Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students

Nor Azziaty Binti Abdul Rahman, Kian Lam Tan, Chen Kim Lim
Faculty of Art, Computing & Creative Industry, Sultan Idris Education University, 35900 Tanjong Malim, Perak Malaysia.
azziaty256@gmail.com

***Abstract*—Data mining techniques are widely used in engineering, medicine, industry, agriculture and even used in education to predict a future situation. In this paper, the used of data mining techniques applied in features selection and determine the best model that can be used to predict the employment status of fresh graduate Public Institutions either employed or unemployed, six months after graduation. In CRISP-DM methodology, six phases were adopted. The algorithm in supervised and unsupervised learning; K-Nearest Neighbor, Naive Bayes, Decision Tree, Neural Network, Logistic Regression and Support Vector Machines were compared using the training data set from Tracer Study to determine the highest accuracy in turn is used as a predictive model. Rapid Miner as a data mining tool was used for data analysis algorithm**

***Index Terms*— Data Mining; CRISP-DM; Rapid Miner Supervised and Unsupervised Machine Learning Algorithm; K-Nearest Neighbor; Naïve Bayes; Decision Tree; Neural Network; Logistic Regression and Support Vector Machine.**

I. INTRODUCTION

The unemployment rate among graduates has increased from year to year. A total of 200,000 graduates from university issued every year. 1 out of 4 graduates remain unemployed six months after graduation and majority being degree holders [12]. It is very hard to know the extent to which graduates themselves feel prepared for work, although this area is not researched. Therefore, there is need for a study to create on well awareness of the needs of the industry. In view of the worrying trends in graduate unemployment, it is timely that we begin to reflect upon, research, plan, survey, construct and administer some sort of employability assessment tools to help us to rectify the shortcomings or enhance the quality of students after graduation. There are many studies that have been carried out to detect the trend of output and employment status of graduates. One of these are the graduate tracer study by the ministry of higher education. This study was conducted to look at the employment status of new graduates are either working, studying or actively seeking work still ahead of the season convocation. The results have been recorded to be referenced by researchers in addressing the issue of unemployment among graduates. A total of 107,850 respondents among bachelor's degree holders participated in the online survey were developed by the ministry of higher education for the year 2015. The results showed a total of 58.0% of first degree graduates said they were employed, while 27.9% were still unemployed, 6.2 % waiting for work, 5.6% and 2.2%, further studies are still improving skills. Among the works were 88.7% worked full-time and 11.3% worked part-time according to Kementerian Pengajian Tinggi Malaysia where the survey did in 2015[5].

Based on jobstreet survey, about 64% employers said that they did not care whether graduates were from foreign, private or public universities. From the survey, about 88% of employers stated they are maintaining or increasing hiring in 2016. These percentages show that the election workers more to really fit and have a variety of skills and advantages, especially in terms of adaptability, multitasking skills, decision making and problem solving skills. According to SEEK Asia Chief Officer, Jake Andrew said that candidates should develop transferable skills and not just a specialist skill to enable them to do other job roles. Transferable skills are the skills acquire that were then applied to the working environment such as interpersonal skills, communications skills, organizational skills and leadership skills. Examples of interpersonal skills are like helping people, solving conflicts and motivation internal while examples for communications skills are like public speaking, advice and persuade or sell. Setting and achieving goals, time management and multitasking are the examples of organizational skills while the examples of leadership skills are like decision making, motivate others and work in groups.

