1, the introduction, background of study, problem statement, study objectives, scope and also significance of this study have been briefly stated. In Chapter 2, a review on literature related to the employability of graduates and sample size will be presented. Chapter 3 will focus on the research methodology, starting with describing the logistic regression model then illustrating the process of determining the appropriate sample together with several indicators that would be used for model evaluation criterion. In Chapter 4, the results of the proposed sample size and sampling ratio will be presented. In addition, these results will be explained and discussed further in terms of factor selection and classification accuracy. The selected best model will represent the final findings in determining the factors affecting graduate employability. Finally, Chapter 5 will contain the main conclusions of this study and proposals for future research.

REFERENCES

Algamal, Z. Y., & Lee, M. H. (2015). High dimensional logistic regression model using adjusted elastic net penalty. *Pakistan Journal of Statistics and Operation Research, 11*(4), 667–676.

Amin, M., Amanullah, M., & Cordeiro, G. M. (2016). Influence diagnostics in the Gamma regression model with adjusted deviance residuals. *Communications in Statistics: Simulation and Computation*, *0918*(September), 1–15.

Ang, M. C. H. (2015). Graduate employability awareness : A gendered perspective. *Procedia - Social and Behavioral Sciences, 211*(September), 192–198.

Beitia-Antero, L., Y´a˜nez, J., & Castro, D. (2018). On the use of logistic regression for stellar classification: An application to colour-colour diagram. *Exp Astron, 45*, 379–395.

Bennett, D. (2017). Employability for music graduates : Malaysian educational reform and the focus on generic skills, *International Journal of Music Education,* 35(4), 588-600.

Bernama (2015). Graduates among 400,000 currently unemployed in Malaysia, says minister. Retrieved June 1 2014 from http://www.themalaymailonline.com/malaysia/article/graduates-among-400000-currently-unemployed-in-malaysia-says-minister.

Bujang, M. A., Nadiah, S., Mohd, T., & Ikhwan, T. (2018). Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population : Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. *Malays Journal Medical Science, 25*(4), 122–130.

Carvalho, M., Sivanandam, H., Rahim, R., & Yunus, A. (2017, March 23). 500.000 currently jobless-Nation: The Star Online. Retrieved May 6, 2019, from https://www.thestar.com.my/news/nation/2017/03/23/500000-currenly-jobless-riot-number-comsidered-low-going-by-international-benchmarks.

Chen, M., Wo, M., Seang, K., Yuen, W., & Tin, C. (2016). Factors affecting the employability in people with epilepsy. *Epilepsy Research*, *128*, 6–11.

Cheong, K., Hill, C., & Fernandez-chung, R. (2016). Studies in Higher Education Employing the ‘unemployable’: Employer perceptions of Malaysian graduates, *Studies in Higher Education*, 41(12), 2253-2270

Cheong, K., Hill, C., Leong, Y., Zhang, C., Hill, C., Leong, Y., & Zhang, C. (2018). Studies in Higher Education Employment as a journey or a destination ? Interpreting graduates ' and employers ' perceptions – a Malaysia case study. *Studies in Higher Education*, *43*(4), 702–718.

Cook, R. D. (1979). Influential Observations in Linear Regression Influential Observations in Linear Regression, *Journal of the American Statistical Association*, 74(365), 169–174.

D'Aguiar, S., & Harrison, N. (2016). Returning from earning : UK graduates returning to postgraduate study, with particular respect to STEM subjects, gender and ethnicity. *Journal of Education and Work*, *29*, 584–613.

Dezˇelan, T., & Hafner, D. F. (2014). First-job educational and skill match An empirical investigation of political science. *International Journal of Manpower,* 35(4), 553-575.

Domenech, R. J. (2018). La incertidumbre de la "significación" estadística, *Revista Med ica De Chile,* (146),1184–1189.

Fan, T., & Cheng, K.-F. (2004). Tests and Variables Selection on Regression Analysis for Massive Datasets. *Proceedings of the Second Workshop on Knowledge Economy and Electronic Commerce*, 229–238.

Grazio, W. S. (2019, February 13). Poor working conditions are main global employment challenge. Retrieved April 5, 2019, from https://www.ilo.org/global/about-the-lo/newsroom/news/WCMS_670171/lange-en/index.htm

Greig, M. (2019). Factors affecting Modern Apprenticeship completion in Scotland, *International Journal of Training and development,* 23(1), 27–50.

Hamzah, A., Nadarajah, K., Mat Noor, M., & Azlan, A. A. (2015). Students' Perception of the Programme Offered by the School of Biosciences and Biotechnology, Faculty of Science and Technology, UKM. *Jurnal Pendidikan Malaysia*, *40*(2), 111–117.

