https://doi.org/10.17576/apjitm-2018-0702-10

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 7 No. 2, December 2018 : 131 - 145
e-ISSN: 2289-2192

CLASSIFICATION MODELS FOR HIGHER LEARNING SCHOLARSHIP AWARD DECISIONS

WIRAWATI DEWI AHMAD
AZURALIZA ABU BAKAR

ABSTRACT

Scholarship is a financial facility given to eligible students to extend Higher Education. Limited funding sources with the growing number of applicants force the Government to find solutions to help speed up and facilitate the selection of eligible students and then adopt a systematic approach for this purpose. In this study, a data mining approach was used to propose a classification model of scholarship award result determination. A dataset of successful and unsuccessful applicants was taken and processed as training data and testing data used in the modelling process. Five algorithms were employed to develop a classification model in determining the award of the scholarship, namely J48, SVM, NB, ANN and RT algorithms. Each model was evaluated using technical evaluation metric, such contingency table metrics, and accuracy, precision, and recall measures. As a result, the best models were classified into two different categories: The best model classified for 'Eligible' status, and the best model classified for 'Not Eligible' status. The knowledge obtained from the rules-based model was evaluated through knowledge analysis conducted by technical and domain experts. This study found that the classification model from SVM algorithm provided the best result with 86.45% accuracy to correctly classify 'Eligible' status of candidates, while RT was the weakest model with the lowest accuracy rate of for this purpose, with only 82.9% accuracy. The model that had the highest accuracy rate for 'Not Eligible' status of scholarship offered was NB model, whereas SVM model was the weakest model to classify 'Not Eligible' status. In addition, the knowledge analysis of the decision tree model was also made and found that some new information derived from the acquisition of this research information may help the stakeholders in making new policies and scholarship programs in the future.

Keywords: scholarship award; classification model; knowledge discovery.

INTRODUCTION

Higher education scholarships are financial aid provided to students at higher education institutes to help them on the spending requirements during their study program. Scholarship award decision is an important process to ensure that financial aid provided to students and scholarship offers can be done efficiently. The inefficiency of higher learning scholarship management will prevent potential students from continuing their learning at a higher education level.

Due to the growing number of applications each year, scholarship providers have been forced to seek new strategies in dealing with the challenges of managing the most qualified candidates in a short period of time. One of the approaches used to manage widely used data is the study of data analytics strategy. Data analytics is a major process in a big data project. In big data initiatives, there are two main processes, namely data management and analytical data. Analytical data encompasses two major processes, namely the process of modelling and the interpretation of knowledge gathered (Gandomi, & Haider 2015). In this study, we explore the classification task on the selection of appropriate candidates for scholarship offer using five different classification algorithms namely J48, Support Vector Machine (SVM), Artificial Neuron Network (ANN), Naïve Bayes (NB), and Random Tree (RT). Each model will be

trained and tested using a set of data taken from actual scholarship application programs from 2013 to 2016.

This study focuses on the development of a single modelling framework for scholarship decision with a data mining approach using five different algorithms. These models will be analyzed and compared to find the best classification model to determine scholarship candidate.

This paper is organized into five sections. We begin this paper with the background on scholarship offer classification problems. The second part is the literature review which focuses on current studies on classification model in various areas. We highlight the studies on scholarship offer classification problems and the techniques that have been employed. We also focus on the methods of evaluation metrics used to evaluate and compare the performance of each model.

The third section describes the methodology of this paper, including the experiment design. The fourth section compares different classifiers; tools which support the techniques and the result of experiments. Finally, the fifth section concludes the paper.

RELATED WORK

KNOWLEDGE DISCOVERY

Knowledge discovery is a process of information acquisition through a systematic approach using machine learning methods to find useful knowledge of existing data. It involves four main processes which are pre-processing data, data mining process, model testing and evaluation, and knowledge analysis (Koturwar, Girase & Debajyoti 2014). Every process should be systematic and precise so that the output at the final stage is useful and correct. One important phase in knowledge discovery is data mining. Data mining is an approach that uses a variety of techniques and algorithms to find hidden patterns, relationships between attributes, and knowledge acquisition that may be useful based on existing data (Jiawei Han 2006).

The knowledge acquisition through this process is then used in a strategic decision-making process in the future. In the context of a scholarship provider organization, obtaining information using this process is capable of supporting the management in making decisions for the administration of future scholarship offers. For example, knowledge discovery of data on scholarship offers can provide the organization with the pattern of scholarship offers in terms of the characteristics of successful candidates, the characteristics of the rejected candidates, the fields offered by scholarships, and others. In addition, knowledge discovery on the financial loan repayment data of students can help the organization view whether students successfully make repayment instalments consistently or not, as well as character traits of students who fail to complete their studies, and so on.

