A Comparative of Predictive Model of Employability

KianLam Tan, Nor Azziaty Abdul Rahman, ChenKim Lim

Abstract: ***In 2017, the global unemployment rate is projected around 5.6% while for 2018 the unemployment rate is 5.5% which is little bit decrease. However, the youth (aged 15 to 24) unemployment rate in Malaysia is over three times higher at around 10.8% in 2017. In addition, Malaysia achieved the second highest rate after Indonesia (15.6%) compare to other countries in Asian including China (10.8%), India (10.5%), Singapore (4.6%), Vietnam (7%), Thailand (5.9%) and Philippines (7.9%). This study aim to present a set of data mining algorithms to find the most important factor of employability among the fresh graduate students. The comparison for six data mining algorithms which are 1) Logistic Regression, 2) Decision Tree, 3) Naive Bayes, 4) K-Nearest Neighbor, 5) Support Vector Machine and 6) Neural Network by using split validation method which is 70-30 as a ratio. Based on the result, Neural Network is the best classifier other than another five algorithms. The Neural Network Model showed 6 majors effect on employability are 1) willing to face challenges of the outside world and work, 2) can communicate effectively, 3) field of technical, 4) convocation on October and 6) Sex (Male). The predictive model of employability will benefit the management of the higher education, Ministry of Education and fresh graduate itself to predict the employability status either employed and unemployed by graduate data.***

I. INTRODUCTION

Employment problem can produce a negative impact on their own community, government and individuals itself especially among the fresh graduate students. In general, unemployment rate in Malaysia remained unchanged at 3.30% in February and January of 2019 which is slightly better than July 2018 as shown in Figure 1. Higher order thinking skills, basic academic skills and personal qualities are the three major requirements by the employers to hire the employees [1].

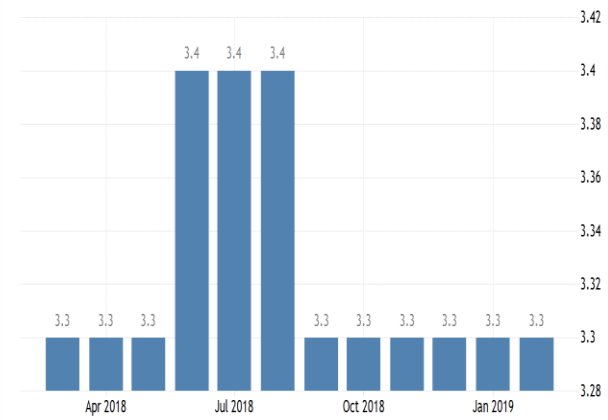

Fig. 1 Overview of Malaysia Unemployment Rate from April 2018 to Jan 2019 from Department of Statistics, Malaysia

Realizing this problem, it is important to identify a good predictive model to determine what sort of intervention is suitable for a particular graduate students so that accurate prediction can be implemented. This study used a technique for knowledge discovery where the computer obtain and learn insight the data. Artificial Intelligence (AI) is the software that drives the Fourth Industrial Revolution. The impact of AI can be seen in homes and businesses. As the rapid progress in Machine Learning (ML), it increase the scale and scope of AI's deployment across all aspects of daily life and as the technology itself that is able to learn and change on its own. Nowadays, the complexity of scientific discovery in the ever-increasing amount of data that can be easily handled using machine learning. Most programmers, statisticians, mathematicians, and engineers are the ones who is familiar with these techniques and algorithms. The research questions is the guidance for this research study .This paper discusses how data is analyzed and results obtained. The data that has been analyzed is used to explore, identify and explain the relationship of the attributes. Data were obtained from Tracer Study, Ministry of Higher Education that involved five public universities. This paper covers data analysis technique and result for the predictive model of employability among the fresh graduate students. The analysis was carried out using Rapid Miner software product.

Predictive Analysis used to analyze historical and current facts to predict unknown or future events.

Revised Manuscript Received on May 22, 2019.

KianLam Tan, Faculty of Art, Computing & Creative Industry, Sultan Idris Education University, Perak, Malaysia

Nor Azziaty Abdul Rahman, Faculty of Art, Computing & Creative Industry, Sultan Idris Education University, Perak, Malaysia

ChenKim Lim, Faculty of Art, Computing & Creative Industry, Sultan Idris Education University, Perak, Malaysia

It is includes a statistical techniques likes machine learning, data mining and predictive modelling. Functional effects that define the technical approach are predictive analysis providing a predictive score (probability) for each graduate student to determine, inform, or influence processes related to the factors affecting graduate students whether they can employed or unemployed.

