International Journal on Advanced Science Engineering Information Technology

Vol.8 (2018) No. 4-2
ISSN: 2088-5334

Classification Techniques for Predicting Graduate Employability

Zalinda Othman*, Soo Wui Shan*, Ishak Yusoff[#], Chang Peng Kee[#]

** Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia.*
E-mail:zalinda@ukm.edu.my, wuishansoo@gmail.com

[#] Career Advancement Centre (UKM-Karier), Level 5, Pusanika, 43600 UKM Bangi, Selangor.
E-mail:iby@ukm.edu.my, chang@ukm.edu.my

Abstract*— Unemployment is a current issue that happens globally and brings adverse impacts on worldwide. Thus, graduate employability is one of the significant elements to be highlighted in unemployment issue. There are several factors affecting graduate employability, traditionally, excellent academic performance (i.e., cumulative grade point average, CGPA) has been the most dominant element in determining an individual's employment status. However, researches have shown that not only CGPA determines the graduate employability; in fact other factors may influence the graduate achievement in getting a job. In this work data mining techniques are used to determine what are the factors that affecting the graduates. Therefore, the objective of this study is to identify factors that influence graduates employability. Seven years of data (from 2011 to 2017) are collected through the Malaysia's Ministry of Education tracer study. Total number of 43863 data instances involved in this employability class model development. Three classification algorithms, Decision Tree, Support Vector Machines and Artificial Neural Networks are used and being compared for the best models. The results show decision tree J48 produces higher accuracy compared to other techniques with classification accuracy of 66.0651% and it increased to 66.1824% after the parameter tuning. Besides, the algorithm is easily interpreted, and time to build the model is small which is 0.22 seconds. This paper identified seven factors affecting graduate employability, namely age, faculty, field of study, co-curriculum, marital status, industrial internship and English skill. Among these factors, attribute age, industrial internship and faculty contain the most information and affect the final class, i.e. employability status. Therefore, the results of this study will help higher education institutions in Malaysia to prepare their graduates with necessary skills before entering the job market.

***Keywords*— graduate employability; multilayer perceptron; support vector machine; decision tree; classification.**

I. INTRODUCTION

According to [1] the concept of marketability refers to various skills in graduates to be hired as an employee. Skills such as communication skills, teamwork, continuous learning, critical thinking, entrepreneurship, and information management are crucial for a graduate to be hired. The numbers of graduates from Malaysian universities have shown a positive increment from 2006 to 2017. In 2006, the total of graduates was 132899 and increased to 299537 in 2017. Based on a report published in [2], the percentage of unemployment of undergraduate students in Malaysia is decreased from 36.4% in 2006 to 26.27.3% in 2017. Even though the rate of unemployment is decreasing, the issues of unemployment in certain disciplines still remain high and the perception of that unemployment of graduates is due to their lack of generic skills. In effort to address these issues and to increase the employability rate, Malaysia Ministry of Education has initiated several steps such as revising curriculum, promoting entrepreneurship courses, emphasizing skill and competencies such as English language, teamwork and analytical skills. Besides that, successful collaborations between university, industries and government may benefit the graduates by promoting their skills to employers in industry [3].

Employability skills have been a subject of research where the skills acquired by graduates could be determined and measured [4]. There are many approaches could be employed in this study, quantitative or qualitative study. One current approach is by employing data mining techniques. Data mining or Knowledge Discovery in Databases (KDD) is a process of extracting knowledge or hidden patterns from a large datasets. It has been proven to be an effective process in solving real-life problems. Several domains, such as financial, climate change, health and safety, stock market and others would benefit from the data mining approach. An example how data mining has been used to predict rainfall has been shown in [5, 6, 7]. Beside prediction task, data

mining has been used to detect e-learning courses anomalies as explored in [8].

In this work, data mining is used to identify graduates employability. This technique requires corresponding data such as the graduates' background, their experiences when studying in university, the effectiveness of the system and self-readiness, current status, employment status (working or unemployed) and others. These data are collected from a Malaysia's tracer study, *Sistem Kebolehpasaran Graduan* (SKPG) that was managed by Ministry of Education. Every graduates need to submit a survey before their convocation day. Data mining approach with the classification technique can produce a model of graduates' employability. By using different classification techniques such as Artificial Neural Networks (ANN), Decision Tree and Support Vector Machine (SVM), factors that affect graduates employability such as academic's achievement, differences in academic's discipline, family's background and many more can be identified.

Bayes Theorem and Decision Tree are used to build a classification model in classifying graduates whether they were working, not working and undetermined [9]. They used data from Maejo University of Thailand. The data were from three academic years that consist 11,853 instances. Ten algorithms used in modeling the classification, i.e. five types of decision tree and five types of Naïve Bayes. In their work, J48 showed the highest accuracy (98.31%) compared to others decision trees. Meanwhile, algorithm WAODE showed the highest accuracy (i.e. 99.77%).

Meanwhile, a research by [10] compared Bayes approach with a number of decision trees based algorithms. Information gain was used to evaluate attributes and found three main attributes affecting the employability. The attributes were job sector, job status and reason for not employ. Data from tracer study for 2009 was used. It contained 12830 instances with 20 background attributes related to19 public and 138 private universities. The results showed J48 has the highest accuracy (i.e. 92.3%) compared to Bayes. They concluded, decision tree algorithm J48 is a suitable algorithm in tracing the data because of its information enquiry strategy.

