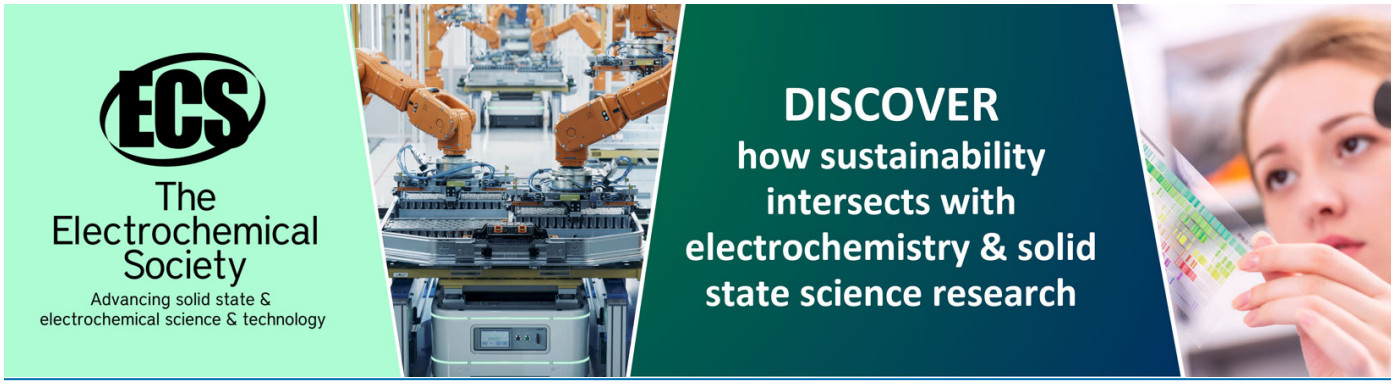
A comparative Study on Graduates' Employment in Malaysia by using Data Mining



A comparative Study on Graduates' Employment in Malaysia by using Data Mining

Nur Iman Natasha Binti A'rifian[2,a], Nur Sakinah Amirah Binti Mohd Daud[2,b], Athirah Faiz Binti Muhamad Romzi[2,c] & Nur Huda Nabihan Binti Md Shahri [1,2,d]

[1]Faculty of Computer and Mathematical Sciences
[2]Department of Mathematical Sciences and Decision Science, Universiti Teknologi MARA

E-mail:[d]huda_nabihan@tmsk.uitm.edu.my

Abstract. This study implements data mining to extract knowledge by analysing the graduates' employment dataset from year 2017 obtained from Ministry of Higher Education (MoHE). The objective of this study is to compare three predictive models which are Decision Tree (DT), Logistic Regression (LR) and Artificial Neural Network (ANN). Besides, this study is also done to determine the best predictive model for predicting graduates' employment sectors whether in public sector or private sectors. Every graduate student wishes to choose the right path in determining which sectors they are going to be entered, either to the public or private sectors. Usually, most graduates in Malaysia prefer the employment in the public sector rather than the private sector. Using data mining to discover the relationship and patterns can help in making a better decision. Prediction model is a must to determine the best performance when dealing with the large data set which helps the graduates to choose a sector based on the type of data or information that he/she furnishes. Based on the analysis, Artificial Neural Network (ANN 5) is the best model in predicting placement of employed graduates whether in public sector or private sector compared to the other models. ANN 5 is the highest accuracy at 81.52% and sensitivity at 65.67% while for the specificity of ANN 5 is 91.44%. The misclassification rate of ANN 5 is 18.48% which is the lowest compared to the other models. Overall, ANN 5 is the best model to predict negative target which is graduates employed in private sector since the value of specificity is higher than sensitivity. The result of this study can be used by government, universities and other responsible agencies in order to predict whether graduates will be employed in public or private sectors.

1. Introduction

In different national context, there is an interest in understanding the career choice that has been developed among students in Malaysia. There are 53 higher education institutions included public and private universities [1]. This is proven that each of the universities will produce a large number of fresh graduates that will be enter the crucial phase in life which is seeking for jobs. Choosing between the public and private sectors may be a problem to the graduates.

By using data mining to discover the relationship and patterns can help in making a better decision. Besides, it is a process of finding predictive information in a large database. Hence, the prediction model is a must to determine the best performance when dealing with the large data set which helps the graduates to choose a sector based on the information obtained. Model predictions are then compared

based on certain criteria such as accuracy, sensitivity, and specificity with the aim of appropriately using the right prediction technique in making prediction.