SCHOLARSHIPS AWARD CLASSIFICATION

Apart from the above problem areas, studies in developing classification and prediction models in the process of awarding higher educational scholarships are also an important aspect of the study to ensure that financial aid is efficient enough to facilitate the production of experts and the betterment of the education sector of a developing country. The classification and prediction studies include a single classification model that recommends appropriate algorithms using the previous set of scholarship packages. Through this study, the selection of multiple features depends on the data set stored by the scholarship provider. On behalf of the grantee, a systematic mechanism is needed for this purpose as the expansion of large size of application and existing data and the awarding process should be conducted within a short period of time (Alhassan & Lawal 2015; Azuraliza & Arshad 2013; Raharja 2014).

According to the study of Azuraliza & Arshad (2013), granting scholarship to unsuitable candidates would create the problem of insufficient funds for the provision of higher education for the less fortunate eligible students, thus denying their chances of pursuing higher education. It is also seen as a deterrent to the primary goal of funding the higher education sector to produce experts in science and technology in the future. In this study, a systematic design for a higher scholarship offer using a decision tree method was proposed. Four main factors that influence the output of the study are as follows: the class label, a nominal-type dependent to determine the label on the data set object; predictors, independent variables or attributes that represent features of the object; training data set that contains the values of the two components used above to determine a suitable class based on a predictor; and testing data set that contains new data to be classified by the model created, where the saccuracy of the classification would be evaluated. This study found that the C4.5 algorithm provided high accuracy in line with increased data training size (90% to 92.31%). This proves that the scholarship classification model using the decision tree method is influenced by the size of the training data used.

The study of Azuraliza & Arshad (2013) also proposed the design of classification model of higher education scholarship award model based on existing scholarship data. This study conducted comparative experiments between the performance of the best classifier models between the two algorithms; rough set and decision tree. Fifteen attributes were chosen to be predictors with class label 'Grant' or 'Not Granted'. The study found that both classifiers show comparative results with certain advantages and drawbacks. The best average accuracy was shown by the rough set model with 90% average compared to the J48 model with 83% average.

A study by Raharja (2014) also used a data mining technique for scholarship disbursement problems. In this study, 800 student records were taken with 9 predictors and 1 attribute namely 'score' as class label to develop a decision tree classification model using the decision tree method. Each leaf would form rules that would be used to determine appropriate candidates for scholarship offers, whether an applicant shall be successful or unsuccessful in his or her application using the applicant data. The tree algorithm used in designing this system has proven to be effective and efficient.

The increasing educational cost is also a factor in the educational scholarship needs of a country. This also affects low-income parents to provide education for their children. Therefore, it is very important to ensure that scholarships are awarded to students in a systematic way (Alhassan, & Lawal 2015; Azuraliza, & Arshad 2013; Raharja 2014; Tun, & Aye 2014).

METHODOLOGY

The quantitative approach in data mining studies has been widely used in determining the performance and capabilities of the classification models (Kaiwen 2018). The methodology used in the present paper is shown in Figure 1. There are four major processes involved, namely data pre-processing (actual data from scholarship recipients), modelling, model testing and evaluation, and knowledge analysis.