A research in [11] used the classification approach with Bayesian technique to build a model of graduates' employability and predict graduates employment status. Graduates data were collected from Khon Kaen University, Thailand in 2009 that consists of 3090 examples and 17 attributes. Six algorithms under the Bayesian technique concluded that *Averaged One-Dependence Estimators with subsumption resolution* (AODEsr) algorithm achieved the highest percentage of accuracy, which is 98.3%. This followed by AODE algorithm (96.1%). This research showed that three factors that affect jobs which are the place of the job, type of jobs and time of jobs.

Another example of employability research used data from 633 students of MARA Profesional College Malaysia [12]. The objective was to classify whether the graduates are working, not working or further study. Five Weka algorithms were used: Naïve Bayes, Logistik Regression, MLP, k-nearest neighbor and J48. The results showed Logistic Regression give the highest accuracy, i.e. 92.5%.

Graduates data from 1400 students of Master of Computer Application (MCA) of colleges in India have been collected and used in [13]. A number of classification techniques used to predict employability of MCA graduates. In their work, they concluded that J48 is the most suitable technique to predict employability with 70.19% accuracy. Beside accuracy, J48 can be easily interpreted, and the time taken to build the model is less (compared to Random Forest). The study identified student empathy, drive and stress management are the main emotional skill parameters that affect employability.

Research by [14] used data mining approach. Two clustering algorithms, X-Means and Support Vector Clustering, and Naïve Bayes as a classification algorithm were used in their study. The study concluded X-Means able to do the prediction better than other algorithms.

Table 1 shows the summary of a number of techniques used in predicting graduates employability.

TABLE I
SUMMARY OF CURRENT WORKS

[9]	Naïve Bayes (AODE, WAODE, BayesNet, HNB, Naïve Bayes)Decision Tree (BFTree, NBTree, REPTree, ID3, C4.5)	Highest accuracy is WAODE (99.77%) followed J48 (98.31%).
[10]	Bayes techniques (AODE, AODEsr, WAODE, Bayes Network, HNB, Naïve Bayesian, Naïve Bayesian Simple and Naïve Bayesian Updateable). Decision trees (ID3, J48, REPTree, J48graft, Random Tree, Decision Stump, LADTree, Random Forest and Simple Cart)	J48 shows highest accuracy (92.3%). Job scope is among the reasons of unemployed.
[11]	Bayesian (AODE, AODEsr, Bayesian Network, Naïve Bayesian, Naïve Bayesian Simple dan Naïve Bayesian Updateable)	Algorithm with highest accuracy is AODEsr (98.3%). Three factors affecting employability are location, type of job and times to find work.
[12]	Naïve Bayes, Logistic Regression, MLP, k-nearest neighbor Decision Tree (J48)	Logistic Regression gave highest accuracy, i.e. 92.5%, with 80% training and 20% testing.
[13]	Bayesian methods, MLP, SMO, Ensemble Methods Decision Tree(J48)	J48 shows highest accuracy (70.19%). Main emotional parameters in affecting employability are emphaty, drive and stress management.
[14]	Cluster model (X-Means dan SVC) and classification model (Naïve Bayes)	X-Means is the most accurate with 83% compared to SVC (81%) and Naïve Bayes (77%).

This paper focuses to identify the factors that affect graduates employability and to compare the classification techniques.

II. Material and Method

Data mining is important, as many sets of data can be extract to usable pattern. The most basic form of data for data mining application are database, data warehouse and transaction data. Most people believe that knowledge discovery of data is used widely and the others believe that data mining is one of the crucial steps in the process of discovery of knowledge [15].

Classification approach is one of the most important data mining task especially for predicting. The approach not only handle a large amount of data sets but also find hidden pattern in making conclusion and reduce data generation structure with ease. It is a process that identify objects categories based on their characteristics. For example, we can use a classification model to classify graduates employability whether they are employed, unemployed or uncertain. Decision tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN) and many other algorithms can be used in classification modeling [16].

In this study, three approaches are used, Decision Tree, ANN and SVM. Decision tree is a tree like flowchart where the internal node represents tests on attributes, every branch represent the test's results and every leave nodes represent the class labels or classification [17]. Leave nodes show the example of classes. Examples are classified by arranging them from the bottom of the tree from root nodes to some leaves nodes.

ANN is a mathematical model that tries to simulate structures and functions of biological neural network. Building blocks of every artificial neural networks is an artificial neural which is the basic mathematic model (function). This kind of model consists of three sets of rules: multiplication, addition and activation. The entry of each value from artificial neural is multiplied with individual weights. On the middle side of the artificial neural is the total function that includes all the inputs' weight. At the end of the artificial neural is the total input that has been weighed and already went through activation phase that is also called transfer function [18].

SVM was first introduced by Vapnik in 1960s as a classification model and recently have been an intense field of research as there is a development in the techniques and theories that are widely range from regression estimation to the density. SVM develops from statistical learning theory with the aim to solve problems without causing greater problem as a mid step [19].

This research consists of three phases. The first phase includes identify the issues, collect and choose data from SKPG. Second phase is to clean and process the data. In this phase the data will be analyzed, grouped, cleaned and transformed. The last phase, pattern identification, is where pattern's interpretation and evaluation take place by using data mining approach with classification technique such as Decision Tree, SVM and ANN.

