Supervised and Unsupervised Data Mining Techniques on Employability of Public Higher Learning Institute Graduates in Malaysia

Wan Nurul Dalilah Wan Pauzi¹*, Haliza Hasan¹ and Zamalia Mahmud¹

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Shah Alam, Selangor, Malaysia.

*wannuruldalilah96@gmail.com

Abstract. It is the students’ dream to secure a job right after graduation. However, there are factors that hinder their employability. This study aims to predict Malaysian graduates’ employment status based on employability factors and to profile the graduates’ satisfaction towards their curricular activities and information and communications technology (ICT) skills. A total of 375,507 student records were obtained based on tracer studies conducted by the Malaysian Ministry of Higher Education between 2015 and 2018. Due to the large amount of data with various categories, supervised and unsupervised data mining techniques were used to unmask the underlying variables and reveal hidden information about graduates’ employability for better tracing the employment status of graduates. Various types of consolidation techniques were also used to reduce the number of levels for categorical inputs in the dataset, namely, classifiers without consolidation, with manual consolidation, and with tree consolidation. Three types of data mining variable selections were used to improve the performance of the classifiers in predicting employment status. The results show that logistic regression (LR) without variable selection is the best classifier for data without consolidation, while LR using variable selection with LR stepwise is the best classifier for data with manual and tree consolidations. In profiling the satisfaction of graduates, K-Means Clustering was used, which revealed seven clusters. The most prominent cluster consisted of graduates who were highly satisfied with their ICT skills but less satisfied with their curricular activities. These data mining techniques were able to trace graduates’ employment status and identify the success factors of graduates’ employability.

Keywords: Consolidation, data mining technique, Malaysia, tracer study

1. Introduction

There are many reasons why a graduate wants to be hired. But some of them prefer to further their studies because they aim higher by sharpening their skills further in order to secure an ideal job. By reducing the rate of unemployment, the nation can contribute more towards the economy in terms of the goods or services that could have been produced. Other than that, it can reduce the number of the homeless and the crime rate. Thus, it is very important to take any action to lessen the unemployment rate. Unfortunately, there have not been many studies conducted on employment status prediction using information from tracer studies provided by the Malaysian Ministry of Higher Education. The tracer
studies can provide valuable information in evaluating the outputs of education, which can be used towards further changes or development of institutions in producing the best quality of graduates. This research focused on tracing the employment status of all bachelor’s degree graduates of a public university in Malaysia from 2015 to 2018. To discover the factors that might contribute in reducing the number of unemployment in Malaysia, this research was conducted to trace graduates’ employment predictions based on several attributes using unsupervised and supervised data mining techniques. The study was conducted because it can be more comprehensive since more inputs can be included in the analysis. This approach can be used to unmask the underlying variables in order to provide more comprehensive factors that can explain the pattern of graduates’ employment status. Other than that, the hidden relationships between inputs can be identified by using Neural Network, which processes complex relationships between inputs, a process that can hardly be done via experimental designs. Moreover, a consolidation approach was involved to reduce the large number of levels for categorical inputs that existed in the dataset. To improve the performance of the classifiers in predicting employment status, a variety of variable selections were used in this study. Apart from that, this study would choose the best classifiers in predicting employment status. Furthermore, new information could be obtained when graduates were segmented into similar groups based on their satisfaction towards curricular activities and information and communications technology (ICT) skills.

2. Related Works

2.1 Employment Status
The research done on employment status by [1] classified the status into graduates who are employed, furthering their studies, upgrading skills, waiting for work placement, and unemployed. The study by [2] reported that different groups of age have a different characteristic of the labour market, such that some age groups are more or less likely to be employed, unemployed, or outside the labour force. It also stated that the inclusion or exclusion of the working-age population affects the result of employability.

2.2 Factors Contributing to Employment Status
Inputs of age, CGPA, and gender have been found to be significant attributes to the low employability of Malaysian graduates in an early detection and prediction model that merged the data set from Tracer Studies and Malaysian Soft Skills Scale (My3s) [3]. In contrast, the result by [4] found that gender and final grades of graduates contribute to perceived employability. Other than that, degree types are found to be a determinant with a positive influence on graduates’ employability [3]. It contrasts with the study done by [5], which found that program outcome attainment is not a significant factor in securing employment within the first six months after graduation. From the findings by [6], the chances of employment are found to be significant for graduates associated with high academic performance and who had internship experience. However, high academic performance by itself does not exhibit any significance on employment chances. In addition, proficiency in English, Malay, and other languages is found to be significant in a study done by [7] in predicting employment status within the first six months after graduation. English language proficiency is identified to influence towards predicting the early detection and identification of the employability of graduates [3]. The study by [8] stated that knowing a foreign language plays an important role in job hunting, as it becomes an advantage for graduates as one of their special skills. According to the study done by [9], graduates’ race and year of graduation are determined to significantly influence unemployment status and identified as the strongest determinants. The research by [3] revealed that satisfaction towards counselling on career and job interviews is found to be a significant determinant on graduates’ employability. Research is done by [10] found that 89.8% of male statistics graduates at Bahir Dar University, Ethiopia was employed compared with the female with only 60% were employed. Other than that, it is found that the increases in CGPA is not associated with any statistically significant changes in employment likelihood but the probability of being employed for graduates with second upper CGPA class is higher about 7-8% than those who are not in second upper class holders [11].
2.3 Methods Used in Previous Studies