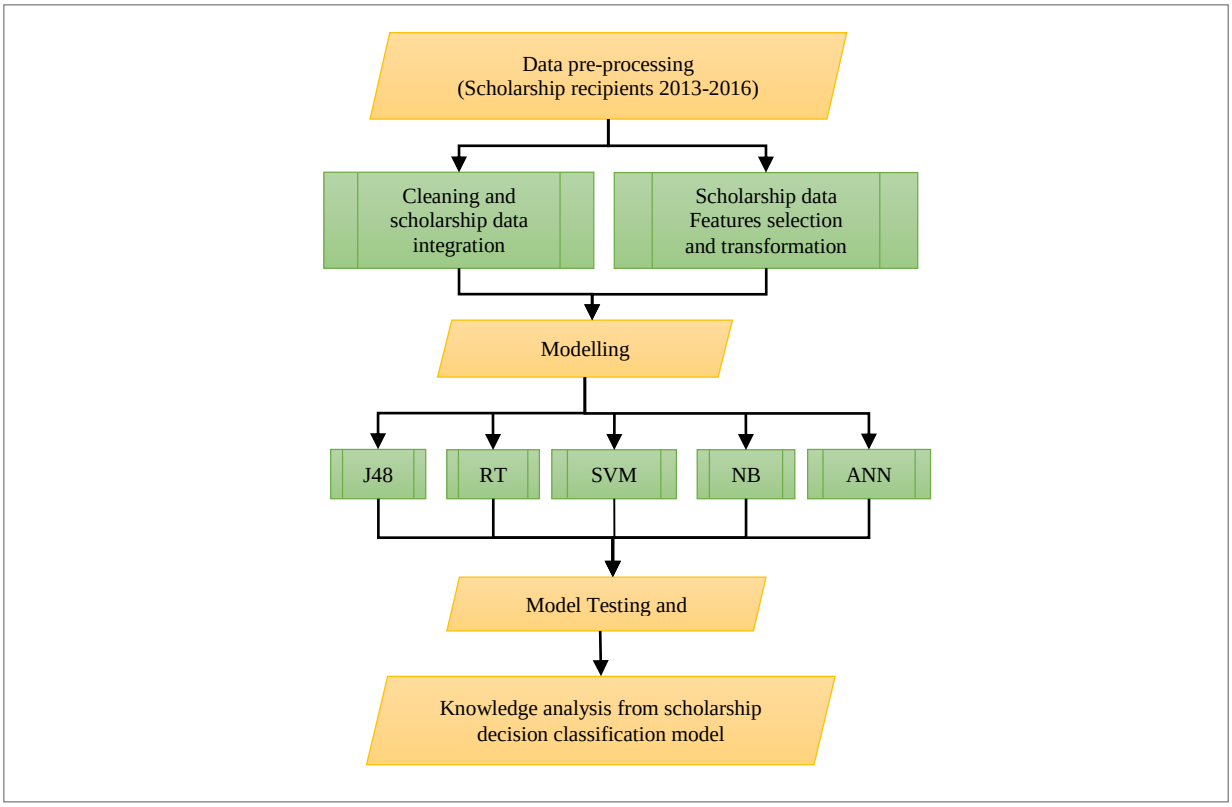

FIGURE 1. Modelling process

DATA PRE-PROCESSING

The data preparation process was conducted to provide a sufficient and completed set of data for modelling purposes. The two main processes undertaken in this stage were the process of clearing and integration, and the process of selecting features and data transformation.

From an actual scholarship recipients' database, raw data was obtained and analyzed to see the type and size of the attributes and their importance. A total of 87,000 data was recorded with over 100 attributes. After a preliminary study of on the data, several steps had been taken to ensure the obtainment of clean and complete data such as eliminating noise in data, completing incomplete data, eliminating outliers, and deleting inconsistent data. Data noise is usually caused by poorly managed database. Records that did not store the correct values for the attribute type and size were deleted. It is most likely that actual databases are sometimes used to carry out system entry tests that receive meaningful test data. A record that has no value for the final decision of a committee is defined as incomplete data and eliminated from the net data set. Similarly, a set of data that is of no value to the upload document indicates that the applicant is not committed to applying for a scholarship.

An important attribute that determines the candidate's outcome is the 'Eligible' or 'Not Eligible' status, thus, data that stores a value other than 1 or 2 was also eliminated. This odd value was due to a system testing process or an applicant who did not resubmit a complete document during the application process. This data record was eliminated because it is considered unimportant in this classification experiment modelling process. Inconsistent data records were also found during the data analysis process and were eliminated, such as the applicant's age exceeds 64 years, the applicant's age is less than 22 years, non-existent identity card numbers, or a record with multiple application. As a result of the above process, a total of 57000 net records was successfully provided. Attributes can be categorized into three

categories, which are deleted attribute, attributes that are retained by 11 attributes, and a new attribute created to replace the attributes deleted by 4 attributes.

In this study, the initial set of data was processed into a new data set with attributes of the correct type and value for the purpose of modelling. This new set of data was then analyzed using feature selection techniques and transformed into appropriate forms to meet modelling algorithm requirements. Feature selection is a step in providing data for modelling purposes. Feature selection is able to provide guidance and suggestions on features that impact on the objectives of the study (Daud et al. 2017). Additionally, the selection of these features also helps reduce the cost of time and resources needed to analyze large data. Unpredictable features can be eliminated in the early stages, thus increasing the efficiency during the modelling process. In this study, feature selection was done with two techniques, expert views and technical analysis. During the process of data clearing and integration, expert views were taken to select essential attributes based on expert knowledge, while technical analysis was also made as an extension of expert opinions. For the technical analysis of data feature selection, two (2) common feature selection methods were used in data mining approach: Info Gain and Gain Ratio. The results are shown in table 1.

TABLE 1. Result for features selection using Info Gain and Gain Ratio Techniques

	Name of Attribute	
Rank	Info Gain	Gain Ratio
1	Age_range	Marital_status
2	Yearofgraduated_range	Age_range
3	University	fieldofstudy
4	Marital_status	Yearofgraduated_range
5	Approval_status	Approval_status
6	Fieldofstudy	Empolyment_status
7	Empolyment_status	university
8	Program_structure	Program_structure
9	Stateoforigin	religion
10	Gender	gender
12	Religion	Disable_status
13	Levelofstudy	Stateoforigin
14	Disable_status	levelofstudy
15	Sponsorship	Sponsorship

After the feature selection process, the data transformation process succeeded in establishing a set of training and test data that fulfils the requirements of the classifier algorithms by identifying and altering the label data sets into nominal form. The final result of the transformation process is shown in Table 2. In this study, attributes worthy of committees or deemed qualified are identified as labelled data classes with a value of either 1 (‘Eligible’) or 2 (‘Not Eligible’).