Hashim, R. A., Eam, L. H., Yatim, B., Ariffin, T. F. T., Zubairi, A. M., Yon, H., & Osman, O. (2015). Estimating a prediction model for the early identification of low employability graduates in Malaysia, *Singapore Economic Review,* 60(4), 1–22.

Heckmann, T., Gegg, K., Gegg, A., & Becht, M. (2014). Sample size matters :

investigating the effect of sample size on a logistic regression susceptibility model for debris flows, *Natural Hazards and Earth System Science*.259–278.

Hosmer, D. W. Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression.* Hoboken, NJ: Wiley.

Hsu, H., Chang, Y. I., & Chen, R. (2019). Greedy active learning algorithm for logistic regression models. *Computational Statistics and Data Analysis*, *129*, 119–134.

Jackson, D. (2014). Factors influencing job attainment in recent Bachelor graduates : Evidence from Australia, *High Education*, 135–153.

Jackson, D. A. (2019). Encouraging students to draw on work experiences when articulating achievements and capabilities to enhance employability, *Australian Journal of Career Development,* 28(1), 39–50.

Kaneko, H., & Funatsu, K. (2016). Preparation of Comprehensive Data from Huge Data Sets for Predictive Soft Sensors, *Chemometrics and Intelligent Laboratory Systems*, 153, 75-81.

Kirby, J., & Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics*, *70*, 70–85.

Kumar, R., & Khurana, K. (2017). ScienceDirect Employability Skills among Information Technology Professionals : A Literature Review. *Procedia Computer Science*, *122*, 63–70.

Lazic, S. E. (2017). Four simple ways to increase power without increasing the sample size, *Laboratory Animals*, 52(6), 621-629.

Lorca-puls, D. L., Gajardo-vidal, A., White, J., Seghier, M. L., Le, A. P., Green, D. W., … Price, C. J. (2018). The impact of sample size on the reproducibility of voxel-based lesion- de fi cit mappings, *Neuropsychologia*, 101–111.

Machart, R. (2017). The implementation of industrial training in tertiary education in Malaysia : Objectives , realisations and outputs in the case of foreign language students, *International Revision Education,* (63),103–122.

Malec, L., & Kiráľová, A. (2018). Evaluating competencies of graduates in Tourism as a prerequisite for future employability, *Prague Economic Papers, 27*(2), 196–214.

Matsuda, S. (2019). Young Men ' s Employment and Their Marriage : A Comparison among Japan , South Korea , *Comparative Sociology,* (18), 204–228.

Misra, R. K., & Khurana, K. (2017). Employability Skills among Information Technology Professionals: A Literature Review. *Procedia Computer Science*, *122*, 63–70

Mugwisi, T., & Hikwa, L. (2015). A tracer study of Master of Science in Library and Information Science graduates from the National University of Science and Technology, Bulawayo, Zimbabwe. *African Journal of Library, Archives & Information Science*, *25*(2), 173–183.

Nad, L., & Ka, P. (2018). Why sampling ratio matters : Logistic regression and studies of habitat use, *Plos One,* 13(7), 1–9.

Neumann, J. von. (2016). Model selection and overfitting. *Nature Methods, 13*(9), 703–704.

Noko, P., & Ngulube, P. (2013). A vital feedback loop in educating and training archival professionals: a tracer study of records and archives management graduates in Zimbabwe. *Information Development*, *31*(3), 270–283.

Ohyver, M., Moniaga, J. V., Yunidwi, K. R., & Setiawan, M. I. (2017). Logistic regression and growth charts to determine children nutritional and stunting status: A review. *Procedia Computer Science, 116,* 232–241.

Pandit, S. A., G., P., Wallack, D. C., & Vijayalakshmi, C. (2015). Towards Understanding Employability in the Indian Context : A Preliminary Study. *Psychology and Developing Societies,* 27(1), 81-103.

Park, Y., Jeon, S., & Tae Yeon Kwon. (2018). A sample size calibration approach for the p -value problem in huge samples A sample size calibration approach for the p -value problem in huge samples, *Communication for Statistical Application and Methods,* 25(5), 545–557.

Piad, K. C., & Ballera, M. A. (2016). Predicting IT Employability Using Data Mining Techniques, *Third International Conference on Digital Information Processing, Data Mining, and Wireless Communication,* 26–30.

Pinto, L. H., & Ramalheira, D. C. (2017). Perceived employability of business graduates : The effect of academic performance and extracurricular activities, *Journal of Vocational Behaviour, 99*, 165–178.

Piper, S. K., Grittner, U., Andre, R., Riedel, N., Felix, F., Nadon, R., … Dirnagl, U. (2019). Exact replication : Foundation of science or game of chance ?, *Plos Biology,* 17(4), 1–9.