II. LITERATURE REVIEW

Predictive studies in educational institutions also get researchers' attention in assessing the parameters that influenced the academic quality of the faculty in the career of students [1]. This study used data mining methods and used data from a college in India. The data used consists of thirty one attribute of eight groups such as Quality of Teaching, Faculty profile, Learning Assessment, Maintaining Relationships, Organizational Qualities and Outcome Counselling and Mentoring, Administrative Functions, Research and Development. By using universal kernel function (linear, radial, sigmoid, polynomial and Pearson based on kernel functions (PUKF) in Support Vector Machines (SVMs) for model learning. A Support Vector Machine analyze data for regression and classification analysis associated learning algorithms. [2] predicted the trend of unemployment based on web information by using Novel Neural Network (NN). Academic achievement, emotional skills and socio-economic conditions are the attributes tested using Sequential Minimal Optimization (SMO), Bayesian, Multilayer perceptrons, Decision Trees and Ensemble Method by Mishra [3]. Naive Bayes, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbor and J48 Decision Tree used by Tajul et al. [4] to test the datasets collected from the curriculum unit, examinations unit and Alumni unit. Gao [5] used employment information and Decision Tree classifier by using WEKA to build a data mining model. Jantawan & Tsai [6] conducted a research by build graduates employment model to predict whether graduates are employed, unemployed or undetermined situation through comparison between Decision Tree and Bayesian. Affendey et al. [7] used Naïve Bayesian algorithms, AODE and compared the results of the experiment using Bayesian to predict student academic performance factors. [8] used Decision Tree classifications techniques to identify the factors of employability through performance comparison for the models under Bayesian methods.

III. ALGORITHMS

This study is a quantitative experimental that used supervised machine learning algorithms to investigate the existing factors that significantly causes fresh graduate students to be employed or unemployed as well as identify the relationship within the attributes of employability among fresh graduate students. This study used Rapid Miner Studio Software to build predictive flow analysis and analyze the entire data obtained. The designed for the model to use RapidMiner tool, which is predictive analytical tool and data mining business. This study used six algorithms which are 1) Logistic Regression, 2) Decision Tree, 3) Naïve Bayes, 4) k-Nearest Neighbor, 5) Support Vector Machine and 6) Neural Network.

1. **k-Nearest Neighbor Algorithm**: K-Nearest Neighbor is an algorithm that compares an unknown examples to a value of k which is the nearest neighbor with an unknown examples. First step is to find k closet training examples. Then the second step, the unknown examples is classified by a majority vote of the found neighbours.
2. **Naïve Bayes Algorithm:** Naïve Bayes is a low-variance classifier and high-bias that even can build a good model with a small data set. The assumption of Naïve Bayes is that the label value (class) and any attribute value is independent compared to other attribute.
3. **Decision Tree Algorithm:** A Decision Tree is like collection of nodes intended to estimate a value of numerical target and create decision on affiliation value to a class. Each node represents a splitting rule for one specific attribute. A particular property is represented by each node for the separation rule. New nodes are developed repeatedly so the criteria are met. Estimates for numerical values are obtained based on the target on the leaf values if the predictions for the class lab are determined by the majority of the sample that attained this leaf during generation.
4. **Neural Network Algorithm:** Neural Network learns a model by multi-layer perceptron which is a feed-forward trained by a back propagation algorithm. This algorithm use connectionist approach to compute the information processes and consists of interconnection of artificial neurons groups.
5. **Logistic Regression Algorithm:** Logistic regression is used for predicting the categorical outcome based on one or more predictor variables. The probability of a single probable result is modelled using the logistic function as an explanatory variable function. Logistic regression measures the relationship by converting the dependent variable to probability scores between a categorical dependent variable (continuous independent variable).
6. **Support Vector Machine (SVM) Algorithm:** Support Vector Machine predicts each input with two possible classes making the Support Vector Machine as a non-probabilistic binary linear classifier.

IV. EXPERIMENT SETUP

The data consists of 16729 examples and 31 attributes for 2015. The Power BI software is used to visualize the data distribution as shown in Figure 2.