Algorithms have been identified to determine the best classification of graduates’ employment status based on tracer studies. From literature findings, the tree-based classifier with the use of Information Gain (J48) is more suitable compared with Naive Bayes, with accuracy of 92.3% and 91.3%, respectively [7]. The study by [12], in building the most suitable algorithm for employability, found that decision tree J48 with pruned C4.5 is the best classifier. The study by [13] on graduates’ employment revealed that logistic regression had the highest accuracy with 92.47% when using a testing set of 80%, while Naive Bayes had 62.4% accuracy when 70% testing set was used. The recent study by [14] found that Artificial Neural Network (ANN) is the best model, since this method obtained the highest result with accuracy at 81.52%, sensitivity at 65.67%, and specificity at 90.52%.

Consolidation was used in the present study, since there were several large categorical inputs. The research by [14] on the analysis of profit and customer satisfaction in consumer finance used mega models that had a chain of models, such as consolidation tree followed by selection tree and consolidation neural network or regression. It also mentioned that building these consolidations was worth the effort even though they were hard to build.

3. Methodology

3.1 Background of Data

This study obtained secondary data from the Malaysian Ministry of Higher Education, which was gained originally from graduates of a public university before their convocation event with variables of interest based on the objectives of the study. This study only focused on four years of duration starting from 2015 to 2018. This research utilized R programming for data cleaning, Tableau for visualizing, and SAS Enterprise Miner for analysis of supervised and unsupervised data mining. There were 25 inputs and one target used from the data obtained. Employment status acted as the target variable, and inputs included age, gender, race, OKU status, marital status, year of convocation, CGPA, course, prerequisite, study mode, learning method, working experience, funding, internship status, band of MUET (Malaysian University English Test), grades for Bahasa Melayu and English for SPM (Sijil Pelajaran Malaysia, or Malaysian Certificate of Education), experience with guidance and counselling, and third language. ICT skills, satisfaction towards library, satisfaction towards facilities, satisfaction towards facilitators, satisfaction on curricular activities, and satisfaction on the system in the institution were scored on a scale. Detailed description about the data and its measurement is displayed in Table 1.

| Table 1. Description of Data |
|-----------------------------|
| Variable                    | Role               | Measurement unit |
| Employment Status           | Target variable    | Nominal          |
| Age                         | Input variable     | Ratio            |
| Race                        | Input variable     | Nominal          |
| OKU Status                  | Input variable     | Nominal          |
| Marital Status              | Input variable     | Nominal          |
| Year of Convocation         | Input variable     | Nominal          |
| CGPA                        | Input variable     | Ordinal          |
| Course taken during bachelor| Input variable     | Nominal          |
| Prerequisite before joining bachelor | Input variable | Nominal          |
| Study Mode                  | Input variable     | Nominal          |
| Learning Method             | Input variable     | Nominal          |
| Working Experience          | Input variable     | Nominal          |
| Funding during degree       | Input variable     | Nominal          |
| Internship status           | Input variable     | Nominal          |
| Band MUET                   | Input variable     | Ordinal          |
3.2 Analysis and Model Performance

Data preparation was done before data analysis was conducted. For analysis, supervised data mining partitioned the data into 50:50 due to the massive dataset. The algorithms that were used in this study were decision tree, logistic regression, and artificial neural network, which were mainly used for classification and prediction purposes. Consolidation was introduced in this process because a major problem was that the categorical inputs appeared to be large. There were three types of consolidations used: dataset without consolidation, dataset with manual consolidation, and dataset with tree consolidation. Variable selection was applied on all types of datasets used. The variable selection models used were a model without variable selection, a model with variable selection using stepwise, a model with variable selection using decision tree, and a model with variable selection using logistic regression stepwise. These were applied to compare and obtain the best models at classifying the prediction of employment status. The steps involved in supervised data mining for predicting employment status was described in figure 1. Figures 2, 3, and 4 show the process of supervised data mining ran using SAS Enterprise Miner.

![Figure 1. Steps diagram for supervised data mining](image1.png)

![Figure 2. Model without consolidation](image2.png)
To choose the best model for model performance (accuracy, misclassification rate, sensitivity, and specificity), model selection was chosen based on parsimony, as a simpler model was preferred and easy to interpret, and model deployment for future use was conducted. As for unsupervised data mining, K-Means Clustering was used to find the overall patterns that might reveal hidden factors to give some information. To analyse using K-Means Clustering, cluster node is used and then it is applied to segment profile node. To be specific, it was utilized to classify the graduates based on their satisfaction towards curricular activities and ICT skills.

4. Results and Discussion
The percentage of the graduates’ composition comprised 54.2% who are employed and others at 45.8% for IPTA (Institut Pendidikan Tinggi Awam, or Public Higher Learning Institutes) graduates from 2015 to 2018. Figure 5 illustrates the trends in the percentage of graduates according to employment status. The trends indicate that the percentage of employed graduates increased from year 2015 to 2018, i.e., from 9.7% to the highest peak at 15.4%, respectively; for the percentage for others, there was an increase from 11.1% to 12.1% (2015–2017) but then the percentage dropped by about 1.3% in 2018. By increasing the number of several programs as MySTEP and PROTÉGÉ can reduce the rate of unemployment and at the same time can train fresh graduates to be more marketable.
Figure 5. Trends in percentage of graduates according to employment status

Figure 6 shows the percentage of graduates based on working experience as grouped by employment status. Based on the figure shown, both “Employed” and “Others” present a high reading in no working experience and a low percentage in working experience.