A. Data Pre-processing

The first step until the fourth are different phases of pre-processing data that were used to prepare sets of data for mining. Pre-processing is important in the process of finding results as the quality of the results depends on the quality of data. Detect data anomalies and correct them earlier and diminish some sets of data to be analyzed can brings advantage when deciding on a conclusion.

Data collection phase is the first phase in model's development methodology. This research used data from the SKPG, particularly data of University Kebangsaan Malaysia (UKM) graduates as a case study. These data sets include the seven years of data from 2011 to 2017. Table 2 shows the total amount of data that was collected from SKPG's report. The total data instances are 43863.

TABLE II
TOTAL NUMBER OF DATA

Year	No. of Data
2011	6925
2012	6789
2013	6044
2014	5538
2015	5325
2016	5889
2017	7353
Total	43863

Data integration process is the first step in the planning and preprocessing data. It is a technical combination that is use to combine sets of data from different sources and information. Data integration from 2011 to 2017 has been carried out as the data are from different datasets. These data have been rearranged by years in excel format. These seven years of data have been integrated by using WEKA with *Simple CLI* in application menu. *Append* method have been used in this research.

Graduates with other certificates than degree have been removed as this research only focuses on undergraduate students. Data from other level of studies such as Diploma, Ph.D., Master, Advanced Diploma, Medical Degree and other certificates have been removed from this research.

Data cleaning process is to remove or correct data error, incosistency data, missing data, overlapping records and to identify outliers. Missing data can be replaced with the mean for every attributes involved. The average values were taken and calculated based on overall sets of data. Average values were used to reduce disturbance in the sets of data. Outlier that were found in the sets of data also were replaced by the average values. This research uses sets of data that have been processed through statistic method and this resulting in clean, consistent sets of data and no overlapping records.

Data transformation is a process to ensure that all sets of data that were in continuous form are changed into nominal, numbered and divided into specific scales. This process is to make the modeling process easier where existing sets of data can be understood and can be used to study the pattern for building model's forecast.

Data discretion process converts continuous attributes into numbered, nominal and divided by specific scale. The purpose of this process is to simplify the data analysis process. Next, the last step for data preparation is to transform data that involving normalizations of data and construction of attributes. Normalization process is a process that classifies values of data into specific values by using minimum and maximum steps. This process is also to simplify sets of data by using scales 0.0 to 1.0.

In this work, some of the attributes have been transformed into different category, such as *cgpa* attribute. Originally this attribute is continuous, but in this project, it is transformed into grade range. The range is classed into four parts: 2.00 - 2.49, 2.50 - 2.99, 3.00 - 3.66 dan 3.67 - 4.00. Meanwhile, *e_umur* is also being transformed intro four range: 16-25, 26-35, 36-45 dan >46. *e_pendapatan* has been classified into three classes: less than RM1501, RM1501 - 3000 and more than RM3000. The continuous attributes have been transformed into nominal in preparing the data for classification. For example, the attributes *e_bidang* and *e_40* have been changed to nominal from previous numeric code values.

Feature selection is used to discrete irrelevant attributes in building a model. It helps to choose the best and useful attributes in building a model. By using related attributes, classification algorithms will increase the accuracy of prediction, shorten the duration of research and also form an easier concept. The aim of features selection process is to choose important and useful attributes to increase the percentage of accuracy in building models.

Before features selection takes place, 357 original attributes have been reduced to 26 attributes. Attributes such as *e_nama*, *e_kp*, *e_bulan_umur*, *e_hari_umur*, *e_matrik*, *e_alamat*, *e_emel*, *e_tel_rumah* and others unuseful attributes have been removed before the selection of features. Table 3 shows 26 total attributes of graduates employability before feature selection process.

TABLE III
LIST OF ATTRIBUTES AFTER ELIMINATION OF IRRELEVANCE

No	Attribute	Value	Description
1	e_jantina	Man, Woman	Gender
2	e_umur	20-29, 30-39, 40-49 dan >49	Age
3	e_keturunan	Malay, Chinese, Indian, Others	Race
4	e_negeri	Johor, Kedah, Kelantan, Selangor, Perak, Pahang, Negeri Sembilan, Terengganu, Wilayah Persekutuan Kuala Lumpur, Melaka, Pulau Pinang, Sarawak, Sabah and others.	State
5	e_muet	Band 1 to Band 3, Band 4 to Band 6, Not Applicable	Muet
6	e_fakulti	Islamic Study,Economy and Management, Social Science and Humanity, Science and Technology, Education, Health Science,Engineering an and Build, Pharmacy,Information Science and Technology, Law	Faculty
7	e_bidang	Art and Social Science, Science and Technology, Information Technology & Communication and Education	Field
8	e_cgpa	2.00-2.49, 2.50-2.99, 3.00-3.66, 3.67-4.00	CGPA
9	e_pendapatan	<RM1501, RM1501-RM3000 and >RM3000	Family Income
10	e_15_a_i	Not Active, Active, Not Applicable	Co-curriculum (Society)
11	e_15_a_ii	Not Active, Active, Not Applicable	Co-curriculum (Club)
12	e_15_a_iii	Not Active, Active, Not Applicable	Co-curriculum (Sport)
13	e_status_kahwin	Single, Married, Others	Marital Status
14	e_17	Yes, No	Industrial Internship?
15	e_32_a	Yes, No	Join any enterpreneurship programs?
16	e_25_b	Satisfactory, Not Satisfactory	Bahasa Melayu skill
17	e_25_c	Satisfactory, Not Satisfactory	English Language Skill
18	e_25_e	Satisfactory, Not Satisfactory	Interpersonal Skill
19	e_25_f	Satisfactory, Not Satisfactory	Critical and creative thinking
20	e_25_g	Satisfactory, Not Satisfactory	Problem solving skill
21	e_25_h	Satisfactory, Not Satisfactory	Analytical skill
22	e_25_i	Satisfactory, Not Satisfactory	Team work
23	e_25_j	Satisfactory, Not Satisfactory	Positive values
24	e_25_k	Satisfactory, Not Satisfactory	General knowledge
25	e_40	Employed, Not Employed	Current Status
26	e_terima_bantuan_kewangan	Yes, No	Financial assistance