Rahmat, N., Ayub, A. R., & Buntat, Y. (2016). Employability skills constructs as job performance predictors for Malaysian polytechnic graduates : A qualitative study Employability skills constructs as job performance predictors for Malaysian polytechnic graduates : A qualitative study. *Malaysian Journal of Society and Space*, *12*(3), 154–167.

Sapaat, M. A., Mustapha, A., Ahmad, J., Chamili, K., & Muhamad, R. (2011). A classification-based graduates employability model for tracer study by MOHE. *Communications in Computer and Information Science*, *188 CCIS*(PART 1), 277–287.

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research : Differences Between Sub-Disciplines and the Impact of Potential Biases, *Journal Frontiers and Psychology,* 10, 1–13.

Shanmugam, M. (2017, March 25). Unemployment among graduates needs to be sorted out fast. *The Star Online*. Retrieved May 6, 2019, from https://www.thestar.com.my/business/business-news/2017/03/25/unemployment-among-graduates-needs-to-sorted-out-fast.

Singh, P., Thambusamy, R. X., & Ramly, A. (2014). Assessing graduates' generic skills: An indicator of employability. *Pertanika Journal of Social Science and Humanities*, *22*(3), 845–860.

Solmi, M., Correll, C. U., Carvalho, A. F., & Ioannidis, J. P. A. (2019). The role of meta-analyses and umbrella reviews in assessing the harms of psychotropic medications : Beyond qualitative synthesis. *Epidemiology and psychiatric Sciences*, 27, 537-542.

Suleman, F. (2018). The employability skills of higher education graduates : Insights into conceptual frameworks and methodological options, *High Education*, 76, 263–278.

Suryani, K., Khairudin, & Syahmaidi, E. (2017). Online tracer study of bung hatta university. *International Journal of Geomate*, *13*(37), 20–27.

Verd, J. M., Barranco, O., & Bolíbar, M. (2019). Youth unemployment and employment trajectories in Spain during the Great Recession : what are the determinants ? *Journal for Labour Market Research,* 53(4), 1-20.

Verma, P., Nankervis, A., Priyono, S., Moh, N., Connell, J., & Connell, J. (2018). Graduate work-readiness challenges in the Asia-Pacific region and the role of HRM. *Equality, DIversity and Inclusion: An International Journal*, *37*(2), 121–137.

Wang, H. Y., Zhu, R., & Ma, P. (2018). Optimal Subsampling for Large Sample Logistic Regression. *Journal of the American Statistical Association*, *113*(522), 829–844.

West, J. (2018, October 11). Employability in the 21st Century: The Global Graduate Skills Gaps and Mismatched Expectations. Retrieved May 4, 2019, from https://www.qs.com/the-global-graduate-skills-gaps.

Wo, M. C. M., Lim, K. S., Choo, W. Y., & Tan, C. T. (2016). Factors affecting the employability in people with epilepsy. *Epilepsy Research*, *128,* 6–11.

World Bank. (2019). Unemployment, youth total (% of total labor force ages 15–24) (modeled ILO estimate). Washington, DC: World Bank Group. Retrieved 21 July 2019, from http://data.worldbank.org/indicator/SL.UEM.1524.ZS.

Yorke, M. (2006). *Employability in Higher Education: What It Is - What It Is Not*. Heslington, York, United Kingdom: The Higher Education Academy United Kingdom.

Yusof, N., & Jamaluddin, Z. (2015). Graduate employability and preparedness: A case study of University of Malaysia Perlis (UNIMAP), Malaysia. *Malaysian Journal of Society and Space 11*, *11*(11), 129–143.

Zhou, Q., & Li, Z. (2016). How many samples are needed ? An investigation of binary logistic regression for selective omission in a road network. *Cartography and Geographic Information Science*, *43*(5), 405–416.

Zwaan, G. L. Van Der, Hengel, K. M. O., Sewdas, R., Wind, A. De, & Steenbeek, R. (2019). The role of personal characteristics , work environment and context in working beyond retirement : a mixed-methods study. *International Archives of Occupational and Environmental Health*, *92*(4), 535–549.

LIST OF PUBLICATIONS

Indexed Journal

1. Mohamed, T. S. T. and Lee, M. H., (2020). Robust logistic regression for graduate employability from public universities in Malaysia. *(Accepted for publication in Matematika).*

Non-Indexed Conference Proceedings

1. Mohamed, T. S. T. and Lee, M. H., (2018). Logistic regression for employability rates modelling among graduates from public universities in Malaysia. 7th *International Graduate Conference on Engineering, Science and Humanities (IGCESH),* 646-648.