Figure 6. Bar chart of percentage of graduates based on working experience according to employment status

For supervised data mining, to find the attributes that contributed towards employment status, the significant inputs were obtained from logistic regression based on their significant values; for decision tree and artificial neural network, the significant inputs were obtained through variable importance. All these outputs for determining the significant variables were executed without using any variable selection and without consolidating the inputs’ categories.

Logistic regression result shows that age, course, CGPA, year of convocation, English (SPM), third language, experience with counselling, facilitators, facilities, funding, gender, ICT skills, library, OKU status, race, study mode, system, and working experience were contributors to employment status since the $p$-values for each input were lower than the alpha value of 0.05. By referring to the variables’ importance provided by decision tree, course, prerequisite, working experience, year of convocation, age, race, study mode, funding, marital status, and gender were found to be important towards employment status. As for artificial neural network, significant inputs identified were course, working experience, year of convocation, prerequisite, third language, age, race, marital status, facilities, English (SPM), CGPA, and funding. The common inputs that were found to be significant in all three models
were course, working experience, year of convocation, age, and race, which highly contributed towards employment status.

This section provides the analysis of logistic regression (LR), decision tree (DT), and artificial neural network (ANN) models without variable selection, with variable selection using stepwise, with variable selection using DT, and with variable selection using LR stepwise. The models’ performances are presented in this section. Model comparison and the best models are also shown in this section to determine the best models in predicting employment status.

Table 2. Model performance (without consolidation)

| Model                                | Accuracy | Misclassification Rate | Sensitivity | Specificity |
|--------------------------------------|----------|------------------------|-------------|-------------|
| Without Variable Selection           |          |                        |             |             |
| Logistic Regression                   | 0.6142   | 0.3858                 | 0.7290      | 0.4784      |
| Decision Tree                        | 0.6081   | 0.3919                 | 0.7414      | 0.4506      |
| Artificial Neural Network             | 0.6160   | 0.3840                 | 0.7293      | 0.4820      |
| Variable Selection Using Stepwise    |          |                        |             |             |
| Logistic Regression                   | 0.6041   | 0.3959                 | 0.7339      | 0.4505      |
| Decision Tree                        | 0.6048   | 0.3952                 | 0.7779      | 0.4001      |
| Artificial Neural Network             | 0.6101   | 0.3899                 | 0.7399      | 0.4566      |
| Variable Selection Using Decision Tree|         |                        |             |             |
| Logistic Regression                   | 0.6097   | 0.3903                 | 0.7323      | 0.4645      |
| Decision Tree                        | 0.6077   | 0.3923                 | 0.7427      | 0.4480      |
| Artificial Neural Network             | 0.6167   | 0.3833                 | 0.7356      | 0.4760      |
| Variable Selection Using Logistic Regression Stepwise | | | | |
| Logistic Regression                   | 0.6142   | 0.3858                 | 0.7289      | 0.4785      |
| Decision Tree                        | 0.6081   | 0.3919                 | 0.7414      | 0.4506      |
| Artificial Neural Network             | 0.6192   | 0.3808                 | 0.7238      | 0.4955      |

Table 2 shows the models’ performances evaluated from the confusion matrix. All models generated accuracy above 60% and their relative misclassification rates were below 40%. ANN with variable selection using logistic regression stepwise outperformed the other models, since its accuracy was the highest at 61.9% with a misclassification rate of 38.08%. This was followed by ANN with variable selection using decision tree (accuracy=61.67%, misclassification rate=38.33%), ANN without variable selection (accuracy=61.60%, misclassification rate=38.40%), LR without variable selection and LR with variable selection using logistic regression stepwise (accuracy=61.42%, misclassification rate=38.58%), and ANN with variable selection using stepwise (accuracy=61.01%, misclassification rate=38.99%). The remaining models (DT without variable selection, LR and DT with variable selection using stepwise, LR and DT with variable selection using decision tree, and DT with variable selection using logistic regression stepwise) achieved accuracy below 61%.

As for sensitivity, all DT models generated the highest magnitude in all methods. All produced sensitivity above 74% of graduates correctly classified as employed. For specificity, all ANN models show the highest results in all approaches used.
Table 3. Model comparison for models without consolidation