TABLE 2. List of Attributes

Attribute name	Definition
age range	1 = 20 to 25 years 2 = 25 to 29 years 3 = 30 years to 34 years 4 = 35 years to 39 years 5 = 40 years to 49 years 6 = 50 years to 64 years
gender	1 = male 2 = female
marital status	1 = single 2 = married 3 = Others
disabilities	1=Yes 2= No
sponsorship status	0 = Yes 1=No
work status	0=Employed 1=Unemployed
approval status	1=Eligible 2=Not Eligible
field of study	1=Social science 2=Science & Technology
program structure	1=Course work 2=Mix mode 3=Research
graduated years of study	1=2014 and 2015 2=2013 3=2012 4=2011 5=between 2006 and 2010 6=between 1990 and 1999 7=Others
Level of study	8=Master 13=PhD

The complete and correct data preparation of the classification modelling process is important to ensure that the developed model is accurate (Aruna & Nandakishore 2011; Jiawei Han 2006; Yang et al. 2017). This complete set of data was used as input to the classification process of the classifier, so the set data that satisfies the needs of the classifier algorithms was also determined.

MODELLING

In this study, to ensure diversity exists in the classification design, two approaches were applied: ensuring the diversity of decision boundaries by diversifying the training data sets and using the diversity of classifier by applying several different classification algorithms. Using the set data provided, this classification task involves developing models for determining scholarship award candidate using five algorithms which are J48, Support Vector Machine (SVM), Naïve

Bayes (NB), Artificial Neuron Network (ANN) and Random Tree (RT). The modelling process uses a 10-fold cross-validation technique and a percentage split of training and testing data set of 90 to 10 percent. Each algorithm will produce 11 different classifier models and overall this experiment would produce 55 single classification models which was then evaluated using chosen appropriate evaluation metrics.

The decision tree algorithm is compatible with both types of data whether numerical or nominal. In addition to whatever data size, higher accuracy in the decision tree classification technique always produces good performance of a classification model. The decision tree can handle large quantities of input data, such as text with numerical or nominal numerical data. It is a supervised learning approach that has the ability to extract information from a large amount of data based on rules or decision tree. In this study, two basic algorithms under this technique were used which are J48 and Random Tree (RT). J48 and RT were selected as they are easily processed and produce easy-to-understand output, such as rules of if-then-else, which is easy to interpret.

J48 decision tree was developed by Ross Quinlan in 1993. It is a tree-shaped classifier used to represent a model with an attribute relationship in a data set. Different decision tree algorithms have many advantages over various learning algorithms such as sound noise, low cost calculation for model generation, and the capability for different properties and modules (Aruna & Nandakishore 2011; Kaur & Gangwar 2017). The percentage of records that are properly classified will determine whether the model is performing well or not. Rules in the form of IF-THEN can be extracted from this model. Each path from root to leaf node can be written as a rule. The previous rules (part IF) are formed by combining the separation criteria along the path provided by the AND connection. Leaf nodes containing class predictions form the rules that occur (THEN section).

RT is one of the decision tree categories of algorithm. In the main approaches, the 'K' attribute is randomly selected to classify the data. It does not contain any trimming techniques to minimize errors. Random tree algorithm has the option of estimating class probabilities for classification (Wang et al. 2011).

SVM is a machine learning technique introduced by Boser, Guyon & Vapnik and has been widely used in various classification problems to date(Abdullah Khalid & Omar 2012; Ayub & Karnalim 2017; Hamsagayathri & Sampath 2017)SVM is generally seen as a hyper-plane that separates the object (the eye) belonging to the class (positive object) rather than the class (negative object). This separation is performed by the SVM algorithm during the learning step where the hyper-plane is obtained to divide the positive and negative objects with maximum margins. Margins show space from hyper-plane to the nearest positive and negative object.

NB is a classification technique based on the probability theory of a data or an attribute with data frequency or the same attribute in a data set (Goyal, Thakur & Chowdhury 2016). The NB model is easy to use for very large data sets. In summary, NB assumes that the value of a particular feature is not related to the presence or absence of any other characteristics, given in the class variable.

ANN is a classification model represented by interconnected nodes. It can be seen as a rounded node represented as an artificial neuron that represents the output of one neuron to another input (Statnikov et al. 2005). The ANN model works in showing hidden links in historical data, thereby facilitating classification and prediction of new data. The ANN model is precise enough to make clear and relevant decisions about existing data patterns. However, this algorithm takes longer than other classification algorithms as it builds interconnected nodes.

TESTING AND EVALUATION

The evaluation metric is used as a measure that focuses on the model's prediction capabilities (Rathore, & Gupta 2014). In this study, evaluation that would be used are parameters from the contingency table; True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), and parameters extended from this table; Accuracy, Precision, Recall and F-measure as shown in Table 3 and formula in (1) (2) and (3).