Most organizations find that employees who have good qualifications, great personality won the favor of the employer. According to a survey recently conducted by JobStreet.com, only 14% of employers prefer qualifications while the majority of 51% of employers prefer personality as the main criteria for selecting employees [13]. There are six top reasons for fresh graduate unemployment of 59% of employers said poor attitude or character among graduates, 64% poor command in English, 60% poor in communication skills as well as problems in terms of multitasking skills, decision making skills, and problem solving skills. The study will link the attributes of the dataset with all six skills required to meet the needs of work by the employer. Thus, whether the methods of predictive analysis and data mining can help various parties in planning the graduate job opportunities after graduation. Data mining is an essential process of discovering pattern and knowledge from a large amount of data. Data mining is known for the usage in technology firm in the world such as Facebook, Microsoft, and Google etc. Data mining is about explaining the past and predicting the future trends and behavior by means of data analysis. Data mining has its roots in the fields of database technology and uses machine learning, statistics, visualization techniques and artificial intelligence which is easily comprehensible to human to present knowledge and discovery in a form. There are includes many tasks such as concept of description, association analysis, classification and prediction, cluster analysis, outlier analysis, statistical analysis, trend and evaluation analysis, regression and etc.

Predictive modelling is the process by which a model is created to predict the outcome. If the outcome is categorical then it is called classification but if the outcome is numerical then it calls regression. Under descriptive modelling, clustering is the assignment of observations into clusters so that observations in the same cluster are similar and association rules can find interesting associations amongst observation. Classification and prediction are the most important tasks in data mining. Selection algorithm often depends on the type of data (ie nominal, ordinal, ratio or interval) that will be used. Machine learning category is provided for every data mining algorithms and different data mining algorithms will be used to set data based on knowledge.

II. Preliminary Study

Classification and prediction techniques is one of the core data mining task that have been widely used by researchers in the area of prediction analysis. For studies on graduate employment, classification and prediction models commonly used such as Naive Bayes, Neural Network, Decision Tree, Logistic Regression, Random Forest and more. The prediction analysis also use WEKA as many data mining tools to build predictive models. The prediction analysis that have been done conducted using KDD (Knowledge Discovery in Databases) and CRISP-DM (Cross Industry Standard Process for Data Mining) as a method of study. Xu et al. [10] used Novel Neural Network (NN) to predict the unemployment rate based on information from the web. The study found that the effectiveness and efficiency of the method proposed as a potential alternative for forecasting the unemployment trend. By Mishra [7] who used five methods which are Bayesian, Multilayer perceptrons, Sequential Minimal Optimization (SMO), Ensemble Method and Decision Trees. The attributes such as academic achievement, emotional skills and socio-economic conditions were used to build predictive models in the data set. The study found that the decision tree algorithm is more suitable to predict the employability of students on the basis of acceptable accuracy. Sapaat et al. [8] have built a Graduate Employability model. This study used the data from Tracer Study conducted by the Ministry of Higher Education, Malaysia involving all graduates of polytechnics, public and private institutions. Sapaat et al. [8] used Bayesian and Decision Tree algorithm to determine whether in the first six months after graduation graduates have been employed, unemployed or in circumstances which cannot be determined.

The results of the study found that classification J48 a variant of Decision trees provide the highest accuracy it compared to Bayes algorithm.

Tajul et al. [9] used five data mining algorithms; Naive Bayes, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbor and J48 Decision Tree. The datasets collected from the Examinations Unit, Alumni unit and curriculum unit. The results showed that the Logistic Regression is the best classifier that can be used to predict whether graduates will work in the private or public sector, unemployed or continue their education. Shahiri et al. [11] used five classification tasks such as ANN, Naive Bayes, KNN, SVM and Decision Tree to predict student performance. Attributes are used for this study are demography, internal and external evaluation. The experimental results showed that the prediction error of Neural Network is less and has the highest classification accuracy than other tasks.