| 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 | 
|------------------|------------------------------|------------------------------|-----|-------------------------------|-------------------------------|-----|-----------------|-----------------|-----| 
| 1: ANN           | 0.2285                       | 0.2295                       | 0.0010 | 0.3782                       | 0.3808                       | 0.0027 | 0.661           | 0.656           | 0.005 | Best model      |
| 1: ANN           | 0.2291                       | 0.2308                       | 0.0016 | 0.3795                       | 0.3833                       | 0.0039 | 0.658           | 0.651           | 0.007 | Overfit         |
| 1: ANN           | 0.2291                       | 0.2301                       | 0.0010 | 0.3791                       | 0.3840                       | 0.0049 | 0.657           | 0.653           | 0.004 | Best model      |
| 1: LR            | 0.2295                       | 0.2307                       | 0.0011 | 0.3805                       | 0.3858                       | 0.0053 | 0.656           | 0.651           | 0.005 | Best model      |
| 1: LR            | 0.2295                       | 0.2307                       | 0.0012 | 0.3807                       | 0.3858                       | 0.0051 | 0.656           | 0.651           | 0.005 | Best model      |
| 1: LR            | 0.2319                       | 0.2327                       | 0.0008 | 0.3881                       | 0.3899                       | 0.0018 | 0.645           | 0.642           | 0.003 |                  |
| 1: LR            | 0.2316                       | 0.2328                       | 0.0012 | 0.3874                       | 0.3903                       | 0.0030 | 0.646           | 0.641           | 0.005 |                  |
| 1: DT            | 0.2348                       | 0.2355                       | 0.0007 | 0.3893                       | 0.3919                       | 0.0025 | 0.628           | 0.624           | 0.004 |                  |
| 1: DT            | 0.2348                       | 0.2355                       | 0.0007 | 0.3893                       | 0.3919                       | 0.0025 | 0.628           | 0.624           | 0.004 |                  |
| 1: DT            | 0.2349                       | 0.2356                       | 0.0007 | 0.3897                       | 0.3923                       | 0.0026 | 0.627           | 0.624           | 0.003 |                  |
| 1: DT            | 0.2363                       | 0.2365                       | 0.0002 | 0.3946                       | 0.3952                       | 0.0005 | 0.62            | 0.619           | 0.001 |                  |
| 1: LR            | 0.2340                       | 0.2346                       | 0.0006 | 0.3952                       | 0.3959                       | 0.0007 | 0.636           | 0.633           | 0.003 |                  |

According to table 3, there is one overfit model, which belongs to ANN with variable selection using decision tree (1: ANN). An overfit model can be determined when the performance of data in Valid is greater than the training data. Thus, the ANN model with variable selection using decision tree (1: ANN) is said to be an overfit model because it has the largest value in the gap between Train and Valid of average squared error (ASE), misclassification rate, and ROC Index.

There is no underfit model, since the performances of all models have the best validation results. To choose the best models, three candidates were chosen as the best models in predicting employment status for models without consolidation. Table 4 narrows down the information based on each best model.

Table 4. Candidates for best model without consolidation

| Model Description | Accuracy | Sensitivity | Specificity | Misclassification Rate |
|-------------------|----------|-------------|-------------|------------------------|
| LR Without Variable Selection (1: LR) | 0.6142 | 0.7290 | 0.4784 | 0.3858 |
| ANN Without Variable Selection (1: ANN) | 0.6160 | 0.7293 | 0.4820 | 0.3840 |
| ANN with Variable Selection Using LR Stepwise (1: ANN) | 0.6192 | 0.7238 | 0.4955 | 0.3808 |

Based on accuracy, ANN with variable selection using logistic regression stepwise has the highest magnitude with 61.92% compared with LR without variable selection and ANN without variable selection with 61.42% and 61.60%, respectively. For sensitivity, ANN without variable selection has a 0.03% higher value than does LR without variable selection and a 0.55% higher value than does ANN with variable selection using logistic regression stepwise. To choose the best model, the parsimony criterion was used, and LR without variable selection was chosen, since the ANN model is complicated due to its hidden relationships and is hard to interpret. Therefore, LR without variable selection was chosen as the best model for models without consolidation.

The next section explains the models using manual consolidation on large categorical inputs based on researchers’ literacy.

Table 5. Model performance (manual consolidation)

| Model Description | Accuracy | Misclassification Rate | Sensitivity | Specificity |
|-------------------|----------|------------------------|-------------|-------------|
| Logistic Regression | 0.6121 | 0.3879 | 0.7325 | 0.4698 |
| Decision Tree | 0.6048 | 0.3952 | 0.7521 | 0.4304 |
| Artificial Neural Network | 0.6175 | 0.3825 | 0.7330 | 0.4810 |

Variable Selection using Stepwise
Logistic Regression 0.5993 0.4007 0.7389 0.4342
Decision Tree 0.6032 0.3968 0.7638 0.4132
Artificial Neural Network 0.6056 0.3944 0.7542 0.4298

