Graduates employment classification using data mining approach

Mohd Tajul Rizal Ab Aziz’ and Yuhanis Yusof’

Citation: **1761**, 020002 (2016); doi: 10.1063/1.4960842
View online: http://dx.doi.org/10.1063/1.4960842
View Table of Contents: http://aip.scitation.org/toc/apc/1761/1
Published by the American Institute of Physics

Graduates Employment Classification using Data Mining Approach

Mohd Tajul Rizal Ab Aziz[1, a)] and Yuhanis Yusof [2, b)]

[1]Kolej Professional Mara Indera Mahkota, Kuantan, Pahang, Malaysia
[2]School of Computing, Universiti Utara Malaysia, Kedah, Malaysia

[a)]Corresponding author: tajulrizal@mara.gov.my
[b)]yuhanis@uum.edu.my

Abstract. Data Mining is a platform to extract hidden knowledge in a collection of data. This study investigates the suitable classification model to classify graduates employment for one of the MARA Professional College (KPM) in Malaysia. The aim is to classify the graduates into either as employed, unemployed or further study. Five data mining algorithms offered in WEKA were used; Naïve Bayes, Logistic regression, Multilayer perceptron, k-nearest neighbor and Decision tree J48. Based on the obtained result, it is learned that the Logistic regression produces the highest classification accuracy which is at 92.5%. Such result was obtained while using 80% data for training and 20% for testing. The produced classification model will benefit the management of the college as it provides insight to the quality of graduates that they produce and how their curriculum can be improved to cater the needs from the industry.

INTRODUCTION

Graduates employment is one of the issues in Malaysia as there are as many as 53 higher education institutions that include public and private university. The large number of institution will then produce a large number of human resources; hence obtaining information on the whereabouts of the graduates can contribute to the strategic planning of the particular institution. Even though Malaysian education institutions are reported to produce more than 180,000 graduates each year, unfortunately there isn't any complete statistics on all of these graduates.

According to MOHE, employment can be defined as the potential to secure a job at workplace and employability can be defined as the potential to maintain, secure and grow in a particular job at workplace. Based on Robinson, in order to make sure that the graduates are employed, it depends on the employability skills such as model of understandings, achievements and personal attitude to get employment and successful in career. Another researcher, Bush and Barrick defined graduate employability as an attitude based on the personal value, decision making skills, problem solving, communication skills, relations with other people and commitment to get a job. Today, most companies hire employee not only based on the academic result and ability to write, listen and communicate, but the employers are also concern about the creative thinking, problem solving, decision making and reasoning of the potential employee.

One of the researches undertaken by Shafie and Nayan [23] revealed that the highest employability, given 100 graduates, is from Universiti Teknologi MARA (i.e. 77). This is followed by Universiti Sains Malaysia (USM) with 74, Universiti Islam Antarabangsa (UIA) with 71, Universiti Malaya (UM) with 63, Universiti Putra Malaysia (UPM) with 61, Universiti Kebangsaan Malaysia (UKM) with 38, Universiti Teknologi Malaysia (UTM) with 35 and Universiti Malaysia Sarawak (UNIMAS) with 34.

To date, various approaches have been used to study graduate employability and this includes data mining. Data mining (DM) is a part of Artificial Intelligence technology where it analyzes raw data and transforms it into knowledge. DM has been applied in various applications such as web mining [4], business intelligence, and decision support system. Data mining offers four main tasks; classification, clustering, regression and association mining.

Proceedings of the International Conference on Applied Science and Technology 2016 (ICAST'16)
AIP Conf. Proc. 1761, 020002-1–020002-7; doi: 10.1063/1.4960842
Published by AIP Publishing. 978-0-7354-1419-8/$30.00

These tasks can be achieved using various methods such as Decision Tree (DT), Linear Regression, Logistic Regression, Naïve Bayes, Neural Network (NN), k-nearest neighbor (K-NN), Relevance Vector Machine (RVM) and Support Vector Machine (SVM).