Thus, this study focuses on Data Mining approaches to analyse graduates' employment dataset using several predictive models including Decision Tree (DT), Logistic Regression (LR) and Artificial Neural Network (ANN)

2. Literature Review

In this globalization era, it is necessary for graduates to have a job that is relevant to their qualifications. The Ministry of Higher Education (MoHE) will conduct a survey every year for the graduates to answer in order to find out the graduates' employment rate. There are a few numbers of previous research that used data mining technique as a tool to analyse and explain the condition of graduates by the dataset obtained from the survey.

Yuhanis Yusof reported that in the study, they tend to investigate the suitable classification model to classify the graduates' employment for one of the MARA Professional College Indera Mahkota (KPMIM) in Malaysia [13] whether they were employed, unemployed or further their study into higher education. The classification models that included in the study was Naïve Bayes, Logistics Regression, Multilayer Perception, K-Nearest Neighbour and Decision Tree J48. From the result, it indicated that Logistic Regression model is the most suitable classifier for KPMIM dataset. This was explained by the author in finding and discussion where the LR had the highest accuracy which was at 92.5%[2].

Next, a study performed by Aida Mustapha reported that the main objective of this paper was to predict a graduate employability status, whether they were employed, unemployed or under undetermined situation within the first month from the graduation date by applying Bayes algorithms and tree-based algorithms [14]. This study also was to construct the graduates' employability model using classification in data mining. Based on the result explained by the author, from the comparison between Bayes algorithms and tree-based algorithms, it showed that tree-based algorithms had the highest and second highest accuracy which was 92.3% and 92.2%. While for Bayes algorithms, the accuracy was 91.3% which made the model fell to the third highest. Nevertheless, the study found that Bayes algorithms and tree-based algorithms complement each other because the Bayes method provided a better view of association or dependencies among the attributes while the result from tree-based method are easier to interpret and easier to understand [1].

Besides, a study by Peter Haddawy [4] reported that the study compared the accuracy of decision tree and Bayesian network algorithms in order to predict the academic performance of undergraduate and postgraduate students at two different academic institutes which were Can Tho University (CTU) and Asian Institute of Technology (AIT). The result from the study showed that the decision tree algorithms were more accurate than the Bayesian network algorithms in predicting students' performance. This was proven from the table that decision tree was 3 to 12% more accurate than the Bayesian network.

A study by Aida Mustapha proposed a clustering analysis using k-means and expectation maximization algorithms on various set of skills specific to different job sectors from government agencies to multinational agencies [14]. The objective of this study was to evaluate relevant job sectors in Malaysian market in order to find out the employability scenario for local undergraduate students. The dataset was obtained by a tracer study done by Ministry of Higher Education (MoHE) that consists of employability information for Diploma and Degree graduates from public universities in year 2011. The purpose of the Ministry of Higher Education (MoHE) done the survey was to improve the quality rate of students in higher education using more strategic manner. According to this study, the analysis will be on clusters emerged from the criteria of students' knowledge and skills that were acquired and their job sectors from the tracer study dataset. Based on the result discussed in this study, they found out that the graduates have only average skills at interpersonal communication, creative and critical thinking, problem solving, analytical and team work. The result provided by the clustering analysis were useful to university management and policy makers so that they can focus on shaping future Malaysian graduates. Besides, it was used to improve the higher educational system in terms of curriculum enhancement as well as teaching and learning structure.[5].

3. Methodology

This study only focuses in nine variables consist of age, gender, education institution, level of study, field of study, employed status, income, employed sector and economy sector. The data used is secondary data and the type of research is content analysis and obtained from Ministry of Higher Education (MoHE) which consists of 119,513 graduates in 2017. Thus, this data strictly focusses on graduates who employed under public sector and private sector in 2017.

The method of data analysis that will be discussed is data mining models DT, LR and ANN. The result will be compared which models in data mining is the best to predict graduate's employment whether in public sector or private sector. In order to obtain the model, four phases were included. The iterative and sequence of phases are shown in Figure 3.1. The details of this study are provided in Section 3.1.