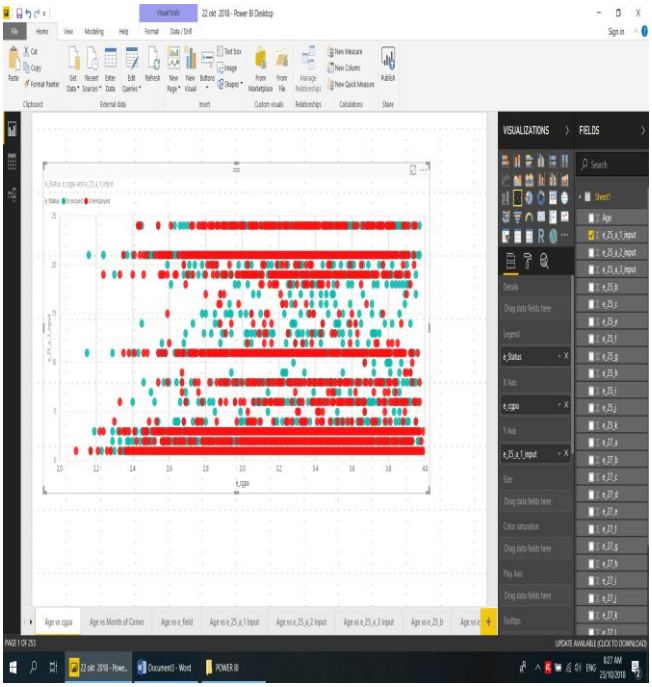

Fig. 2 Overview of Visualizing the Data Distribution between CGPA (x-axis) with Age (y-axis)

The steps of pre-processing as shown in Figure 3 consists of seven steps which are 1) the process of replacing missing values, 2) filtering (age between 21-25 years old, CGPA greater than 0), 3) discretizing by binning four values from the attribute of CGPA (excellent, good, satisfactory and pass), 4) generate attribute, 5) remove the duplication data, 6) set role (employment status as a label), 7) normalize and 8) outlier detection.

The accuracy of the model was tested, the model with the highest accuracy was chosen as a predictive model of employability. Four types of performance measures have been used in this research which are 1) Accuracy (ACC), 2) Recall, 3) Specificity, and 4) Precision.

The accuracy (ACC) is calculated by the number of all correct predictions then divide by the total number of the dataset as shown below:

ACC = (True Employed + True Unemployed) / Total Number of Employed and Unemployed (1)

Recall is the number of correct positive predictions divide by the total number of positives value as shown below:

Recall = True Employed / Total number of Employed (2)

Specificity the number of correct negative predictions divide by the total number of negatives as shown below:

Specificity = True Unemployed / Total number of Unemployed (3)

Precision is the number of correct positive predictions divide by the total number of positive predictions as shown below:

Precision = True Employed / Total number of Employed predictions (4)

In addition, the steps of pre-processing as shown in Figure 2 consists of seven steps which are 1) the process of replacing missing values, 2) filtering (age between 21-25 years old, CGPA greater than 0), 3) discretizing by binning four values from the attribute of CGPA (excellent, good, satisfactory and pass), 4) generate attribute, 5) remove the duplication data, 6) set role (employment status as a label), 7) normalize and 8) outlier detection.

It is crucially important to test a model not only on training data but on new data that has not yet been used for training. With a number of dataset and to avoid over fitting, split validation methods were used with 70/30 percentage split. The ratio for splitting the data is 70% of the complete data for training which is build a model and to the remaining 30% for testing the model which is applies to evaluate the performance. The percentage of accuracy is calculated using a split validation. Split validation has a training and a testing sub process. The training sub process is used for learning and build a training model then the training model is applied in the testing sub process. The model performance is measured during the testing phase.

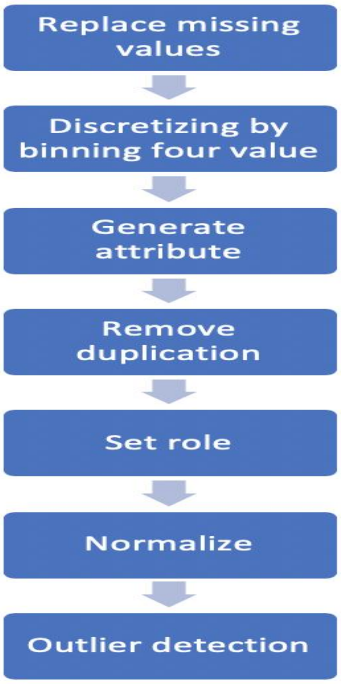

Fig. 3 Steps for Pre-processing

V. RESULT AND DISCUSSION

The accuracy results with different types of algorithms is shown in the Table 1 below which are Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbor, Support Vector Machine and Neural Network. The most improvement is with Neural Network from 52.25% to 59.07% (Gain 13.05%) while the least improvement is with Logistic Regression from 58.44% to 58.50% (Gain 0.10%).