Aziz et al. [1] conducted a study on first year bachelor students in Computer Science course. Aziz et al. [1] proposed a framework for predicting the performance of students. To build predictive models, Aziz et al. [1] used pattern WEKA Naive Bayes classifier. There are six properties were tested, namely income families, university entry mode, race, gender, hometown, and CGPA. The study found that gender, hometown and family income contribute to academic achievement. Jantawan and Tsai [3] presented a model to predict employment status of graduates of Khon Kaen University. Jantawan and Tsai [3] used Bayesian methods included Naïve Bayesian Simple, Naïve Bayesian, Averaged One-Dependence Estimators (AODE), Averaged One-Dependence Estimators with subsumption resolution (AODEsr), Bayesian networks, and Naïve Bayesian Updateable. The results showed Averaged One-Dependence Estimators with subsumption resolution (AODEsr) algorithm has the highest accuracy of 98.3% followed by the AODE with 96.1% accuracy.

Gao [2] build a data mining model to analyze the work using WEKA. Gao [2] using the employment information as the attributes.Classic Desicion Tree classifiers were used for evaluation by doing the analysis and comparisons based on different criteria. There are three conclusions that can be generated which are almost half of graduates do not choose the educational institution, there is a relationship between the origin of students' place and every work is different for each gender. Jantawan & Tsai [4] conducted a research that built Graduates Employment model that can predict whether graduates are unemployed, work or in a state cannot be determined. Jantawan & Tsai [4] made a comparison between the Bayesian and Decision Tree. The study identified the attributes that can affect the employment of graduates 12 months after graduation. The study used a sample consisting of 11,853 samples collected from the Planning Office Maejo University in Thailand. Jantawan & Tsai [4] combined the Process Knowledge Discovery and Cross Industry Standard Process for Data Mining methodologies to build the classifier. The study found that the highest accuracy of 99.77% came from Waode of Bayesian methods. Masethe et al. [6] conducted a research to predict the Work Integrated Learning (WIL) placement based on the student's performance. J48, Bayes Net, Naive Bayes, Mobile Cart, and REPTREE algorithm were used in this research to classify students. Gender, attendance, sponsors, subjects and semester grades are used as the attributes to this research. The results showed that Bayes Net and Naive Bayes algorithm gives a good results and classification techniques successfully predict the number of students who passed or failed the examination for WIL placement. The overall, Decision Tree, Naive Bayes, Neural Networks, Logistic Regression, Bayes Network and Averaged One-Dependence Estimators with subsumption resolution (AODEsr) selected by researchers as a predictive model because of the highest accuracy based on experiments that have been carried out.

III. METHODOLOGY

The purpose of this research is to propose a suitable classification model that can be used in making prediction and assessment of the attributes of the student's dataset to meet the selection criteria of work demanded by the industry of the graduates in the academic field. In order to obtain the model, the CRISP-DM process model was used. There are six phases of CRISP-DM process model were included; business understanding, data understanding, data preparation, modelling, evaluation and deployment as shown in Figure 1.

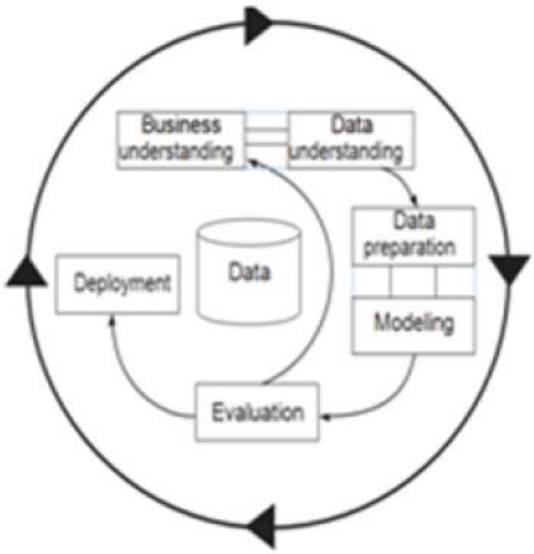

Figure 1 : Research Phases

Business Understanding: This phase focuses on understanding the objectives and requirements from a business perspective. This study aims to find a suitable model used to predict the attributes that affect the selected graduates either employed, further study, upgrading skills, waiting for work placement or unemployed. It uses the classification method in which the resulting model will be able to test certain attributes of the data set that can be used to predict whether graduates will be employed, further study, upgrading skills, waiting for work placement or unemployed. Data classification has two step processes; first step is the training data known as a learning step; a set of training databases were analyzed using a model that will describe the set of predefined concepts or classes. 60% of the data set were used as a training data. The second step is to testing data; the tested model was used to estimate classification accuracy using different data sets. For testing, 40% of the data were used. If the accuracy of the model is accepted, the model can be used to classify future data with unknown class label. Finally, the model can act as a classifier in the decision making process.