| 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 |
|-------------------|------------------------------|------------------------------|-----|------------------------------|------------------------------|-----|--------------------|--------------------|-----|
| 2 1: ANN          | 0.2284                       | 0.2298                       | 0.0014 | 0.3782                     | 0.3825                     | 0.0043 | 0.661 | 0.655 | 0.006 |
| 2 4: ANN          | 0.2289                       | 0.2302                       | 0.0013 | 0.3783                     | 0.3829                     | 0.0046 | 0.659 | 0.654 | 0.005 |
| 2 3: ANN          | 0.2305                       | 0.2309                       | 0.0004 | 0.3828                     | 0.3832                     | 0.0004 | 0.653 | 0.651 | 0.002 |
| 2 4: LR           | 0.2311                       | 0.2319                       | 0.0008 | 0.3853                     | 0.3879                     | 0.0026 | 0.65  | 0.647 | 0.003 |
| 2 1: LR           | 0.2311                       | 0.2319                       | 0.0008 | 0.3851                     | 0.3879                     | 0.0028 | 0.65  | 0.647 | 0.003 |
| 2 3: LR           | 0.2335                       | 0.2340                       | 0.0005 | 0.3929                     | 0.3929                     | 0.0000 | 0.639 | 0.637 | 0.002 |
| 2 2: ANN          | 0.2342                       | 0.2345                       | 0.0004 | 0.3931                     | 0.3944                     | 0.0013 | 0.635 | 0.633 | 0.002 |
| 2 1: DT           | 0.2355                       | 0.2359                       | 0.0005 | 0.3932                     | 0.3952                     | 0.0021 | 0.625 | 0.623 | 0.002 |
| 2 4: DT           | 0.2355                       | 0.2359                       | 0.0005 | 0.3932                     | 0.3952                     | 0.0021 | 0.625 | 0.623 | 0.002 |
| 2 3: DT           | 0.2355                       | 0.2360                       | 0.0005 | 0.3932                     | 0.3954                     | 0.0022 | 0.625 | 0.622 | 0.003 |
| 2 2: LR           | 0.2359                       | 0.2363                       | 0.0004 | 0.3993                     | 0.4007                     | 0.0013 | 0.627 | 0.626 | 0.001 |

Table 5 shows the models’ performance for manual consolidation. ANN without variable selection has the highest accuracy at 61.75% and it is closely followed by ANN with variable selection using LR stepwise at 61.71%, ANN using variable selection decision tree at 61.68%, and LR without variable selection and LR with variable selection using LR stepwise both at 61.21%. The remaining models mostly hit accuracy below 61%. For sensitivity, the models’ performances are similar to models without consolidation, of which DT remains the best at correctly classifying the model for employed graduates compared with other models. ANN for all approaches comes second after DT, followed by LR, based on sensitivity values. In classifying the “Others” group of graduates, most of the models lie between 41.32% to 48.10% in terms of specificity.

Under manual consolidation, there are two overfit models identified based on table 6: ANN without variable selection (2 1: ANN) and ANN with variable selection using LR stepwise (2 4: ANN). ANN without variable selection (2 1: ANN) has a large gap between Train and Valid of ASE and ROC Index values with 0.0014 and 0.006, respectively. As for ANN with variable selection using LR stepwise (2 4: ANN), it has a large gap between Train and Valid of misclassification rate with 0.0046 and between Train and Valid of ROC Index with 0.005. Thus, these two models were identified as overfit models. There were two candidates for the best model in employment status prediction. Both ANN with variable selection using DT and LR with variable selection using LR stepwise generated low values for Valid of ASE (0.2309 and 0.2319, respectively), small values for Valid of misclassification rate (0.3831 and 0.3876, respectively), and high values for Valid of ROC Index (0.651 and 0.647, respectively).
Table 7. Candidates for best model with manual consolidation

| Model                                      | Accuracy | Sensitivity | Specificity | Misclassification Rate |
|--------------------------------------------|----------|-------------|-------------|------------------------|
| ANN with Variable Selection Using DT (2:ANN) | 0.6168   | 0.7328      | 0.4796      | 0.3832                 |
| LR with Variable Selection Using LR Stepwise (2:LR) | 0.6121   | 0.7326      | 0.4697      | 0.3879                 |

ANN with variable selection using DT (2: ANN) has a slightly higher accuracy value than does LR with variable selection using LR stepwise (2: LR), where 61.68% were correctly classified as employed and “Others” compared with LR with variable selection using LR stepwise (2: LR) with accuracy of 61.21%. As for sensitivity, both models have a very small difference in classifying the employed group, with values of 73.28% and 73.26%. The same goes with specificity, in which there is a 1% difference in value between both models. Since the magnitude for every performance has a slightly small difference in values, thus the parsimony characteristic was preferred to determine the best model at predicting the employment status of graduates between 2015 to 2018. Other than that, the fact that ANN is known as a complex model, which has hidden relationships between the inputs, thus contributed to the selection of LR with variable selection using LR stepwise as the best model among the models with manual consolidation. Table 8 shows the result for the models’ performances under tree consolidation.

Table 8. Model performance (tree consolidation)

| Model                                      | Accuracy | Misclassification Rate | Sensitivity | Specificity |
|--------------------------------------------|----------|------------------------|-------------|-------------|
| Without Variable Selection                 |          |                        |             |             |
| Logistic Regression                         | 0.6100   | 0.3900                 | 0.4447      | 0.7497      |
| Decision Tree                              | 0.6082   | 0.3918                 | 0.4506      | 0.7414      |
| Artificial Neural Network                   | 0.6139   | 0.3861                 | 0.4489      | 0.7535      |
| Variable Selection Using Stepwise          |          |                        |             |             |
| Logistic Regression                         | 0.6048   | 0.3952                 | 0.4535      | 0.7327      |
| Decision Tree                              | 0.6048   | 0.3952                 | 0.4001      | 0.7779      |
| Artificial Neural Network                   | 0.6124   | 0.3876                 | 0.4773      | 0.7266      |
| Variable Selection Using Decision Tree     |          |                        |             |             |
| Logistic Regression                         | 0.6093   | 0.3907                 | 0.4649      | 0.7313      |
| Decision Tree                              | 0.6077   | 0.3923                 | 0.4480      | 0.7427      |
| Artificial Neural Network                   | 0.6150   | 0.3850                 | 0.4677      | 0.7395      |
| Variable Selection Using Logistic Regression Stepwise |          |                        |             |             |
| Logistic Regression                         | 0.6102   | 0.3898                 | 0.4450      | 0.7499      |
| Decision Tree                              | 0.6082   | 0.3918                 | 0.4506      | 0.7414      |
| Artificial Neural Network                   | 0.6114   | 0.3886                 | 0.4521      | 0.7461      |