In this work, WEKA is used to select attributes by employing *Attribute Evaluator*. *InfoGainAttributeEval* has been selected to evaluate the attributes. It evaluates the worth of an attribute by measuring the information gain with respect to the class.

$$\text{InfoGain(Class,Attribute)} = H(\text{Class}) - H(\text{Class} \mid \text{Attribute})$$

Meanwhile *Ranker* is used to rank the most informative attributes and *Attribute Selection Mode* used 10-folds cross-validation. Results from the WEKA feature selection shows that attribute *e_umur*, *e_17*, and *e_fakulti* are the top three attributes with highest average reading value. Table 4 shows attributes that were chosen to be used in the classification modelling after considering the experts opinion, feature selection and past researches.

TABLE IV
SELECTED ATTRIBUTES

No	Attribute	Value	Description
1	e_umur	20-29, 30-39, 40-49 dan >49	Age
2	e_fakulti	Islamic Study,Economy and Management, Social Science and Humanity, Science and Technology, Education, Health Science,Engineering an and Build, Pharmacy,Information Science and Technology, Law	Faculty
3	e_bidang	Art and Social Science, Science and Technology, Information Technology & Communication and Education	Field
4	e_pendapatan	<RM1501, RM1501-RM3000 and >RM3000	Family Income
5	e_15_a_ii	Not Active, Active, Not Applicable	Co-curriculum
6	e_status_kahwin	Single, Married, Others	Marital Status
7	e_17	Yes, No	Industrial Internship?
8	e_25_c	Satisfactory, Not Satisfactory	English Language Skill
9	e_40	Working, Not Working	Employability Status

B. Performance Measurement

There are a number of measurements used to evaluate the performance of classifiers. Beside accuracy, root mean squared error, time and ROC are used to measure the classifiers' performance.

Accuracy

Accuracy measured the number of correct predictions made divided by total number of predictions made, usually in percentage. Accuracy is measured as follows;

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Where,
TP - true positive
TN - true negative
FP - false positive
FN - false negative

Root Mean Squared Error, RMSE

RMSE is used to measure the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.

ROC

ROC is used to differentiate between true positive and false positive.

$$\text{ROC for TP rate} = \frac{TP}{TP+FN}\times 100 \quad (2)$$

$$\text{ROC for FP rate} = \frac{FP}{FP+TN}\times 100 \quad (3)$$

III. Result And Discussion

In this section, results for the three algorithms are compared. The testing mechanism used is 10-folds cross validation. In WEKA, with cross validation the data samples are divided once, say 10 pieces. Then, 9 pieces are taken for training and the last piece is for testing. Then, with the same division, another 9 pieces are taken for training and the held-out piece for testing. The whole thing is repeated 10 times, using a different segment for testing each time. In other words, the dataset is divided into 10 pieces and then hold-out each of these pieces in turn for testing, train on the rest, do the testing and average the 10 results. This would be "10-fold cross validation".

The performances of the algorithms are compared based on the accuracy, ROC, RMSE and the time taken to build the model.

TABLE V
10 FOLD CROSS VALIDATION RESULTS FOR J48

	Train	Test	Time (s)	Acc (%)	RMSE	ROC
Decision Tree (J48)	90	10	0.19	64.8624	0.4622	0.703
	80	20	0.06	64.4936	0.4633	0.696
	70	30	0.05	65.0376	0.4609	0.703
	60	40	0.06	65.3508	0.4601	0.706
	50	50	0.12	65.6772	0.4595	0.706
	40	60	0.05	65.7519	0.4591	0.707
	30	70	0.28	65.6226	0.4598	0.708
	20	80	0.08	65.7024	0.4607	0.706
	10	90	0.05	65.5421	0.4625	0.697
	10-fold cross validation		0.05	66.0651	0.4584	0.707

Based on Table 5, the average accuracy for J48 is 66.0651%.

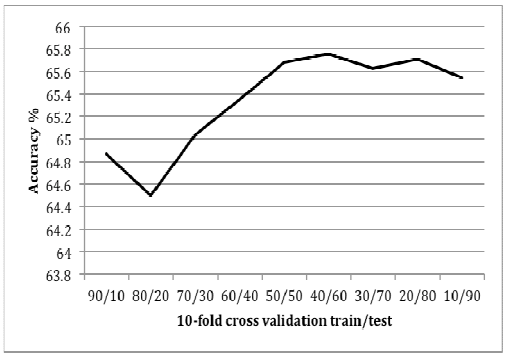

Fig. 1 10-folds cross validation for J48 accuracy

Fig. 1 shows the trend of J48 accuracy in 10-folds cross validation. The accuracy shows an increasing trend. The best accuracy is at 40 training and 60 testing fold.