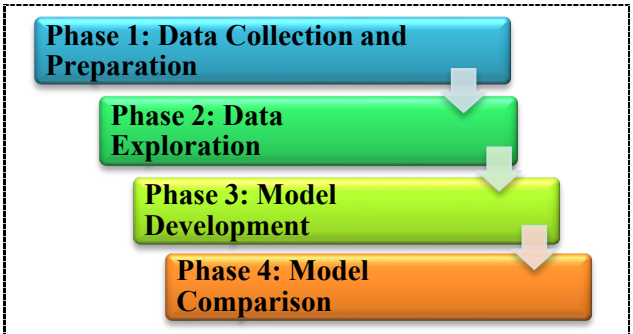


Figure 3.1 Research framework

3.1 Method

3.1.1 Decision Tree (DT)

DT is a flowchart-like tree structure, where each internal node denotes a test on a variable, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [8]. During the tree construction, variables selection measures are used to select the variables which have best partitions. The result can be explained using several outputs which are, tree output, table of variable importance and table of fit statistics. Figure 3.2 below shows a simple decision tree diagram.

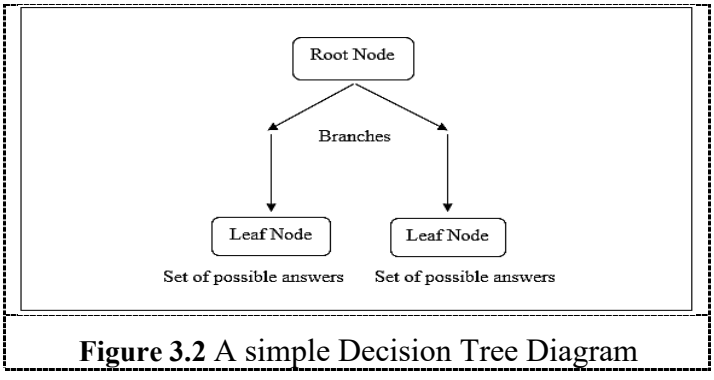


Figure 3.2 A simple Decision Tree Diagram

3.1.2 Logistic Regression (LR)

LR is a statistical classifying method to analyse dataset that has one or more independent variables that are later to determine an outcome (dichotomous or binary) [9]. Both continuous and numerical input variables can also be used. Several explanations in this logistic regression analysis which are table of likelihood ratio test, table of fit statistics of the model, table of Wald Chi-Square and table of odd ratio estimates. The logistic regression model is shown below.

$$\ln\left(\frac{p}{p-1}\right) = \beta_0 + \beta_1 + \ldots + \beta_n \quad (1)$$

where,

- p = probability that the event Y occurs p(Y=1)
- [range = 0 to 1]
- $\left(\frac{p}{p-1}\right)$ = the "odds ratio"
- $\ln\left(\frac{p}{p-1}\right)$ = log odds ratio or "logit"
- [range = -∞ to +∞]
- β_0 = the value of constant variable
- β_1 = the value of parameter

3.1.3 Artificial Neural Network (ANN)

ANN is computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes based on input and output [10]. ANN has 3 layers which are input layer, hidden layer and output layer. Input layer is the beginning of the workflow for the ANN. Each variable is connected to each hidden variable and the hidden variable is connected to each of the output variable. The hidden layer means a neuron whose output is connected to the inputs and it does not visible as a network output. Output layer is the last layer of the neuron that produces output for the study.

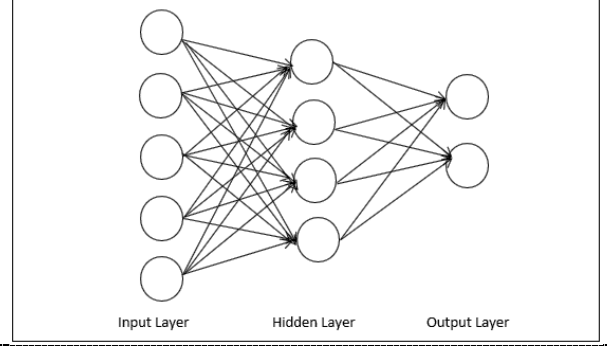


Figure 3.3 A simple Neural Network Topology

3.2 Model Evaluation

The result will be compared between those three methods of analysis. The best algorithm will be decided based on the comparison made. In this phase, prediction performance was evaluated using Model Assessment and Misclassification Rate. The detail of the prediction performance is explained as below.

3.2.1 Model Assessment

Model assessment consider three methods to get the best model which are underfit, overfit and the best model. Underfit is model that perform very well in validation set as compared to training set. While, overfit is defined by the absolute gap between training and validation results. The absolute gap between training and validation of the misclassification rate, average square mean and ROC index should be the largest result and looks for the majority results. For the model to be assumed as best model, misclassification rate and average square mean in the validation set should be the lowest value and for the ROC index at the validation set should be the largest value.