Table 8 reveals that ANN with variable selection using DT produced the highest accuracy among other methods (accuracy=61.50%). The difference between ANN without variable selection, ANN with variable selection using stepwise, ANN with variable selection using LR stepwise, and LR without variable selection in terms of accuracy are minimal and all of their accuracy values are above 61%. The remaining models achieved accuracy of 60% and above. For sensitivity, tree consolidation was found to produce sensitivity below 50%, which contrasted with the models without consolidation and with manual consolidation. This is shown from the sensitivity values of the models without consolidation and with manual consolidation as above 70%, while all values of sensitivity using tree consolidation
appeared to be at 40% and above. For specificity, all models correctly classified the “Others” group of employment status with high values (>72%).

| 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 |
|-------------------|-------------------------------|-----------------------------|-----|-------------------------------|-------------------------------|-----|-----------------|-----------------|-----|
| 3: ANN            | 0.2293                        | 0.2309                      | 0.0016 | 0.3821                       | 0.3850                       | 0.0029 | 0.656          | 0.649          | 0.007 |
| 3: DT             | 0.2348                        | 0.2355                      | 0.0007 | 0.3893                       | 0.3918                       | 0.0025 | 0.628          | 0.625          | 0.003 |
| 3: LR             | 0.2316                        | 0.2328                      | 0.0012 | 0.3879                       | 0.3907                       | 0.0028 | 0.646          | 0.641          | 0.005 |
| 3: ANN            | 0.2317                        | 0.2324                      | 0.0007 | 0.3867                       | 0.3876                       | 0.0009 | 0.646          | 0.643          | 0.003 |
| 3: LR             | 0.2304                        | 0.2315                      | 0.0011 | 0.3881                       | 0.3900                       | 0.0019 | 0.651          | 0.647          | 0.004 |
| 3: DT             | 0.2348                        | 0.2355                      | 0.0007 | 0.3893                       | 0.3918                       | 0.0025 | 0.628          | 0.625          | 0.003 |
| 3: LR             | 0.2349                        | 0.2356                      | 0.0007 | 0.3897                       | 0.3923                       | 0.0026 | 0.627          | 0.624          | 0.003 |
| 3: LR             | 0.2340                        | 0.2346                      | 0.0006 | 0.3952                       | 0.3959                       | 0.0007 | 0.636          | 0.633          | 0.003 |

Table 9. Model comparison for models with tree consolidation

Table 10. Candidates for best model with tree consolidation

| Model Description | Accuracy | Sensitivity | Specificity | Misclassification Rate |
|-------------------|----------|-------------|-------------|-------------------------|
| ANN Without Variable Selection (3: ANN) | 0.6139 | 0.4489 | 0.7535 | 0.3861 |
| LR with Variable Selection Using LR Stepwise (3: LR) | 0.6102 | 0.4450 | 0.7499 | 0.3898 |
| ANN with Variable Selection Using LR Stepwise (3: LR) | 0.6114 | 0.4521 | 0.7461 | 0.3886 |

Under tree consolidation, referring to table 9, one overfit model was found, which was ANN with variable selection using DT (3: ANN). This overfit model has higher values of the gap of ASE, misclassification rate, and ROC Index (0.0016, 0.0029, and 0.007, respectively) than do other models. To select the best model, there were three qualified approaches: ANN without variable selection (3: ANN), ANN with variable selection using LR stepwise (3: 4: ANN), and LR with variable selection using stepwise (3: 4: LR).

Table 10. Candidates for best model with tree consolidation

ANN without variable selection (3: 1: ANN) has the highest accuracy at 61.39% compared with LR with variable selection using LR stepwise (3: 4: LR) and ANN with variable selection using LR stepwise (3: 4: ANN) with accuracy of 61.02% and 61.14%, respectively. For sensitivity, ANN with variable selection using LR stepwise (3: 4: ANN) is the most sensitive in correctly classifying employed graduates with 45.21%. With 75.35% specificity, ANN without variable selection (3: 1: ANN) is better at classifying the “Others’” group of graduates in determining employment status. These three approaches have better performances in predicting employment status. However, an ANN model is complex, and it is hard to interpret its hidden relationships. Therefore, LR with variable selection using LR stepwise (3: 4: LR) was selected as the best model.

Table 11 displays the summary of the best models. It can be concluded that LR is the best model for all types of consolidation approaches used. For models without consolidation, LR without variable selection is the best classifier, whereas for models with manual consolidation, LR with variable selection using LR stepwise is the best model, and for models with tree consolidation, LR with variable selection using LR stepwise is the most suitable classifier.
Table 11. Summary of all best models obtained

| Consolidation | Model | Accuracy | Sensitivity | Specificity | Misclassification Rate |
|---------------|-------|----------|-------------|-------------|------------------------|
| Without Consolidation | LR without Variable Selection (1: LR) | 0.6142 | 0.7290 | 0.4784 | 0.3858 |
| Manual Consolidation | LR with Variable Selection using LR Stepwise (2: LR) | 0.6121 | 0.7326 | 0.4697 | 0.3879 |
| Tree Consolidation | LR with Variable Selection using LR Stepwise (3: LR) | 0.6102 | 0.4450 | 0.7499 | 0.3898 |

To fulfil the third objective, K-Means Clustering was used to group the same behaviours of graduates based on their satisfaction towards curricular activities and ICT skills.