Furthermore, the RAPIDMINER [14] was used for data mining. Rapid Miner is a tool for experimenting with machine learning and data mining algorithms including data loading and transformation (ETL), data processing and visualization, modelling, evaluation and utilization. Rapid miner tool will be used to compare various data mining classification techniques. There are several techniques for classification such as K-Nearest Neighbor, Naive Bayes, Decision Tree, Neural Network, Logistic Regression and Support Vector Machine to obtain a suitable model for predictive analysis.

Data Understanding: In order to classify the attributes of student data set to determine the status of student either employed, further study, upgrading skills, waiting for work placement or unemployed using data sourced from the Tracer Study database from the Ministry of Higher Education (MOHE) for the year 2014 and 2015. The total are 16, 729 instances and 68 attributes related to graduate profile from five public universities; USM, UPM, UM, UKM and UTM were used as a samples of prediction. Table 1 below shows the complete attributes for the Tracer Study data set. Data Preparation: After the data is collected, the process provides data that has been achieved. In this phase, the data that will be used will be first cleaned and formatting. Data cleaning is data pre-processing to reduce noise and handle missing values. Cleaning data also involved correcting inconsistent data, data with identifying outliers as well as removing duplicate data. For example, some attributes like Job Sector and Industry Classification, have been entered in zero values because it is related for unemployed status. From the total of 68 attributes in the raw data, the data cleaning process ended up 41 attributes. This type of data has been modified and study. In order to construct the attributes, Rapid Miner tools were used. Rapid Miner only accepts data set in Binary File, Excel and Csr format. Prediction performance are checked by Rapid Miner. Rapid Miner is a data mining tool will then remove excess properties and public and normalize the data.

Modeling: Modeling phase can be done by testing and trying different techniques model with different properties. This test can be repeated with little change parameters, monitor results and make some preliminary conclusions on the model. A variety of tests, the results for six predictive model will be compared with existing techniques and seek the best outcome that meets the needs of problem analysis. K-Nearest Neighbor, Naïve Bayes, Decision Tree, Neural Network, Logistic Regression and Support Vector Machine were used as a machine learning for the section of modelling and experiments. K-Nearest Neighbor is an algorithm based on learning by analogy that compared the sample test provided with similar to it examples of exercises. A Naive Bayes is based on Bayes theorem that only requires a small amount of training data to estimate the variance and mean of variables. A Decision Tree is a tree like model or graph used to predict the target attribute value from example set by creating a classification model. Neural Networks are commonly used to find patterns in data or model complex relationships between inputs and outputs. Logistic regression was used to predict the outcome of one or more predictor variables and measuring the relationship between dependent variable and categorical variable. A Support Vector Machine takes a set of input data and predicts for each given input, which of the two possible classes comprises the input.