MARA Professional College (KPM) was formerly known as MARA Institute of Commerce (IPM), which was established on May 1977. KPM offers accredited diploma program from Malaysian Qualification Agency (MQA). To date, KPM have 6 branches which includes Seri Iskandar (KPMSI) Perak, Beranang (KPMB) Selangor, Bandar Melaka (KPMBM) and Ayer Molek (KPMAM) Melaka, Bandar Penawar (KPMBP) Johor and Indera Mahkota (KPMIM) Pahang. For this project, the study is focusing on KPMIM which offers four diploma programs such as Diploma in Accountancy (DIA), Diploma in English Communication (DEC), Diploma in Computer Networking (DCN) and Diploma in Business Digital Media Creative (DDC). At the moment, there are 1500 students and every year, a total of 300 graduates are produced by KPMIM. In practice, information on employment of KPMIM graduates is obtained directly from the graduates. It is the responsible of the alumni unit to acquire the employment status from every graduate, and to date, the process is done manually. Such a time consuming process is not efficient as the number of KPMIM graduates are increasing and there exist graduates who do not respond to the request from the alumni unit. As the graduate employment status is important in determining the quality of the program (i.e how relevant is the program to the market) offered by an education institution, there is a need to automatically capture the statistics.

In this project, data mining (i.e. classification task) is performed on KPMIM data in order to produce a classification model that best suits KPMIM. It is hoped that the obtained model can be used by the KPMIM management to forecast its graduate employment. Upon completing a program, KPMIM will know in advance whether a graduate will be employed in public or private sector, unemployed or continue study.

LITERATURE REVIEW

Graduates employment has been defined in various forms. In general, graduates employment means that a graduate is able to get a job once he has completed his training or education program. In higher education, it is important that graduates are able to find job that is relevant to their qualification. Today, the number of graduates is increasing vastly as compared to the demands from the market. One of the issues of why graduates fail to secure a job is personal attributes of the graduate which includes communication skills, confidence level, teamwork, professional and decision making skills. These skills are consider competent in application and practice. Other than that is understanding for scientific and technologies (Life Long Learning).

The Malaysian Government conducted a survey on Malaysian graduates and it was discovered that about 60,000 Malaysian graduates were unemployed due to a lack of experience, poor in English, poor communication skills and because they had pursued studies irrelevant to the market. The research further mentioned that the typical unemployed graduate was female, mainly from the Malay ethnic group and from the lower income group. Most unemployed graduates had majored in business studies or information technology. A total of 81 percent of the unemployed graduates had attended public universities where the medium of instruction in many courses was the Malay Language. The Ministry of Human Resource recently reported that a large number of graduates are still jobless. According to the report, 70 percent graduates of from public universities and institutions of higher learning are still unemployed. This is in contrast with 26 percent from private institutions of higher learning and 34 percent who are foreign graduates.

Peter Knight from the Institute for Educational Technology at the Open University is quoted in the Hobsons Directory 2005 (www.get.hobsons.co.uk), for graduate-level vacancies, discussing skills looked on favorably amongst employers: ‘‘When hiring, employers generally value good evidence of ability to cope with uncertainty, ability to work under pressure, action-planning skills, communication skills, IT skills, proficiency in networking and team working, readiness to explore and create opportunities, self-confidence, self-management skills and willingness to learn’’.

Based on article deaf college student, the Conference Board of Canada divided graduate employability into three layers of fundamental skills, personal management and teamwork skills. Reich (1991) made a list of necessary skills consists of abstraction, system thinking, experimentation and collaboration. A domestic scholar divided graduate employability into social compatibility, pre-professional image and characters for job. For deaf college student can conclude in positive perception IT will improved their initiative of study and conversation, social network, get the job information and negative is daily used are not suitable because get the harmful information.

According to David Rae [10], there are five approaches to develop the skills of employability with enterprise. First, is the personal development where student must understand the goal of learning, for example to develop a career plan. Other than that is applying the learning to learn activity theoretical and cognitive, and transfer skill between university and workplace. Another is the skill development that takes the opportunity to practice and get the credit for the development such as personal, people and task. In addition, the work based learning is an essential aspect of every degree providing opportunity for personal development, applied learning and skills development. Lastly is career management, it will be encouragement to participate in ongoing career development activity.