Segment 1 consisted of 43.33% of the composition and shows that the graduates had high satisfaction towards ICT skills, while having low satisfaction towards curricular activities. Based on figure 9, Segment 4 has the second-highest segment composition at 34.94%. According to its distribution, this cluster had higher satisfaction towards ICT skills and moderate satisfaction towards curricular activities. This figure indicates that these graduates were highly interested in ICT skills and moderately interested in curricular activities.
Segment 10 (figure 10) consists of graduates who were highly satisfied with ICT skills only, which means these graduates were interested in ICT only, while Segments 7 and 11 have graduates who were highly satisfied with curricular activities only, which indicates that these graduates actively took part in curricular activities.

Segment 3 in figure 12 consisted of graduates who were highly interested in ICT skills and curricular activities. Thus, Segment 3 shows that these graduates were highly interested in ICT skills and actively participated in curricular activities. For the segment of “Other”, this cluster consisted of graduates who had low satisfaction towards both ICT skills and curricular activities. These graduates can be said to be less interested in both ICT skills and curricular activities compared with the rest of the segments. The following figure shows the variable–worth of Segments 1 and 4.
Figure 14. Variable–worth of Segments 1 and 4

Figure 14 portrays the relative worth of each variable in characterizing each segment. Both Segments 1 and 4 are largely characterized by the ICT skills input, but curricular activities input also plays a role.

5. **Discussion and Conclusion**

Referring to the descriptive analysis part, more than half of the graduates from 2015 to 2018 were employed. As for the trends of graduates’ employment, it was found to be increasing yearly from 2015 to 2018. The percentage of graduates with working experience and are employed is higher than the percentage of graduates with working experience and other employment status.

Based on the analysis, several factors have been determined to be significant towards employment status. These factors or attributes can be obtained from LR, DT, and ANN, all without consolidation. In this study, it was found that age is identified as a significant factor in predicting employment status. This finding is supported by the studies done by [7] and [3], in which age was found to contribute towards employability. However, this finding contradicted the study done by [16], where in age is found as not a significant input in employability. As for CGPA, it is also one of the contributors towards employment status. This is similar to the finding from the studies done by [3], [13], and [6], wherein CGPA was an important attribute in employment.

Courses available in Malaysian IPTAs were also examined in this research. The finding reveals that certain courses are significant in predicting employment status. A previous study done by [8] reached the same conclusion, revealing that graduates who studied engineering and economics and management tend to dominate the labour market. Since this study analysed all courses available in Malaysian IPTAs, courses such as Health Sciences, Arts, Communication, Education, Management, Administration, Medicine, Computer and Mathematical Sciences, and Architecture were found to highly contribute towards employment, while other courses did not contribute to the prediction of employment status.

In contribution towards employment status, race and year of convocation highly affected the model. This result is supported by the research done by [9]. As for the third language, France, Spanish, Hindi, German, Buginese, Arabic, Brunei, Cantonese, Hokkien, Iban, Javanese, Kadazan, Mandarin and Tamil were acknowledged to contribute towards predicting employment status. This finding is supported by previous studies by [7], [3], and [8]. Gender also played a significant role in predicting employment status and was agreed to be important based on the studies by [3] and [16]. ICT skills are one of the contributors towards employment prediction and it is already proven by previous studies by [7] and [8].

Graduates need to seek counselling regarding job interviews, since it is one of the contributors in employment, which is supported by the research done by [3]. The study by [6] resulted in the same finding, as graduates with working experience are much preferred for employment. Satisfaction towards facilitators is also responsible for contributing to employment status. Other inputs that were not studied by previous researchers also were added to determine whether the inputs contribute towards the
prediction of employment status. Some of the inputs were found to be significant, such as English (SPM), satisfaction towards facilities, satisfaction towards library, OKU status, and study mode. When comparing all three classifiers of DT, LR, and ANN (without consolidation), factors found to be significant towards employment status were course, age, working experience, race, and year of convocation.

To achieve the second objective in this study, the supervised data mining technique was used to analyse the prediction of employment status. According to the results generated, LR outperformed the other classifiers. Among models without consolidation, LR without variable selection is the best classifier since it achieved accuracy of 61.42%, misclassification rate of 38.58%, sensitivity of 72.90% and specificity of 47.84%. As for selecting the best model for manual consolidation, there were two candidates: ANN with variable selection using decision tree and LR with variable selection using logistic regression stepwise. These two models appeared to have small differences in model performance. Thus, due to LR being simpler than ANN, therefore LR with variable selection using logistic regression stepwise was chosen due to its parsimony. The model obtained accuracy of 61.21%, relative misclassification rate of 38.79%, sensitivity of 73.26%, and specificity of 46.97%. Lastly, for models that used tree consolidation, LR with variable selection using logistic regression stepwise was selected as the best model with its performance accuracy of 61.02%, sensitivity of 44.50%, and specificity of 74.99%. This best model was deployed to predict employment status using scoring data.