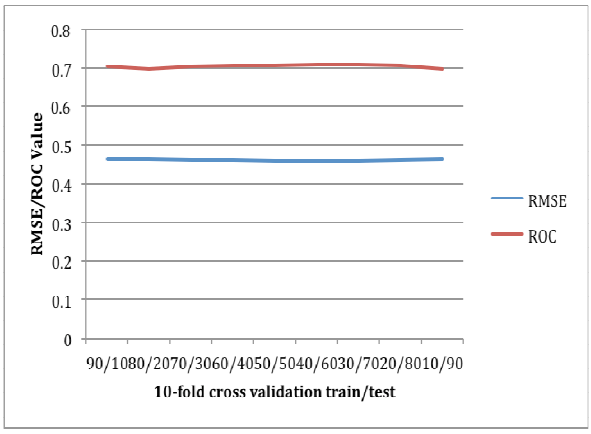

Fig. 2 10-folds cross validation of RMSE and ROC value for J48

Meanwhile, fig. 2 shows the two measures, RMSE and ROC for the 10 iterations of 10-folds cross validation of J48. The results show insignificant changes in both measures.

TABLE VI
10 FOLDS CROSS VALIDATION RESULTS FOR MLP

	Train	Test	Time (s)	Acc (%)	RMSE	ROC
Neural Network (MLP)	90	10	170.75	65.4488	0.47	0.704
	80	20	173.91	63.1852	0.4714	0.678
	70	30	133.54	64.9624	0.4646	0.697
	60	40	80.86	60.9632	0.4736	0.706
	50	50	81.19	65.605	0.4669	0.71
	40	60	80.39	64.609	0.4695	0.694
	30	70	80.39	64.6623	0.467	0.696
	20	80	80.72	64.1177	0.4676	0.698
	10	90	82.27	64.9606	0.4747	0.698
	10-fold cross validation		159.44	65.2937	0.4616	0.702

Based on Table 6, the 10-folds cross validation gives average 65.2937% accuracy.

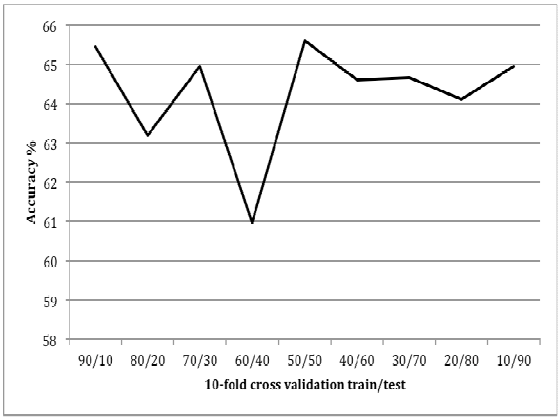

Fig. 3 10-folds cross validation for MLP accuracy

Fig. 3 shows MLP has a lowest accuracy at 60/40 fold and highest at 50/50 fold.

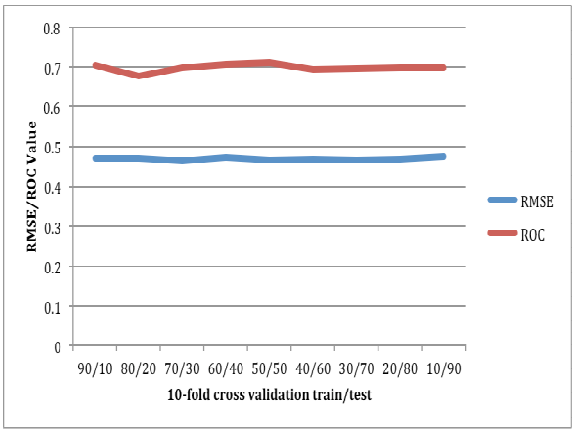

Fig. 4 10-folds cross validation of RMSE and ROC value for MLP

The two measures of RMSE and ROC have shown small changes in the ten folds. The average of RMSE is 0.4616 and ROC is 0.702.

TABLE VII
10 FOLD CROSS VALIDATION RESULTS FOR SMO

	Train	Test	Time (s)	Acc (%)	RMSE	ROC
Support Vector Machine (SMO)	90	10	78.67	65.4939	0.5874	0.655
	80	20	72.23	64.9673	0.5919	0.650
	70	30	77.19	65.3233	0.5889	0.654
	66	34	78.96	65.3264	0.5888	0.654
	60	40	73.89	65.5876	0.5866	0.656
	50	50	79.61	65.7674	0.5851	0.658
	40	60	74.05	65.8496	0.5844	0.659
	30	70	70.48	65.8417	0.5845	0.659
	20	80	71.47	65.9674	0.5834	0.660
	10	90	68.07	65.9231	0.5838	0.659
	10-fold cross validation		73.82	66.0967	0.5823	0.661

Based on Table 7, the split of 20% training and 80% testing for SMO gives the highest correctly classified, i.e. 65.9674%. In average, the accuracy is 66.0967.