TABLE 3. Classification Model For Scholarship Award Decision Confusion Matrix

		Prediction	
		Positive	**Negative**
Actual	**Positive**	(a) **TP –** Records classified 'Eligible' and correct	(b) FP – Records classified 'Not Eligible' but incorrect
	Negative	(c) **FN –** Records classified 'Eligible' and Incorrect	(d) **TN –** Records classified 'Not Eligible' and correct

The calculation for accuracy (1), precision (2) and recall (3) are as below:

$$accuracy = \frac{TP+TN}{n} \ldots\ldots\ldots\ldots\ldots (1)$$

n= total records of testing data

$$precision = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots\ldots (2)$$

$$recall = \frac{TP}{TP+FN} \qquad \ldots\ldots\ldots\ldots (3)$$

By observing the output of these parameters generated by each model, the best and weakest model capabilities can be identified. The best model is the model that achieves high accuracy and less error value. In addition to finding the best and weakest model in classification higher education scholarship decisions, this study also conducts knowledge analysis in finding useful information that may be derived from the developed model.

KNOWLEDGE ANALYSIS

In the process of obtaining information from the models developed, this study uses the expertise of the officers involved in analysing each rule of the obtained rules combined with technical expert in data mining as shown in FIGURE 2.

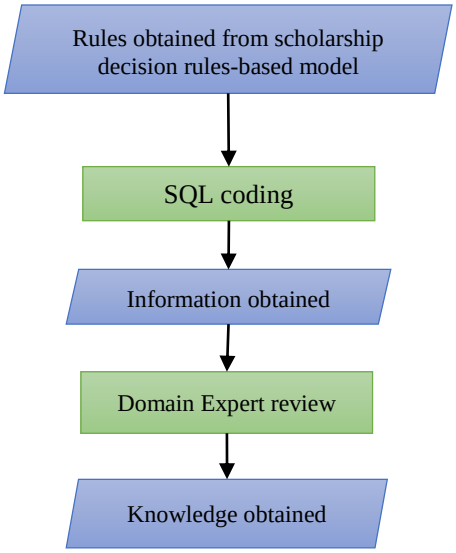

FIGURE 2. Knowledge Analysis Process for the Higher Learning Scholarship Award Classification Model

In this study, the knowledge analysis process was carried out by building a programming code using Structured Query Language (SQL) based on the rules obtained from rules-based scholarship J48 model. The acquisition of information from this process would then be reviewed by the scholarship domain expert who has been involved in scholarship management for more than two years. The domain expert functions are to verify the significant relation between attributes and the mined patterns based on their knowledge and experiences (Zainudin et al. 2015). To ensure that the rules are easily understood by the expert, a code translation was implemented. As result from the expert review, useful knowledge obtained from the rules generated by the model was collected to support the policy maker in making strategic decisions in the future.

RESULT AND DISCUSSION

This section describes the output of each experimental outputs. The decision of the classification model according to the best position to the weakest will be assessed using the assessment metrics specified in the preceding section. This analysis of the decision was made technically or with domain expert knowledge to obtain the information constructed by the developed classification models. Two main objectives of this analysis are to identify model effectiveness and model performance, as discussed in the section below.

MODEL EFFECTIVENESS

These classification models are built to predict two label classes, ‘Eligible’ or ‘Not Eligible’. Based on contingency tables, the rates for the best PB, PS, NB, and NS models per algorithm are analysed. The analysis of contingency tables shows that SVM is the best model for classification for ‘Eligible’ records, with the highest value for TP (7,537) when using 90% split

of training and testing data. This may occur because SVM is very smart in the classification of complete and small-sized data sets. The set of training and test data used in the modelling phase has been through a data-preparation process that ensures the data set used in the modelling phase is sufficient.

FN value shows how the classification model wrongly predict for negative class ('Not Eligible') examples. The result shows that SVM model has the highest wrongly classified examples for negative class ('Not Eligible') with 1,157 instances. This is likely since this model generates lots of rules for positive examples rather than negative examples. Although the SVM model has a high TP rate compared to other models, it has a weakness in classification negative examples. For classification models with two label classes, a good classification model should be able to classify well for both identified label classes.

For TN value, the Naïve Bayes model is the best model in classifying 'Not Eligible' records correctly with highest number of 192 instances with a small percentage of 13.7% from the overall actual number of candidates with 'Not Eligible' status. This rate is considered small and most probably because of this algorithm inability to classify for a small label class. Other models also show poor performance when classifying negative ('Not Eligible') candidates. This may be because the number of records for the 'Not Eligible' class label is small compared to the 'Eligible' class label, causing many of these category rules to be ignored. According to a study by (Chawla et al., 2004), this decision is often the result of unbalanced data, where the rules for the little class are largely ignored. The contingency table summary result of all model average is shown in Table 4.

TABLE 4. Model effectiveness performance

Model/ Metrics	J48	SVM	NB	ANN	RT
TP	7,185	7,537	7,320	7,256	6,873
FP	560	208	425	489	872
FN	1,081	1,157	1,063	1,115	1,081
TN	174	98	192	140	170

MODEL PERFORMANCE

In addition to the contingency tables, model accuracy, precision, recall, and F-measure parameters are also observed. The accuracy of a model can be measured by comparing the actual results with predicted results generated by the model (Wang et al. 2011). Models that give predictive results with high accuracy show that the model is able to perform the best predictive task. Therefore, this percentage of accuracy is very important in choosing the best model developed. The overall result on these parameters is shown in Table 5.