Table 1
Summary of Selected Museum MAR

Attributes	Values	Description
YOB	1950 - 1994	Year of Birth
Sex	Male, Female	Gender of graduate
Age	20-49 years old	Age
Degree	First Degree	Level of Study
e_Specialist	{Computer Science, Computer Application/ Software Development, Hardware & Networking, Computer Design & Graphic, System Analysis, Information Technology atc}	Name of degree
e_year e_month	2014-2015 1 - 10	Year and Month of Practical
e_Field	{0 - 8}	Field of Study
e_SubField	{10 – 863}	Sub Field of Study
Record_Nec	{0 – 8}	Education Programme
E_Nec	{10 – 863}	Specialization
e_cgpa	{2.00-2.49, 2.50-2.99, 3.00-3.66, 3.67-4.00, failed}	CGPA for current qualification
e_YOG e_MOG	2014 - 2015 1 - 10	Year and Month of Graduation
e_YOP e_MOP	2014 - 2015 1 - 10	Year and Month of Practical
e_campus	UM, UPM, USM, UTM, UKM	Name of University
{e_25_a_1_input, e_25_a_1, e_25_a_2_input, e_25_a_2, e_25_a_3_input, e_25_a_3, e_25_b, e_25_c, e_25_e, e_25_f, e_25_g, e_25_h, e_25_i, e_25_j, e_25_k} {e_27_a, e_27_c, e_27_d, e_27_e, e_27_f, e_27_g, e_27_h, e_27_i, e_27_j, e_27_k, e_27_l}	{1 – 5}	1 Very Not satisfied …5 Very Satisfied
e-status	{Employed, Further Study, Upgrading Skills, Waiting for Work Placement, Unemployed}	Current Status
e_43	{1- 5 }	1 Permanent 2 Contract 3 Temporary 4 Self-employed 5 Work with Family
e_45	{1 – 8}	1 Government 2 statutory body 3 private multinational 4 local private 5 Own business 7 government-linked companies 8 non-governmental organizations
e_46_kod	{1 -21}	Economic Sector
e_48_a, e_48_b, e_48_c, e_48_d, e_48_e, e_48_f_1, e_48_f_2, e_48__1, e_48_g_2, e_48_h, e_48_i	{1 -5, 9}	Capacity & Capability (1= Very Low …5= Very High 9 Irrelevant)
e_49_a, e_49_b, e_49_c, e_49_d, e_49_e		Satisfying Level (1= Very Low …5= Very High)
e_50_a	{1-5}	Opinions (1 Very Unhelpful .. 5 Very helpful)
e_54	{1,5, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17, 18}	Reason for not working

Evaluation: This phase involves evaluating the modeling results. Comparisons between different classification models such as K-Nearest Neighbor, Naive Bayes, Decision Tree, Neural Networ k, Logistic Regression and Support Vector Machine to do to get the classification accuracy. The accuracy of six predictive model will be tested where the model which provides the highest accuracy will be chosen to be a model of predictive analysis among the fresh graduate. Data mining tools provide options for the division into three distinct parts: a test, validation and training.

Deployment: The results found that a comparison of the six models will give different accuracy. The model provides the most accurate precision to be used to predict the attributes that really affects employment status of graduates. The model will be implemented a parallel data mining process for another data set.

IV. Implemantation

To obtain a suitable model to predict the employment among fresh graduates, six classification models have been proposed namely K-Nearest Neighbor, Naive Bayes, Decision Tree, Neural Network, Logistic Regression and Support Vector Machine. K-Nearest Neighbor is based on learning by analogy, comparing the test sample provided with examples of exercises that are similar to it. A Naive Bayes classifier is a classifier based on applying simple probability Bayes theorem and it only requires a small amount of training data to estimate the mean and variance of variables. A Decision Tree is a tree like graph or model to predict a target attribute value based on the number of features input from example set by creating a classification model. Neural network (NN), is a mathematical model or computational model that commonly used to model complex relationships between inputs and outputs or to find patterns in data. Logistic regression analysis was used to predict the ultimate outcome of one or more predictor variables and measuring the relationship between a categorical variable with the dependent variable to change the probability score. The standard SVM takes a set of input data and predicts, for each given input, which of the two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier

All six models will be tested, the model that has highest accuracy will be used to predict employment among fresh graduates.