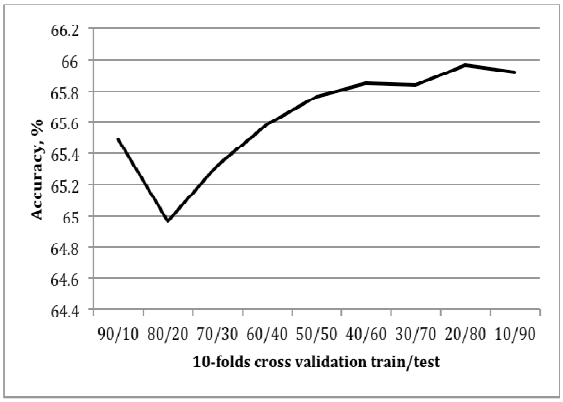

Fig. 5 10-folds cross validation for SMO accuracy

Fig. 5 shows the increasing trend of SMO accuracy. As shown in Table 7, the highest accuracy is at 20/80 fold and the lowest is at 80/20 fold.

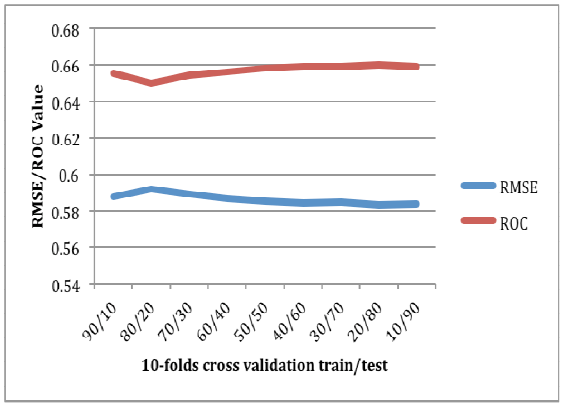

Fig. 6 10-folds cross validation of RMSE and ROC value for SMO

RMSE and ROC values for SMO are shown in Fig. 6 for 10-folds cross validation. RMSE is the highest at fold 80/20 and shows a decreasing trend. Meanwhile, ROC values has the lowest at 80/20 fold, and has a slightly increase.

TABLE VIII
RESULTS FOR 10 FOLD CROSS VALIDATION

Technique	Algorithm	Time (s)	Accuracy	RMSE	ROC
Decision Tree	J48	0.05	66.0651	0.4584	0.707
Neural Network	MLP	159.44	65.2937	0.4616	0.702
SVM	SMO	73.82	66.0967	0.5823	0.661

Table 8 shows the highest accuracy is obtained by applying SMO algorithm compared to other algorithms, i.e. 66.0967%. Second highest is 66.0651%, obtained from J48. These algorithms differ only by 0.03%. But, J48 takes the shortest time to build the model, in 0.05 seconds. Meanwhile, SMO takes 73.82 seconds.

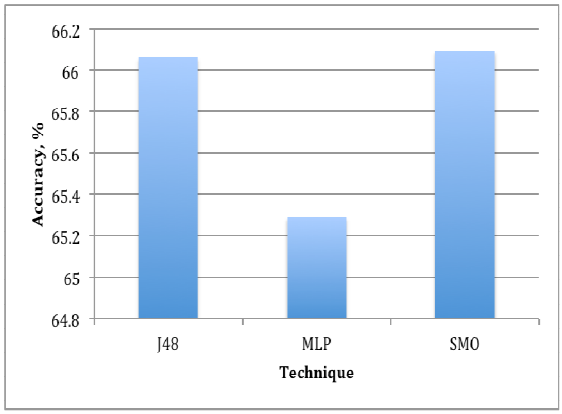

Fig. 7 Comparison of J48, MLP and SMO accuracy

The performance of the three techniques is shown in Fig. 7. The J48 and SMO have shown a good performance in terms of accuracy percentage. Meanwhile, MLP has not perform well in this study.

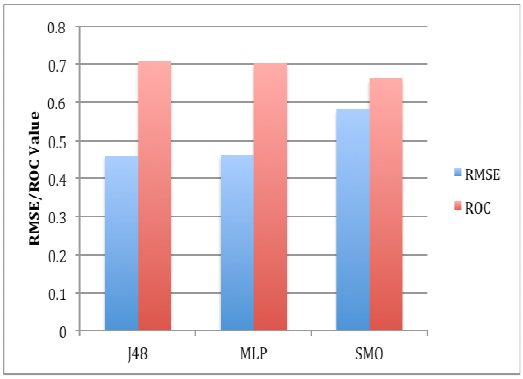

Fig. 8 Comparison of RMSE and ROC value for J48, MLP and SMO

In addition, RMSE value for J48 is the lowest (i.e. 0.4584) compared to SMO and MLP. Performance matrix ROC for J48 shows the highest (ROC value approaching 1 is better), 0.707 compared to SMO, 0.661. This is shown in Fig. 8.

These results show similar findings to [12]. In their work, the accuracy for J48 10-folds cross validation was 63.60% and MLP was 62.65%. Furthermore, this work is comparable to [13], classification accuracy of SMO (63.7%), J48 (70.19%) dan MLP (70.64%), for 10-folds cross validation.

In addition, J48 was tuned to make it perform better by try and error approach. For example, as in Table 9, the value of *confidenceFactor* parameter is between 0.1 and 0.50. This value is manually changed and 0.1 found to be the best value. *binarySplits* parameter with TRUE value means it used a binary division on nominal attributes while building a tree.