3.2.2 Misclassification rate

Misclassification rate refers to the wrong classification of the event [8]. For example, positive event has been wrongly classified at the negative event. The lowest value of the misclassification rate is better for the model. Table 3.1 shows the confusion matrix and also misclassification rate formula to be calculated.

Table 3.1 Confusion Matrix

	Predict 1	Predict 0
Actual 1	True Positive (TP)	False Negative (FN)
Actual 0	False Positive (FP)	True Negative (TN)

$$\text{Misclassification rate} = \left(\frac{FP+FN}{TP+FN+FP+TN}\right) \times 100\% \quad (2)$$

3.2.3 Formula to calculate sensitivity, specificity and accuracy

$$\text{Sensitivity} = \left(\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}\right) \times 100\% \quad (3)$$

$$\text{Specificity} = \left(\frac{\boldsymbol{TN}}{\boldsymbol{FP+TN}}\right) \times 100\% \quad (4)$$

$$\text{Accuracy} = \left(\frac{\boldsymbol{TP+FN}}{\boldsymbol{TP+FN+FP+FN}}\right) \times 100\% \quad (5)$$

The best model is determined by the overall highest percentage of sensitivity, specificity and accuracy.

4 Result and Discussion

4.1 Data Analysis Technique

4.1.1 Data Management

Data management is an administrative process to ensure the reliability and accessibility of the data. In this study, data management is done to remove biasedness of the data for target variable or dependent variable (DV) by choosing the ratio for the sample. The ratio for the graduate's employment data have to be sampled based on which data has the highest value will have 60% from public and the lowest will have 40% from private sector, thus the target data will be balanced. After the data has been sampled, it ready to be analysed.

4*.1.2 Data Cleaning*

Data cleaning is done by removed the missing values in the variables as the data does not affect the entire result. Besides, identifying outliers and also remove the duplicate data to give better result. Data cleaning is done since variable age has outlier in the data. Based on the government policy, the retirement age is 60 years old. Therefore, the age above 60 years old will assumed as outlier if they are still employed under public and private sector. The outlier will be discarded because it has extreme value. After cleaning data is done, the data is ready to be examined.

4.1.3 Data Partition

Data partition is the process of logically partitioning data into segment so that easier to be maintained or access. Partitioning of data helps the performance and utility processing. In this study, the data is divided into 70% for training and 30% for validation. We can specify the required partition percentage by the training has maximum possible percentage and the validation will have the remaining percentage. Sensitivity, specificity and accuracy will be done to calculate the right limit.

The best model is determined based on the highest sensitivity, specificity and accuracy. For misclassification rate, the best model should have the lowest value.

4.2 Model

4.2.1 Decision Tree

Table 4.1 Confusion Matrix of DT

		Predicted		**Total**
		1	**0**	
Actual	1	4550	2705	7255
	0	814	10788	11602
Total		5364	13493	18857

4.2.2 Logistic Regression

Table 4.2 Confusion Matrix of LR

		Predicted		**Total**
		1	**0**	
Actual	1	4739	2516	7255
	0	1100	10502	11602
Total		5839	13018	18857

4.2.3 Artificial Neural Network

Table 4.3 Confusion Matrix of ANN

		Predicted		**Total**
		1	**0**	
Actual	1	4764	2491	7255
	0	993	10609	11602
Total		5757	13100	18857

4.3 Model Evaluation

The results on decision tree, logistic regression and artificial neural network which have the best models are DT CART, LR MAIN + INT and ANN 5 respectively. In order to decide the best algorithm, analysis on the comparison between DT CART, LR MAIN + INT and ANN 5 have been done. The details are shown as below.

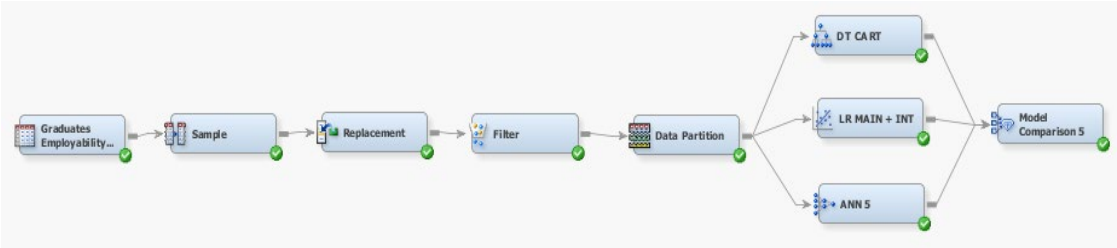


Figure 4.1: Model Comparison

Based on figure 4.1, the model comparison consists off DT CART, LR MAIN + INT and ANN.