The findings of previous studied found that Naive Bayes [1,3,6,9,11], Decision Tree [7,8,9,11,2,4,6], Neural Network [10], K-NN [10] and Logistic Regression [9] give a high accuracy while Support Vector Machine (SVM) suggested machine learning for improvement of existing models [9,10]. Overall, the selection of all six models because that models gave the good accuracy in prediction analysis conducted by most researcher.

V. Workability Tests And Discussions

Six classification techniques method have been applied; K-Nearest Neighbor, Naïve Bayes, Decision Tree, Neural Network, Logistic Regression and Support Vector Machine to build the classification model.

Tables below show the classification performance for each algorithm in Rapid Miner. The classification performance calculates one or more performance measures (performance criteria) and delivers them as performance vector. Performance vector can be used as estimation of a learner prediction accuracy or as fitness for an outer optimization scheme. In table 2, table 3 and table 4 show the result for classification performance for each classification model for the accuracy, classification error and root mean square error.

However, the classification performance of the Neural Network, Logistic Regression and Support Vector Machine can't be tested. This is because Neural Network, Logistic Regression and Support Vector Machine can't handle polynomial attributes. Logistic Regression and Support Vector Machine can be applied on data sets with numeric attributes.

Table 2
Classification Performance using K-Nearest Neighbor Classification Method

Accuracy:97.78%						
	true Upgrading Skills	true Un-employed	true employed	True waiting for job	True further study	Class precision
pred. Up-grading Skills	291	3	7	0	1	96.36
pred. Un-employed	2	2943	52	3	17	97.55 %
pred. employed	9	58	6163	14	26	98.29 %
pred. waiting for job	1	5	21	941	2	97.01 %
pred. Further Study	0	7	28	4	1112	96.61 %
class recall	96.04%	97.58 %	98.28 %	97.8 2%	96.0 3%	

Neural Network applied a sigmoid function as an activation function. The employment status of graduates was a major attributes labelled in the data set was a polynomial data type. Classification performance can only be tested using the Decision Tree, K-Nearest Neighbor and Naive Bayes. From the total 11, 710 testing instances and 28 attributes, the result for K-Nearest Neighbor achieved the highest accuracy with 97.78% compared to Naïve Bayes with 54.90% followed by Decision Tree 53.60%. Table 5 shows the classification error and root mean square error for Decision Tree, Naïve Bayes and K-NN. It noticed that Decision Tree achieved the highest classification error with 46.40% and K-Nearest Neighbor only 2.22% error compared to other algorithms.

Table 3
Classification Performance using Decision Tree Classification Method

Accuracy:53.60%						
	true Up-grading Skills	true Un-employed	true employed	True waiting for job	True further study	Class precision
pred. Up-grading Skills	3	0	1	0	0	75.00%
pred. Un-employed	0	5	1	0	0	83.33%
pred. employed	300	3011	6269	962	1158	53.58%
pred. waiting for job	0	0	0	0	0	0.00%
pred. Further Study	0	0	0	0	0	0.00%
class recall	0.99%	0.17%	99.97%	0.00%	0.00%	

Table 4
Classification Performance using Naïve Bayes Classification Method

Accuracy:54.90%						
	true Up-grading Skills	true Un-employed	true employed	True waiting for job	True further study	Class precision
pred. Up-grading Skills	74	21	83	3	5	39.78%
pred. Un-employed	93	1467	1290	145	387	48.38%
pred. employed	117	1232	4174	299	489	66.14%
pred. waiting for job	6	147	457	496	59	42.56%
pred. Further Study	13	149	267	19	218	32.73%
class recall	24.42%	48.64%	66.56%	51.56%	18.83%	

Table 5
Classification error and root mean square error

Algorithm	Classification Error	Root Mean Square Error
Decision Tree	46.40%	0.651 +/- 0.000
Naïve Bayes	45.10%	0.602 +/- 0.000
K-NN	2.22%	0.149 +/- 0.000