TABLE IX
DEFAULT AND TUNED PARAMETES OF J48 ALGORITHM

Parameter	Default	Tuned Parameter
batchSize	100	100
binarySplits	FALSE	TRUE
collapseTree	TRUE	TRUE
confidenceFactor	0.25	0.1
debug	FALSE	FALSE
doNotCheckCapabalities	FALSE	FALSE
doNotMakeSplitPointActualValue	FALSE	FALSE
minNumObj	2	3
numDecimalPlaces	2	2
numFolds	3	3
reducedErrorPruning	FALSE	FALSE
saveInstancesData	FALSE	FALSE
seed	1	1
subtreeRaising	TRUE	FALSE
upruned	FALSE	FALSE
useLaplace	FALSE	FALSE
useMDLcorrection	TRUE	TRUE

TABLE X
J48 RESULTS AFTER PARAMETER TUNING

Testing		Time (s)	Acc (%)	RMSE	ROC
10-fold cross validation	Before	0.05	66.0651	0.4584	0.707
	After	0.22	66.1824	0.4596	0.695

Table 10 shows the comparison between before and after the parameter tuning. The results show an increase of 0.1173% in accuracy (i.e before the accuracy is 66.0651%, and then increase to 66.1824%). It can be shown from Fig. 9.

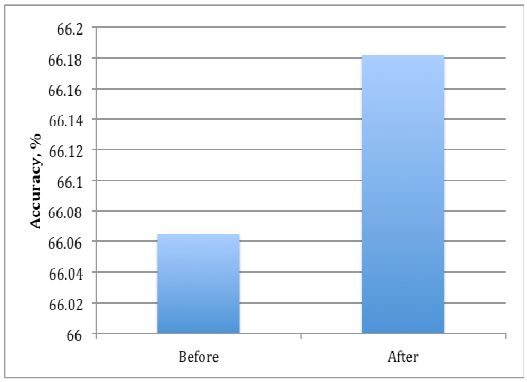

Fig. 9 J48 accuracy before and after parameter tuning

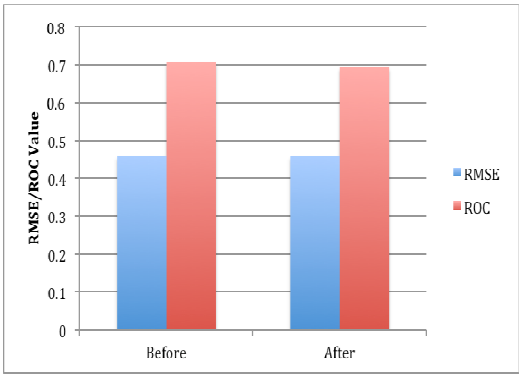

Fig. 10 RMSE and ROC for J48 before and after parameter tuning

In addition, the two measures, RMSE and ROC have shown insignificant difference in both cases, before and after tuning. This can be shown in Fig. 10.

J48 suits the problem of identifying the factors of getting employed; hence it is worthwhile to consider the rules generated by J48. These rules give an insight of the attributes that affect the employability of the students.

Rules derived from J48:

For WORKING class

1. IF age = 20-29 AND industrial internship = Yes AND faculty = Faculty of Economy and Management AND Marital Status = Not Other, THEN Class=WORKING
2. IF age = 20-29 AND industrial internship = Yes AND faculty = NOT Faculty of Economy and Management AND faculty = Faculty of Health Sciences, THEN Class=WORKING
3. IF age = 20-29 AND industrial internship = Yes AND faculty = NOT Faculty of Economy and Management AND faculty = NOT Faculty of Health Sciences AND faculty = Faculty of Engineering and Build Environment, THEN Class = WORKING
4. IF age = 20-29 AND industrial internship = Yes AND faculty = NOT Faculty of Economy and Management AND faculty = NOT Faculty of Health Sciences AND faculty = NOT Faculty of Engineering and Build Environment, AND faculty = Faculty of Information Science and Technology AND English skill = Satisfy, THEN Class = WORKING
5. IF age = 20-29 AND industrial internship = No AND field = Education AND involvement in curriculum activity = Not Applicable, THEN Class = WORKING
6. If age = NOT 20-29, THEN Class = WORKING

For NOT WORKING class

1. IF age = 20-29 AND industrial internship = Yes AND faculty = Faculty of Economy and Management AND Marital Status = Other, THEN Class = NOT WORKING
2. IF age = 20-29 AND industrial internship = Yes AND faculty = NOT Faculty of Economy and Management AND faculty = NOT Faculty of Health Sciences AND faculty = NOT Faculty of Engineering and Build Environment, AND faculty = Faculty of Information Science and Technology AND English skill = Not Satisfy, THEN Class = NOT WORKING
3. IF age = 20-29 AND industrial internship = Yes AND faculty = NOT Faculty of Economy and Management AND faculty = NOT Faculty of Health Sciences AND faculty = NOT Faculty of Engineering and Build Environment, AND faculty = Not Faculty of Information Science and Technology THEN Class = NOT WORKING
4. IF age = 20-29 AND industrial internship = No AND field = Education AND involvement in curriculum activity = Applicable, THEN Class = NOT WORKING

In this work, the generated rules show that the most influential attribute in classifying working or not working is the age attribute. For age = 20 - 29, some instances are working and some are not, but for age other than 20 - 29 (more than 29), the instances are working. In classifying the working class for age 20-29, the factors being considered are industrial internship, faculty, English skill and involvement in curriculum activity. Mean while, two other factors influencing the not working class, are marital status and field of study.