Besides of studies that determine attributes to contribute in determining employability, there also exist studies on the automatic classification of the graduates. In Jantawan and Tsai [14], they reported on graduate students of Maejo University in Thailand, where they determine if graduates are employed within 12 months after graduation. The research also identifies attributes for skilled graduates. The utilized DM methods includes Bayesian Network and decision tree. They concluded that WAODE algorithm in Bayesian Network is a better classifier because it obtained 99.77% accuracy as compared to J48 algorithm that produces 98.31% accuracy.

A study by Tair and El-halees [26] employs other classifiers such as neural network, and k-nearest neighbour to classify graduate employment and their work was based on the Khanyounis College of Science and Technology graduates. The study includes four stages; association, classification, clustering and outlier detection. In the association stage, they associate the student’s grades as either being ‘excellent’, ‘very good’, ‘good’ and ‘average’. In classification their used rule induction and naïve Bayesian. Based on analysis researcher suggest used the naïve Bayesian because this method can predict on time average student. In the clustering, they apply single value decomposition to cluster plot average from 0 to 3. Finally in the outlier detection stage, researchers found two main approaches, it is distance based approach with means student excellent result at the matriculation can be excellent result in the college. Researchers identify association rules and used the classification to predict the graduate employment.

Prior to that, in a study performed by Wook et al. [28], researchers try to combine Artificial Neural Network (ANN) and decision tree. The proposed a classifier was evaluated on students from the Computer Science Department, Faculty of Science and Defence Technology, Universiti Pertahanan Malaysia. There are six stages involved; project understanding, data collection, data preparation, modelling, evaluation and deployment. The research employs 85 students of semester 1 2008/2009 and focuses on two types of information (i.e personnel and academic) to predict the success of the student. The utilized ANN architecture operates based on sigmoid function and the decision tree was employed to represent and understand each cluster. Nevertheless, no data on classification accuracy was reported.

Based on the aforementioned studies, it is noted that data mining helps to automate the classification of graduates. The success of data mining can also be seen in other domains. For example, Zhang et al. [31] reported the use of two algorithms to analyze customer membership card. These algorithms includes the C5.0 clementine decision tree model and Classification and Regression Trees (CART). Based on these classification the highest accuracy algorithm model is C5.0 is 82.53% and CART is 82.24%.

In year 2010, a survey done by, Hyunchul et al. [2] research on the Mobile Telecom Market to develop customer classification model using data mining approach. The researcher divided two sections of DM techniques. Section 1 includes several DM classification such as Logistic Regression (LR), Decision Tree (DT) and Artificial Neural Network (ANN). Section 2 employs the Genetic Algorithm classification to provide probability. It is Simple Averaging of LR, DT and ANN (SAVG), Majority Voting of LR, DT and ANN (MVOTE) and Weighted Averaging of LR, DT and ANN (GAOW). Experiments results showed that DT is highest accuracy with 62.87%. In the work reported by So-In et al. [24] researchers use data mining to investigate intrusion detection. Based on this model he analyzed and find different result among classes with is normal versus attack, type of attack and individual attack. For this case, the best classification model was the k-nearest neighbor as requires the lowest computational complexity compared to other classification models.

Another area that benefits from data mining is the image processing. Wagle et al. [27] wanted to improve medical image classification model using DM techniques. Their used the k-nearest neighbor (KNN) and compare with other techniques such as Naïve Bayes Classifier (NB), Neural Network (NN) and Support Vector Machine (SVM). Based on the undertaken experiment, they concluded that KNN classifier is much easier to use and implemented on normal or severe category of medical images.

METHODOLOGY

The main objective of this project is to find the best classification model for KPMIM graduates employability. The aim is to use the model to forecast the employability of the future graduates. In order to obtain the model, four phases were included; data collection, data preprocessing, data mining and model evaluation, as shown in Figure 1.

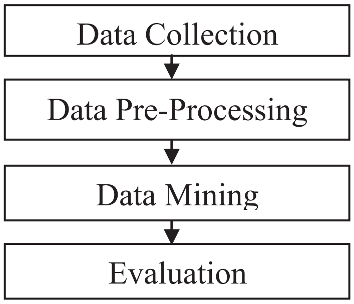

Figure 1. Research Phases

Data Collection

The data for this project is obtained from several sources; academic data from the examination unit, graduate information from alumni unit and also curriculum data from curriculum unit. Academic data is represented by two attributes; program, CGPA and number of semester. An example for the first attribute includes Diploma in English Communication while the values 3.88 and 2.53 are examples for the second attribute. On the other hand, the third attribute represents number of semesters a student has spent in the college in order to complete his/her program. Usually, the total of semester to complete the program is 6 semesters, however there exist students who took more than 6 semesters to complete his study due to specific reasons.