VI. Conclusion And Future Work

Unemployment among graduates is a major issue in this country. This issue should be addressed first, particularly at the university level. In order to meet the needs of industry and the goals of higher education, various methods; K-Nearest Neighbor, Naïve Bayes, Decision Tree, Neural Network, Logistic Regression and Support Vector Machine have been designed and implemented to identify the attributes that affecting the graduates whether employed, further study, upgrading skills, waiting for work placement or unemployed . This study determine which method is actually right in assessing the nature really affect the selection of graduates from UM, UKM, USM, UPM and UTM in the job industry. Results show that K-Nearest Neighbor achieved the highest accuracy of 97.78% with 2.22% classification error and 0.149 +/- root mean square error. K-Nearest Neighbor will be used as a model of prediction to find the attributes from the data set that affect the employments status among the fresh graduate's students.

As a future work, there is a need of more attributes likes grade for every subjects taken during the study period, co-curriculum achievement and the results of the oral test also integrating data sets from different sources of data such as UPSI and other universities.

Acknowledgements

This research is fully supported by number 2015-0184-109-72

References

[1] Aziz, A. A., Ismail, N. H., Ahmad, F., & Hassan, H. (2015). A framework for students??? academic performance analysis using na??ve bayes classifier. *Jurnal Teknologi*, *75*(3), 13–19. https://doi.org/10.11113/jt.v75.5037.

[2] Gao, L. (2015). Analysis of Employment Data Mining for University Student based on Weka Platform, *2*(4), 130–133 A Comparison. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, *3297*(6), 4584–4588. Retrieved from www.ijircce.com

[3] Jantawan, B., & Tsai, C. (2007). A Classification Model on Graduate Employability Using Bayesian Approaches:

[4] Jantawan, B., & Tsai, C. (2013). 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. https://doi.org/10.1016/j.bdr.2015.01.001

[5] Kementerian Pengajian Tinggi Malaysia. (2015). Status Pekerjaan Graduan (Warganegara) 2015 (p. 46).

[6] Masethe, M. A., & Masethe, H. D. (2014). Prediction of Work Integrated Learning Placement Using Data Mining Algorithms, *I*, 22–24.

[7] Mishra, T. (2016). Students ' Employability Prediction Model through Data Mining, *11*(4), 2275–2282.

[8] Sapaat, M. A., Mustapha, A., Ahmad, J., & Chamili, K. (2011). A Data Mining Approach to Construct Graduates Employability Model in Malaysia, *1*(4), 1086–1098.

[9] Tajul, M., Ab, R., & Yusof, Y. (2016). Graduates Employment Classification using Data Mining Approach, *20002*. https://doi.org/10.1063/1.4960842

[10] Xu, W., Li, Z., Cheng, C., & Zheng, T. (2012). Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications*,*7*(1), 33–42. https://doi.org/10.1007/s11761-012-0122-2

[11] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

[12] Michelle, L. (2016). Fresh Graduate Unemployment in Malaysia | EduAdvisor. Retrieved from https://eduadvisor.my/articles/what-didnt-know-fresh-graduate-unemployment-malaysia-infographic/

[13] Employers Pick Personality over Qualification _ JobStreet. (n.d.). Retrieved from http://www.jobstreet.com.my/career-esources/employers-pick-personality-qualification/#.WKMqX9J97IU

[14[http://rapid-i.com/

[15] International, T., & Of, J. (2016). Proposed Model for Predictive Mapping of Graduate ' s Skills to Industry Roles Using Machine Learnining Techniques Fullgence Mwachoo Mwakondo , 2 Dr . Lawrence Muchemi , 3 Prof . Elijah Isanda, 15–24.

[16] Ramamohan, Y., Vasantharao, K., Chakravarti, C. K., & Ratnam, a S. K. (2012). A Study of Data Mining Tools in Knowledge Discovery Process. *International Journal of Soft Computing and Engineering*, *2*(3), 191–194. Retrieved from http://www.ijraset.com/fileserve.php?FID=3979