TABLE 5. Model performance result

Model/ Metric	J48	NB	SVM	ANN	RT
Acc	83.6%	85.46%	86.45%	85.46%	82.9%
Precision	0.799	0.861	0.869	0.855	0.835
Recall	0.836	0.761	0.869	0.806	0.822
F-measure	0.814	0.827	0.787	0.797	0.777

Based on these parameters, the SVM model produces the highest accuracy of 86.45% compared to ANN model and NB model (85.46%), J48 model (83.6%) and RT model with 82.9%. The advantage of classification model using SVM algorithms for small data sizes is the decisive factor for SVM to be the best model in this experiment. This is because the best model of this algorithm uses 90% practice data fractionation, whereas the training data is only 10%. This demonstrates that the increased accuracy is parallel with the large size of training data used during the modelling process. This is supported by a study conducted by (Afram et al. 2017) which stated that the SVM algorithm is very effective in classifying small-sized data. Figure 3 shows the results of performance parameters for the best model of each algorithm generated by this experiment.

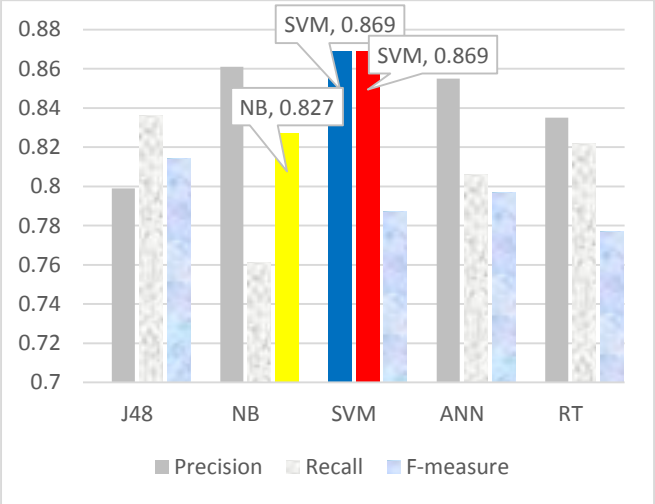

FIGURE 3. Illustration of model performance result comparison

To measure the performance of a classification model, the error rate is also viewed and compared. A high error on a model is not a good indicator of the performance of a classification and prediction model (Thomas, & Vidal 2017) . In examining the error of a model, this study focused on the major error rate parameters of Mean Absolute Error (MER) and Root Mean Squared Error (RMSE). SVM model shows the lowest rate for MER with 0.131 while J48 model shows the highest MER rate with 0.2348. In addition, NB model shows the best rate for RMSE which is 0.3292 with RT model showing the highest RMSE rate as shown in Table 6.

TABLE 6. MER and RMSE measures

Measure/Model	J48	SVM	NB	ANN	RT
Mean absolute error (MAE)	0.2348	0.131	0.215	0.1606	0.2106
Root mean squared error (RMSE)	0.3375	0.3619	0.3292	0.3678	0.3983

According to the study by (Veerasamy et al. 2011), the MER and RMSE rates between 0.5 and 1.0 reflect the classification model's ability to predict the class accurately. For a good classification model, the value of MER and RMSE should be low between 0.5 and 0.3. In conclusion, all the classification models developed in this study are considered as good because the RMSE error ranges of these models are between 0.3292 and 0.4214 according to the definition of study by (Veerasamy et al. 2011) as shown in FIGURE 4.

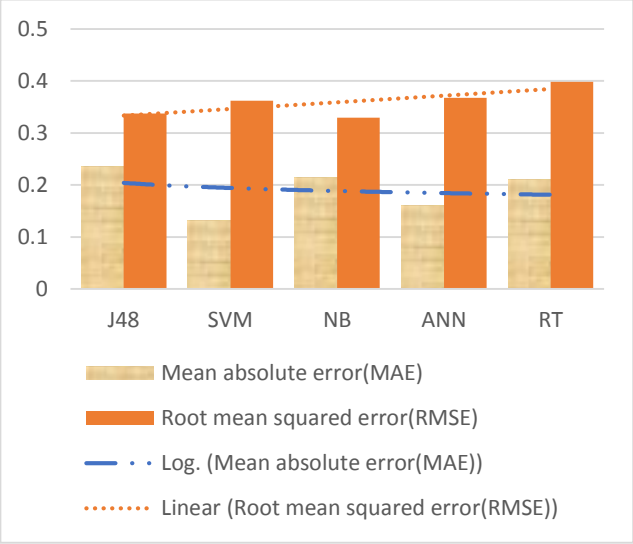

FIGURE 4. Error performance comparison

KNOWLEDGE ANALYSIS

Knowledge analysis is the last step in the cycle of data mining process. In this study, the rules obtained from rules-based J48 model were monitored and analysed to find knowledge that may be useful to stakeholders. Although the experimental result indicates that the scholarship classification model SVM shows a better result than J48 model (86.45% and 83.6%), J48 model is chosen for knowledge analysis because of the following reasons:

1. It is observed that the J48 model is still good and acceptable since it has high accuracy (83.6%), and
2. The rules-based results given by the J48 model are more presentable and easier to be understood by domain expert (Roy, & Garg 2017).