Based on CGPA data we can grouping using the grading system of the MQA to isolate index into several categories. Data obtained from the alumni unit includes gender and status of employment. The second attribute represents whether a graduate is employed or unemployed. If he is employed, the determination is whether he is in the private or public sector. There is also a possibility that the student is continuing his study. Lastly, the curriculum unit provides data on how active a student is in co-curriculum activities such as being a member or holding a post in societies, sports and clubs.

For active student their join the 2 co-curriculum activity and very active student their join the same co-curriculum but their make as a committee member. In total, this project utilizes 633 student records.

Data Pre-Processing

The database that has been used is from the KPM`s academic, alumni and curriculum unit. To build the data pre-processing for classification, firstly the database has to be transferred to excel format sheet and identify the attributes. Next is cleaning the data set with the missing value, removing the duplicate values, identifying the outlier and finally finding the consistent data. For example the duplicate attributes such as name, age, and matric no are removed from the dataset. Some of the collected data need to be discretized which means that the data is required to be represented in categories. This project discretizes the "CGPA" attribute into five categories. The first category is represented as 'excellent' that includes grade points of 4.00 to 3.67 while the points of 3.65 to 3.00 is included in the "Good" category. The "Satisfactory" and "Fail" bin includes the points of 2.99 to 2.33 and 2.32 to 2.00. The final category (i.e. "Fail") then includes the CGPA that is less than 2.00.

These data set must be converted from excel to comma separated values file (.csv) as it is compatible with the WEKA software in order to build the model. Table 1 shows the complete data set of graduate employment after the data pre-processing.

TABLE 1. Attributes for the classification of graduates employment

No	Attributes	Values	Description
1	Gender	Male, Female	Gender of graduate
2	Program	Diploma In English Communication, Diploma In Accountancy, Diploma In Computer Networking, HND Bit, Diploma In Business Digital Media Creative	Program enrolled in KPMIM
3	Semester	Number of Semesters	Number of semesters taken by a student to complete his program
4	Co-curriculum	"active" or "very active"	Student involvement in clubs and societies
5	Academic	Excellent, Good, Satisfactory, Pass and Fail	CGPA
6	Status Employed	Private, Public, Unemployed and Further Study	Status of graduate employment

Data Mining

Here, there are 5 classification models in WEKA were used which are; Naïve Bayes, Logistic Regression, Multilayer Perceptron, K-nearest Neighbor and Decision Tree J48. The results will be compared with each other respectively. The best algorithm will then be decided based on the comparison made. Firstly algorithm is the Naïve Bayes. It is essentially a classification from bayes theorem. It is a classification given based on assumption attribute which is based on individual class. Secondly, the Logistic Regression which is a statistical classifying method to analyze dataset that has one or more independent variables that later determine an outcome. The third algorithm is Multilayer Perceptron which is a type of Neural Network. The architecture includes one input, one hidden and one output layer. Next is the K-nearest Neighbor (KNN). This is a pattern recognition and classification model that works based on distance measure and in this project, the KNN employs the Euclidean Distance. Finally the Decision Tree (J48) produces a classification tree includes decision nodes, change nodes and end nodes.

Evaluation

In this phase, comparison of classification accuracy is performed on the different classification models. Based on the table 2 below, there are percentage of accuracy for classification using training, cross validation and testing set in WEKA software. The finding shows that the highest accuracy classification is Logistic Regression 92.47% using Testing Set=80%. Next, the accuracy percentage using Testing Set=80% is K-Nearest Neighbor with 70.09% and 69.93% for Training Set accuracy. Then, the accuracy percentage for Multilayer Perceptron using Training set is 68.98%. For the lowest percentage of accuracy classification is 62.40% of Naïve Bayes using Testing Set=70%.