| Classifier                  | The Best Classifier                | Accuracy | Misclassification rate | Sensitivity | Specificity |
|-----------------------------|-----------------------------------|----------|------------------------|-------------|-------------|
| Model without Consolidation | LR without Variable Selection     | 61.42%   | 38.58%                 | 72.90%      | 47.84%      |
| Model with Manual Consolidation | LR with Variable Selection using LR Stepwise | 61.21%   | 38.79%                 | 73.26%      | 46.97%      |
| Model with Tree Consolidation | LR with Variable Selection using LR Stepwise | 61.02%   | 38.98%                 | 44.50%      | 74.99%      |

To cluster the behaviours of the graduates, K-Means Clustering was used. It revealed that for the first segment, which had the largest segment composition, the graduates were highly satisfied with ICT skills but had low satisfaction towards curricular activities. This was followed by the fourth cluster, wherein graduates were highly satisfied with ICT skills and moderately satisfied with curricular activities. Segment 10 consisted of graduates who were interested in and highly satisfied with ICT skills only, while Segments 7 and 11 consisted of graduates who were highly satisfied with and actively participated in curricular activities only. For Segment 3, these graduates were grouped together based on their high satisfaction towards curricular activities and ICT skills, which indicates that they participated actively and had a high interest in ICT skills that were provided by the institution. Lastly, Segment “Other” consisted of graduates who had low satisfaction towards ICT skills and curricular activities.

In conclusion, all objectives have been achieved in this study. However, the models’ performances were not high as expected and the reasons need to be further studied. Although there was a lack in the models’ performances, the study is believed to be important in providing useful and meaningful insights towards the prediction of employment status among IPTA graduates in Malaysia.

Acknowledgments
The authors would like to extend their appreciation to the Malaysian Ministry of Higher Education for providing the data; to Universiti Teknologi MARA, Shah Alam, for providing the Research Entity Initiative (REI) Grant No. 600-IRMI/REI 5/3 (005/2019) enabling this paper to be accepted for publication.
References

[1] Abdul Rahman N A, Tan K L and Lim C K 2017 Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students Journal of Telecommunication, Electronic and Computer Engineering 9(2-12) 155–61

[2] International Labour Organization 2019 Quick Guide on Interpreting the Unemployment Rate (International Labor Organization)

[3] Awang-Hashim R, Lim H E, Yatim B, Tengku Ariffin T F, Zubairi A M, Yon H and Osman O 2015 Estimating a Prediction Model for the Early Identification of Low Employability Graduates in Malaysia Singapore Economic Review 60(4) 1–22

[4] Pinto L H and Ramalheira D C 2017 Perceived Employability of Business Graduates: The Effect of Academic Performance and Extracurricular Activities Journal of Vocational Behavior 99 165–178

[5] Gamboa R A, Namavivayam S N, Singh R, Vaithilingam C A and Faouladi M H 2018 PO Attainment vs. Employment Status: Correlation Study for Taylor’s University School of Engineering 2018 IEEE 10th International Conference on Engineering Education (ICEED) 100–104

[6] Nunley J M, Pugh A, Romero N and Seals R A 2016 College Major, Internship Experience, and Employment Opportunities: Estimates from a Résumé Audit Labour Economics 38 37–46

[7] Sapaat M A, Mustapha A, Ahmad J, Chamili K and Muhammad R 2011 A Data Mining Approach to Construct Graduates Employability Model in Malaysia International Journal on New Computer Architectures and Their Applications 1(4) 1086–98

[8] Jun K 2017 Factors Affecting Employment and Unemployment for Fresh Graduates in China Unemployment—Perspectives and Solutions pp 53–70

[9] Baldry K 2016 Graduate Unemployment in South Africa: Social Inequality Reproduced Journal of Education and Work 29(7) 788–812

[10] Woya A A. 2019. Employability among statistics graduates: Graduates’ attributes, competence, and quality of education. Education Research International, 2019.

[11] Soon J, Lee A. S, Lim H and Idris I 2019 Studies in Higher Education Cubicles or corner offices Effects of academic performance on university graduates’ employment likelihood and salary. Studies in Higher Education, 0(0) 1–16

[12] Mishra T, Kumar D and Gupta S 2017 Students’ Performance and Employability Prediction Through Data Mining: A Survey Indian Journal of Science and Technology 10(24) 1–6

[13] Ab Aziz M T R and Yusof Y 2016 Graduates Employment Classification Using Data Mining Approach AIP Conference Proceedings 1761 020002

[14] A’rifian N I N, Mohd Daud N S A, Muhamad Romzi A F and Md Shahri N H N 2019 A Comparative Study on Graduates’ Employment in Malaysia by Using Data Mining Journal of Physics: Conference Series 1366 012120 https://doi.org/10.1088/1742-6596/1366/1/012120

[15] Wang C and Zhuravlev M 2014 An Analysis of Profit and Customer Satisfaction in Consumer Finance Case Studies in Business, Industry and Government Statistics 2(2) 147–56

[16] Piad K C 2018 Determining the Dominant Attributes of Information Technology Graduates Employability Prediction Using Data Mining Classification Techniques Journal of Theoretical and Applied Information Technology 96(12) 3780–90