SQL coding method and expert reviews were used in achieving this goal. The expert knowledge was obtained from an experienced officer from a higher education scholarship management office and combined with a technical expert in data mining to help seek the best knowledge from the rules generated. The rules generated by this algorithm were observed one at a time and were recorded by class; class 1 label ‘Eligible’ and class 2 label ‘Not Eligible’. The rules generated by this model are 2771 rules with the tree size of 3265. The rules which successfully classified more than 15 candidates are categorized as important rules from a total of 79 rules. The following are examples of hidden information for successful candidates generated by this model as shown in Table 7.

TABLE 7. Information gathered from rules obtained examples

Rules 1 & 2	Information
age_range = 2 AND level_of_study = 8 AND university = M0100103 AND year_of_graduation = 1: 1 (1.0) OR year_of_graduation = 2: 2 (2.0) OR year_of_graduation = 3: 1 (4.0/1.0) OR year_of_graduation = 4 *age_range = 2 AND level_of_study = 8 AND university = M0100103 AND year_of_graduation = 5 AND sponsorship_status = 1 AND prog_structure = 1*	Graduates from 2013 are most widely offered scholarships and are found to have the most number of studies in *M0100103* with 408 candidates. This indicates that continuing studies after three years of completing a bachelor's degree can be seen from these rules. From these, 65% were from working groups and 35% were non-working groups. It can be concluded that Malaysian graduates choose to work after completing a Bachelor's Degree and resume studies after having a stable job. Additionally, a group of scholarship graduates from 1990 to 1999 is found as the highest number of studies in UPM with age ranging from 40 to 49 years. It also shows that UPM became the university preferred by the student group.
age_range = 2 AND level_of_study = 8 AND university = M0100401 AND gender = 2 AND year_of_graduation = 5 AND field_of_study = 1 AND sponsorship_status = 0	At *M0100401*, a total of 527 candidates who are pursuing studies in the field of social science / literature comprise candidates aged between 25-29 years at the Master's degree. This group has never gained any sponsorship during Bachelor's degree studies. The group is also found to be working and part-time studying.
age_range = 2 AND level_of_study = 8 AND university = M0100101 AND religion = 1 AND marital_status = 2 AND sponsorship_status = 0 AND year_of_graduation = 5 AND gender = 1	In *M0100101*, the results of the model-led rules found that 44% of candidates have been awarded scholarships amongst graduates from various fields from 2006 to 2010, which is 44% with 78% of the candidates are single. From the rules, it also found that 93% of the candidates had a sponsored record during Bachelor's degree studies compared to 17% who never got any sponsorship

The acquisition of information from the acquired rules as in Table 7, may help policy makers and universities work together to improve the strategy in organizing higher education programmes and higher education scholarships programmes in the future. Higher education programmes that meet the needs of candidates will attract more students to pursue higher education while the support of a suitable scholarship programme will increase the number of students in higher education in the future. This is in line with the policy of the nation's higher education policy to create more qualified researchers in the future.

FUTURE WORK AND CONCLUSION

This study found that the SVM classification model, with highest accuracy rate 86.9%, is the best model in classifying the results of higher education scholarships compared to the models of J48, NB, ANN, and RT algorithms. However, the NB model demonstrates the best ability to classify unqualified candidates. The rules generated by the J48 model also provide many useful new knowledges to stakeholders. The data mining approach in this study may reveal thousands of information patterns from the experiments conducted. However, most of the patterns found

may not appeal to the organization, either because it represents a common knowledge or it may not be something new.

With the development of this scholarship decision classification model, the efficacy of higher education scholarship offers can be enhanced by time saving and efficient selection of candidates. New knowledge gained from the knowledge analysis process may support policy makers to determine the appropriate conditions for scholarship programmes in the future. On the part of the University, the knowledge gained from this process may be the basis for developing a special higher education programme for working groups to attract them to pursue their study to higher level of education in the future. This will certainly contribute to the country's productivity in producing researchers and skilled people for the national development.

Some challenges still relate to the development of techniques in assessing the extent of the importance of the patterns found, particularly in relation to the needs of the organization. The use of interesting measures or user-specific constraints to guide the discovery process and reduce search space is another active field of research.

ACKNOWLEDGMENT

We would like to thank Ministry of Higher Education for their support and encouragement. This project is supported by Research University Grand Challenge Grant CP-2017-15/1.