TABLE 2. Classification Result

Classifier	Accuracy					
		Cross Validation		Testing		
	Training	5	10	70	80	90
Naïve Bayes	65.34	63.60	63.76	62.40	79.40	63.95
Logistic Regression	66.77	65.03	64.55	61.65	**92.47**	64.17
Multilayer Perceptron	68.98	63.44	62.65	63.90	68.91	69.23
K-Nearest Neighbor	69.93	63.92	63.76	64.66	70.09	69.45
J48	67.24	64.24	63.60	63.15	66.53	65.93

CONCLUSION

Every year, the number of graduates produced by higher education institutes is increasing. The optimal scenario is that the number matches job opportunities in the market. However, in reality, this is hard to achieve. Hence the goal of this study is to assist KPM management in estimating their graduate employability. This is achieved by creating a classification model that is able to automatically forecast whether a graduate will be employed in public sector, private sector, continue study or even unemployed. Based on the experiments, it is learned that the Logistic regression model is the best classifier for the KPMIM dataset. The model produces as high as 92.5% accuracy and has outperformed the other four classifiers.

As for future work, there is a need to include more attributes in the model, such as the grades for major courses in the program and the types of co-curriculum involvement (being a president, secretary or treasurer). Inclusion of these attributes would provide better understanding on the employment pattern of KPIM graduates. Furthermore, the new set of attributes need to be tested on other machine learning classifiers such as Support Vector Machine and Least Suqares Support Vector Machine.

ACKNOWLEDGMENTS

The author would like to thank the management of Kolej Professional MARA in particular from the Indera Mahkota branch in providing the data and financial support for the study.

REFERENCES

1. J.S. Aguilar-Ruiz, J.C. Riquelme, and M. Toro. Evolutionary learning of hierarchical decision rules. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society*, **33**(2), 324–31 (2003). doi:10.1109/TSMCB.2002.805696
2. A. Hyunchul, C.W. Song, J.J. Ahn, H.Y. Lee, T.Y. Kim and K.J. Oh. Using Hybrid Data Mining Techniques for Facilitating Cross-Selling of a Mobile Telecom Market to Develop Customer Classification Model. *The 43rd Hawaii International Conference on System Sciences (HICSS),* 1–10 (2010). doi:10.1109/HICSS.2010.429
3. A. Haug, J.S. Arlbjørn, A. Pedersen. A classification model of ERP system data quality, Industrial Management & Data Systems, **109** (8), 1053 – 1068 (2009)
4. M. André, and P. Henriques. A Web Mining, **6**(9), 1523–1532 (2010).
5. N.A. Aris, Z. Baharum, Z.M. Sanusi, and I.K.A. Rahman. Assessment of critical success factors for accounting graduates employability. *2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC)*, 526–531 (2013). doi:10.1109/BEIAC.2013.6560183
6. B. Menendez, S. Okazaki, A.D. Martin, and M. Rozano. Using twitter to engage with customer: A data mining approach. **29**(5), 494–519 (2009). doi:10.1108/JFM-03-2013-0017
7. K. Chen, Q. Huang, M. Hu, and R.M. Randoy. Perspective : Employment Conditions for Computer Engineering Related College Graduates in China, (ICCSE) (2011).
8. S. Chen, and Z. Wang. A Major Study on Quality Improvement of Employment of College Graduate. *Journal of Indian Society of Periodontology*, **18**(4), 425 (2014). doi:10.1109/EBISS.2009.5137894
9. S.I. Conference, M. Learning, C. Science, and E. Engineering. A New Classification Mining Model Based on the Data, 2–5 (Nov. 2003).
10. D. Rae. Connecting enterprise and graduate employability Challenges to the higher education culture and curriculum? Lincoln Business School, University of Lincoln, Lincoln, UK (2007).
11. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases, **17**(3), 37 (1996). doi:10.1609/aimag.v17i3.1230
12. M. Fenton, and A. Barry. Enhancing Young Graduate Intention Towards Entrepreneurship Development, **56**(8/9), 733–744 (2014).
13. L. Gao. Analysis of Employment Data Mining for University Student based on Weka Platform, **2**(4), 130–133 (2015).
14. B. Jantawan, and C. Tsai. The Application of Data Mining to Build Classification Model for Predicting Graduate Employment. *International Journal of Computer Science and Information Security*, **11**(10), 1–8 (2013).