IV. CONCLUSIONS

In this work, data mining techniques were used to classify factors affecting graduates employability, particularly UKM. Three methods were used, i.e. J48, MLP and SOM. The results showed that J48 performed better compared to other techniques with 66.0651% and it increased to 66.1824% after the parameter tuning. This paper identified several factors affecting UKM graduate employability such as age, faculty, field of study, co-curriculum, marital status, industrial internship and English skill. Among these factors, attribute *age*, *industrial internship* and *faculty* contain the

most information and affect the final class, i.e. *employability status*.

ACKNOWLEDGMENT

We would like to thank University Kebangsaan Malaysia for granting this project under KRA-2017-003.

REFERENCES

[1] Morshidi Sirat, Chan Lean Heng, Munir Shuib, Shukran Abdul Rahman, Seri Rahayu Ahmad Kamil, and Jasvir Kaur Nachatar Singh, "Employability of graduates in Malaysia, Graduate Employability in Asia", UNESCO Bangkok: Asia and Pasific Regional Bureau for Education, pp. 30-37, 2012.

[2] (2017) Sistem Laporan Kajian Pengesahan Graduan Kementerian Pendidikan Tinggi website. [Online]. Available: http://graduan.mohe.gov.my/v/.

[3] M. S. Salleh and M. Z. Omar, "University-Industry Collaboration Models in Malaysia", *Procedia - Social and Behavioral Sciences*, 102, pp. 654 - 664, 2013.

[4] H. Mohd. Yusof, Ramlee Mustapha, Syed A. Malik Syed Mohamad, Seri Bunian, "Measurement Model of Employability Skills using Confirmatory Factor Analysis", *Procedia - Social and Behavioral Sciences*, 56, pp. 348 - 356, 2012.

[5] Suhaila Zainudin, Dalia Sami Jasim, and Azuraliza Abu Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction", *International Journal on Advanced Science Engineering Information Technology*, vol. 6, pp. 1148-1153, 2016.

[6] Alhamdi Mohammed Alshareef, Azuraliza Abu Bakar, Abdul Razak Hamdan, Sharifah Mastura Syed Abdullah, Mohammed Alweshah, "A Case-based Reasoning Approach for Pattern Detection in Malaysia Rainfall Data", International Journal of Big Data Intelligence, vol. 2, no. 4, 2015.

[7] Zulaiha Ali Othman, Noraini Ismail, Abdul Razak Hamdan, Mahmoud Ahmed Sammour, "Klang Valley Rainfall Forecasting Model using Time Series Data Mining Technique", *Journal of Theoretical and Applied Information Technology*, vol. 92 No. 2, , pp. 372-379, 31st October 2016.

[8] Fatiha Elghibari, Rachid Elouahbi and Fatima El Khoukhi, "Data Mining for Detecting E-learning Courses Anomalies: An Application of Decision Tree Algorithm", *International Journal on Advanced Science Engineering Information Technology*, vol. 8, pp. 980-987, 2018.

[9] B. Jantawan, and C. F. Tsai, "The Application of Data Mining to Build Classification Model for Predicting Graduate Employment", *International Journal of Computer Science and Information Security*, vol. 11(10), pp. 1–8, 2013.

[10] M. A. Sapaat, A. Mustapha, J. Ahmad, K. Chamili, and R. Muhamad, "A Data Mining Approach to Construct Graduates Employability Model in Malaysia", *International Journal on New Computer Architectures and Their Applications*, vol. 1(4), pp. 1086-1098, 2011.

[11] B. Jantawan, and C. F. Tsai, "A Classification Model on Graduate Employability Using Bayesian Approaches: A Comparison", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2(6), pp. 4584-4588, 2014.

[12] M. T. R. A Aziz and Y. Yusof, "Graduates Employment Classification Using Data Mining Approach" in *Proc. of the International Conference on Applied Science and Technology*, 2016, p. 1–7.

[13] T. Mishra, D. Kumar, and S. Gupta, "Students' Employability Prediction Model through Data Mining", *International Journal of Applied Engineering Research*, vol. 11(4), pp. 2275-2282, 2016.

[14] G. Suganthi and M. V. Ashok, "Predicting Employability of Students using Data Mining Approach", International Journal of Information Research and Review, vol. 04 (02), pp. 3798-3801, February 2017.

[15] J. Pei Han and M. Kamber, *Data Mining: Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, 3rd Ed. Amsterdam: Elsevier, 2011.

[16] V. Manjula, M. D. U. Sankari, A. P. Nayaki and R. Saranya, "Predicting Employability of a Student In R Programming", *International Journal of Advanced Research Trends in Enginerring and Technology*, vol. 4(13), pp. 35-39, 2017.

[17] L. Rokach, and O. Maimon, Data Mining with Decision Trees: Theory and Applications. 2nd. Edition, Volume 81 of Series in Machine Perception and Artificial Intelligence, World Scientific, 2014.

[18] A. Krenker, J. Bešter, and A. Kos, *Introduction to the Artificial Neural Networks, Artificial Neural Networks - Methodological Advances and Biomedical Applications*, Prof. Kenji Suzuki: InTech, 2011.

[19] C. M. Bishop, C.M., *Neural Networks for Pattern Recognition*, Advanced Texts in Econometrics, United States:Oxford University Press, 1995.