In addition, Neural Network provides the best accuracy among the six algorithms.

Table 2 shows the confusion matrix for Neural Network since Neural Network provides the best result in Accuracy. Basically, a confusion matrix is formed by four outcomes that produced a result of binary classification. A classifier predicts all test data either Employed or Unemployed which produces four outcomes (True Employed, True Unemployed, False Employed and False Unemployed). Based on Table 2, the value for Specificity is 53.73%, Recall is 63.89% and Precision is 57.32% for Neural Network.

Table. 1 A Comparative of Six Data Mining Techniques for Classification Accuracy

Algorithms	Accuracy (Default Parameter)	Accuracy (Tuning Parameter)	Accuracy Gain
Logistic Regression	58.44%	58.50%	0.10%
Decision Tree	58.55%	58.84%	0.50%
Naïve Bayes	56.94%	57.66%	1.26%
k-Nearest Neighbor	52.45%	57.40%	9.43%
Neural Network	52.25%	59.07%	13.05%
Support Vector Machine (SVM)	58.93%	59.01%	0.14%

Table. 2 Confusion Matrix for Neural Network

Neural Network	True Employed	True Unemployed	Class Precision
Predict Employed	885	639	57.32%
Predict Unemployed	762	1166	60.48%
Class Recall	53.73%	63.89%	

VI. CONCLUSION AND FUTURE WORK

In this study, six data mining techniques were compared on Tracer Study, Ministry of Higher Education dataset with little change parameter. Graduate dataset have 16729 examples, 31 attributes with two class label, employed and unemployed. Neural Network achieved the highest accuracy of 59.07% followed by Support Vector Machine 59.01% and Decision Tree 58.84%. From the experimental result, Neural Network can be used as a predictive model for predicting employability among the fresh graduate students. The accuracy (59.07%) was a result for 70% training and 30% testing data. The performance is measured by four evaluation measures in terms of its 1) Accuracy (ACC), 2) Recall, 3) Specificity, and 4) Precision. In addition, the experiment showed that 6 majors effect on employability are 1) willing to face challenges of the outside world and work, 2) can communicate effectively, 3) field of technical, 4) convocation on October and 6) Sex (Male). As for future work, in order to improve the efficiency of Neural Network algorithm, other statistical indicators and statistical techniques based feature selection can be integrated.

ACKNOWLEDGEMENTS

This research is fully supported by University Research Grant from Sultan Idris Education University under the grant number of 2018-0134-109-01.

REFERENCES

1. Deepak, E., Pooja, G. S., Jyothi R. N. S., Venkatrama, P. K. S.: SVM Kernel based Predictive Analystics on Faculty Performance Evalation, 1-4 (2016)
2. Shafie, L.A, Nayan, S.: Employability Awareness among Malaysian Undergraduates. International Journal of Business and Management, 5(8):119—123 (2010)
3. 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
4. Mishra, T. (2016). Students ' Employability Prediction Model through Data Mining, *11*(4), 2275–2282.
5. Tajul, M., Ab, R., & Yusof, Y. (2016). Graduates Employment Classification using Data Mining Approach, *20002*. https://doi.org/10.1063/1.4960842
6. Gao, L. (2015). Analysis of Employment Data Mining for University Student based on Weka Platform, *2*(4), 130–133.
7. 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
8. Affendey, L. S., Paris, I. H. M., Mustapha, N., Sulaiman, M. N., and Muda, Z, "Ranking of influencing factors in predicting student academic performance", Information Technology Journal, Vol. 9, No. 4, pp. 832-837, 2010.
9. Kumar, V. and Chadha, A., "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, Vol. 2, pp. 80-84, 2011.
10. Arsad, P. M., Buniyamin, N., & Manan, J. A. (2014). Neural Network and Linear Regression Methods for Prediction of Students ' Academic Achievement, (April), 916–921.
11. Huang, J. (2014). Hardiness , Perceived Employability , and Career Decision Self-Efficacy Among Taiwanese College Students, (415), 1–14. https://doi.org/10.1177/0894845314562960