4.4 Model Assessment

Table 4.4: Model Assessment of the best model

Model Description	Train: Average Squared Error	Valid: Average Squared Error	Gap	Train: Misclassification Rate	Valid: Misclassification Rate	Gap	Train: Roc Index	Valid: Roc Index	Gap
ANN	0.1306	0.1328	0.0021	0.1820	0.1847	0.0027	0.8730	0.8700	-0.0030
DT CART	0.1330	0.1352	0.0022	0.1841	0.1866	0.0025	0.8660	0.8630	-0.0030
LR MAIN + INT	0.1346	0.1372	0.0027	0.1866	0.1917	0.0051	0.8670	0.8640	-0.0030

Table 4.5 Comparison of the Models

Method	Model Evaluation			
	Sensitivity	**Specificity**	**Accuracy**	**Misclassification Rate**
DT	62.72%	92.98%	81.34%	18.66%
LR	65.32%	90.52%	80.82%	19.18%
ANN	65.67%	91.44%	81.52%	18.48%

Table 4.5 shows that the artificial neural network model which is ANN is the best model since the value of sensitivity and accuracy which are 65.67% and 81.52% have the highest value compared to DT CART and LR MAIN + INT. The value of misclassification ANN 5 which is 18.48% is also the lowest compared to the other models.

5. Conclusion

ANN is better in predicting the graduates who employed under private sector (Y=0) as specificity value 91.44% is highest compared to sensitivity value 65.67%. The accuracy value of ANN 5 is 81.52% while the misclassification rate is 18.47% which is good enough compared to other models. Lastly, artificial neural network (ANN 5) is the best algorithm in predicting graduates under private sector.

Acknowledgement

The authors are grateful to the financial support provided by the Faculty of Computer and Mathematical Sciences Universiti Teknologi MARA, Shah Alam, Selangor.

References

[1] 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.

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

[3] Jantawan, B., & Tsai, C. (2013). The Application of Data Mining to Build Classification Model for Predicting Graduate Employment. CoRR, abs/1312.7123.

[4] Nghe, N. T., Janecek, P., & Haddawy, P. (2007). *A Comparative Analysis of Techniques for Predicting Academic Performance, (November)*. https://doi.org/10.1109/FIE.2007.4417993

[5] Aieda, N., Bakar, A., Mustapha, A., & Nasir, K. (2014). *Clustering Analysis for Empowering Skills in Graduate Employability Model*, *7*(December 2013), 21–28.

[6] Abdul, I., Azmi, G., & Hashim, R. C. (2018). The Employability Skills Of Malaysian University Students, 1–14.
[7] Ahmad, N. W., Mawar, M. Y., & Ripain, N. The Exploration Study on Employability of Islamic Banking and Finance Graduates.
[8] R. Hamzah, N. Jamil, K. A. F. A. Samah, N. N. A. Mangshor, N. Sabri and R. Roslan, "Comparing statistical classifiers for emotion classification," *2017 7th IEEE International Conference on System Engineering and Technology (ICSET)*, Shah Alam, 2017, pp. 183-188. doi: 10.1109/ICSEngT.2017.8123443
[9] Wah, Y. B., Rahman, H. A. A., He, H., & Bulgiba, A. (2016, June). *Handling imbalanced dataset using SVM and k-NN approach*. In *AIP Conference Proceedings* (Vol. **1750**, No. 1, p. 020023). AIP Publishing.
[10] Embong, R., Aziz, N. N. A., Karim, A. A., & Ibrahim, M. R. (2017, September). *Colour application on mammography image segmentation*. In *Journal of Physics: Conference Series*(Vol. **890**, No. 1, p. 012066). IOP Publishing.
[11] Abu Bakar, N. (2018). Managing economic and Islamic research in big data environment: from computer science perspective. Journal of Emerging Economies & Islamic Research, 6(1), 1-5.
[12] Samsurim, S., Kamal, N. A. M., Ismail, M., & Diah, N. M. (2018). Prediction Outcome for Massive Multiplayer Online Games Using Data Mining. Indonesian Journal of Electrical Engineering and Computer Science, 11(1), 248-255.
[13] Tajul, M., Ab, R., & Yusof, Y. (2016). Graduates Employment Classification using Data Mining Approach, *20002*. https://doi.org/10.1063/1.4960842
[14] Aieda, N., Bakar, A., Mustapha, A., & Nasir, K. (2014). Clustering Analysis for Empowering Skills in Graduate Employability Model, *7*(December 2013), 21–28.