15. Kementerian Pengajian Tinggi Malaysia. *The National Graduate Employability Blueprint 2012-2017 (2012)*. http://doi.org/10.1007/s13398-014-0173-7.2
16. S. Ken, T. Ting, and C.Y. Ying. Business Graduates ' Competencies in the Eyes of Employers : An Exploratory Study in Malaysia. *World Review of Business Research*, **2**(2), 176–190 (2012).
17. M. Alwazae, H. Kjellin, and E. Perjons. A synthesized classification system for best practices, VINE: The journal of information and knowledge management systems, **44**(2), 249 – 266 (2014)
18. M. Mukhtar, Y. Yahya, S. Abdullah, A.R. Hamdan, N. Jailani, and Z. Abdullah. Employability and service science: Facing the challenges via curriculum design and restructuring. *2009 International Conference on Electrical Engineering and Informatics*, 357–361 (Aug, 2009). doi:10.1109/ICEEI.2009.5254712
19. N. Hazilah, A. Manaf, K. Ahmad, S. Ahmed. Critical factors of service quality in a graduate school of Malaysia, International Journal of Quality and Service Sciences, **5**(4) 415 – 431 (2013).
20. T.O. Oresanya, O.S. Omodewu, T.T. Kolade, and A.O. Fashedemi. Vocational Education and Employability : The Nigerian Situation, **5**, 2013–2015 (2014).
21. K. Patel, and J. Donga. Practical Approaches : A Survey on Data Mining Practical Tools, **2**(9), 1–10 (2015).
22. X. Ren, and J. Malik. Learning a classification model for segmentation. *Proceedings of Ninth IEEE International Conference on Computer Vision*, **1**(c), 10–17 (2003). doi:10.1109/ICCV.2003.1238308
23. L.A. Shafie, and S. Nayan. Employability Awareness Among Malaysian Undergraduates. *International Journal of Business and Management*, **5**(8), 119–123 (2010).
24. C. So-In, N. Mongkonchai, P. Aimtongkham, K. Wijitsopon, and K. Rujirakul. An evaluation of data mining classification models for network intrusion detection. *2014 Fourth International Conference on Digital Information and Communication Technology and Its Applications (DICTAP)*, 90–94 (2014). doi:10.1109/DICTAP.2014.6821663
25. P. Spooren. On the credibility of the judge. *Studies in Educational Evaluation*, **36**(4), 121–131 (2010). doi:10.1016/j.stueduc.2011.02.001
26. M.M.A. Tair, and A.M. El-halees. Mining Educational Data to Improve Students ' Performance : A Case Study. *International Journal of Information and Communication Technology Research*, *2*(2), 140–146 (2012).
27. S. Wagle, J.A. Mangai, and V.S. Kumar. An improved medical image classification model using data mining techniques. *2013 7th IEEE GCC Conference and Exhibition (GCC)*, 114–118 (2013). doi:10.1109/IEEEGCC.2013.6705760
28. M. Wook, Y.H. Yahaya, N. Wahab, M.R.M. Isa, N.F. Awang, and H.Y. Seong. Predicting NDUM Student's Academic Performance Using Data Mining Techniques. *2009 Second International Conference on Computer and Electrical Engineering*, **2**, 357–361 (2009). doi:10.1109/ICCEE.2009.168
29. L. Ying. On the Construction of College Graduates' Employment Promotion System and the Realization of its Function (2011).
30. A. Zaharim, M.Z. Omar, Y.M. Yusoff, N. Muhamad, A. Mohamed, and R. Mustapha. Practical framework of employability skills for engineering graduate in Malaysia. *2010 IEEE Education Engineering Conference, EDUCON 2010*, 921–927 (2010). doi:10.1109/EDUCON.2010.5492478
31. L. Zhang, Y. Chen, Y. Liang, and N. Li. Application of Data Mining Classification Algorithms in Customer Membership\nCard Classification Model. *2008 International Conference on Information Management, Innovation Management and Industrial Engineering*, **1**, 211–215 (2008). doi:10.1109/ICIII.2008.168