REFERENCES

Abdullah, S. N. H. S., Khalid, M. & Omar, K. 2012. Performance Comparison of License Plate Recognition System Using Multi- *10*(2011), 53–62.

Afram, A., Janabi-Sharifi, F., Fung, A. S. & Raahemifar, K. 2017. Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system. *Energy and Buildings*, *141*, 96–113. doi:10.1016/j.enbuild.2017.02.012.

Alhassan, J. K. & Lawal, S. A. 2015. Using Data Mining Technique for Scholarship Disbursement *9*(7), 1511–1514.

Aruna, S. & Nandakishore, L. V. 2011. K NOWLEDGE B ASED A NALYSIS OF V ARIOUS S TATISTICAL T OOLS IN D ETECTING B REAST 37–45. doi:10.5121/csit.2011.1205.

Ayub, M. & Karnalim, O. 2017. Predicting outcomes in introductory programming using J48 classification. *World Transactions on Engineering and Technology Education*, *15*(2), 132–136.

Azuraliza & Arshad, A. 2013. Rough Set and Decision Tree Model for Determining Scholarship Award Qualification *12*(April), 65–70. doi:10.4156/rnis.vol12.11.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F. & Alowibdi, J. S. 2017. Predicting Student Performance using Advanced Learning Analytics. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 415–421. doi:10.1145/3041021.3054164.

Gandomi, A. & Haider, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. doi:10.1016/j.ijinfomgt.2014.10.007.

Goyal, A., Thakur, S. & Chowdhury, R. 2016. Using Ensemble Learning and Association Rules to Help Car Buyers Make Informed Choices. *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies - BDAW '16*, 1–5. doi:10.1145/3010089.3010093.

Hamsagayathri, P. & Sampath, P. 2017. Decision Tree Classifiers for Classification of Breast Cancer. *International Journal of Current Pharmaceutical Research*, *9*(2), 31. doi:10.22159/ijcpr.2017v9i1.17377.

Jiawei Han, M. K. 2006. *Data Mining Concept and Techniques*.

Kaiwen, W. 2018. A Quantitative Analysis Method of Ideological and Political Instructors Work Based on Data Mining. doi:10.1109/ICITBS.2018.00069.

Kaur, R. & Gangwar, R. 2017. A Review on Naive Baye ' s (NB) , J48 and K-Means Based Mining Algorithms for Medical Data Mining.

Koturwar, P., Girase, S. and & Debajyoti, M. 2014. A Survey of Classification Techniques in the Area of Big Data. *International Journal of Advance Foundation and Research in Compute*, *1*(11), 1–7.

Raharja, Y. P. 2014. Rancang Bangun Sistem Rekomendasi Beasiswa Menggunakan ALgoritma Klasifikasi C4.5 pada Universitas Dian Nuswantoro. *Undinus*, 1–4. Retrieved from http://eprints.dinus.ac.id/13408/.

Rathore, S. S. & Gupta, A. 2014. A comparative study of feature-ranking and feature-subset selection techniques for improved fault prediction. *Proceedings of the 7th India Software Engineering Conference on - ISEC '14*. doi:10.1145/2590748.2590755.

Roy, S. & Garg, A. 2017. Analyzing performance of students by using data mining techniques a literature survey. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, 130–133. doi:10.1109/UPCON.2017.8251035.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D. & Levy, S. 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631–643. doi:10.1093/bioinformatics/bti033.

Thomas, R. W. & Vidal, J. M. 2017. Toward detecting accidents with already available passive traffic information. *2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017*, 1–4. doi:10.1109/CCWC.2017.7868428.

Tun, K. T. & Aye, A. M. 2014. Selection of Appropriate Candidates for Scholarship Application Form using KNN Algorithm. *International Journal of Scientific Engineering and Technology Research*, *3*(6), 1019–1026.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P. & Agrawal, R. K. 2011. Validation of QSAR Models - Strategies and Importance. *International Journal of Drug Design and Disocovery*, *2*(3), 511–519. doi:10.1016/j.febslet.2005.06.031.

Wang, G., Hao, J., Ma, J. & Jiang, H. 2011. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230. doi:10.1016/j.eswa.2010.06.048.

Yang, X., Lo, D., Xia, X. & Sun, J. 2017. TLEL: A two-layer ensemble learning approach for just-in-time defect prediction. *Information and Software Technology*, *87*, 206–220. doi:10.1016/j.infsof.2017.03.007.

Zainudin, S., Hamdan, A. R., Mohamed, Z. A., Idayu, N. O. R. & Shukri, A. 2015. Wound Up Malaysian Companies ' Pattern Exploration Using Data Mining Methods *4*(2), 87–98.

Wirawati Dewi Ahmad
Azuraliza Abu Bakar
Research Center for Software Technology and Management
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia
P86693@siswa.ukm.edu.my, azuraliza@ukm.edu.my

Received: 16 June 2018
Accepted: 20 August 2018
Published: 30 December 2018