Ensemble Machine Learning Model for Higher Learning Scholarship Award Decisions

Wirawati Dewi Ahmad[1], Azuraliza Abu Bakar[2]
Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology
University Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, MALAYSIA

***Abstract*—The role of higher learning in Malaysia is to ensure high quality educational ecosystems in developing individual potentials to fulfill the national aspiration. To implement this role with success, scholarship offer is an important part of strategic plan. Since the increasing number of undergraduates' student every year, the government must consider to apply a systematic strategy to manage the scholarship offering to ensure the scholarship recipient must be selected in effective way. The use of predictive model has shown effective can be made. In this paper, an ensemble knowledge model is proposed to support the scholarship award decision made by the organization. It generates list of eligible candidates to reduce human error and time taken to select the eligible candidate manually. Two approached of ensemble are presented. Firstly, ensembles of model and secondly ensembles of rule-based knowledge. The ensemble learning techniques, namely, boosting, bagging, voting and rules-based ensemble technique and five base learners' algorithm, namely, J48, Support Vector Machine (SVM), Artificial Neuron Network (ANN), Naïve Bayes (NB) and Random Tree (RT) are used to develop the model. Total of 87,000 scholarship application data are used in modelling process. The result on accuracy, precision, recall and F-measure measurement shows that the ensemble voting techniques gives the best accuracy of 86.9% compare to others techniques. This study also explores the rules obtained from the rules-based model J48 and Apriori and managed to select the best rules to develop an ensemble rules-based models which is improved the study for classification model for scholarship award.**

***Keywords*—Scholarship classification; ensemble learning; rules-based classification; rules-based ensemble**

I. INTRODUCTION

Since the year 2013, the number of Higher Education scholarship applications has increased in line with the increasing number of graduates at under graduate level every year. Therefore, the scholarship provider institution need to use the best approach or technique to select the best candidate for awarding education scholarship in appropriate time. Some of the scholarship provider institution have begun to adopt analytical data approach with appropriate classification techniques in managing and selecting the best candidates for scholarships award. By implementing this approach, the best eligible candidate can be selected automatically and provide better decision than manual methodology.

Through preliminary studies, the scholarship application processing method is using three key methods [9], namely, (1) pre-applicant checking that need to meet the general requirements using the online application system; (2) manual checking by responsible officer to verify applicant information and (3) final selection by a voting committee to select eligible candidates based on the information or characteristics of the applicant manually. These methods have two (2) major weaknesses, namely, (1) requiring skilled system programmers to maintain the system code and (2) required a long time and a large number of employees to process and do the checking for each application for decision within a specified time frame. Hence, this study aims to propose an analytical data approach using ensemble learning to select the best eligible candidate based on existing scholarship recipient's data. This approach helps to provide a solution for time and people issues in the current method by providing eligible candidate based on classification model of ensemble learning technique.

This study conducted a research on the best ensemble classification techniques for awarding scholarships to graduate students in Malaysia who applied for scholarships provided by the Ministry of Higher Education Malaysia. A set of real data contains a collection of application including the succesfull and unsuccesfull candidate that were collected from the online scholarship application system are processed and transformed according to each technique used. In this process, there are variety of filters to narrow down the important features. It is also important that any preparation of the data prior to fitting with the algorithm used. In order to develop the model, we carried out two types of experiments, namely, common ensemble learning model and rules-based ensemble model. These model are been trained and tested using two data distribution techniques namely strata distribution and k-folds distribution to ensure best learning and performance model.

Moreover, in the model evaluation task, we used two different evaluation, namely technical evaluation and expert evaluation to ensure that the assessment of performance performance of each technique is evaluated. For technical evaluatian, common performance measurement namely accuracy, precision, recall and F-measure are used to evaluate model performance in terms of accuracy and error for models produced using the selected technique. In addition, evaluations by domain specialists are also done to ensure better results are made. After evaluation of the model developed with the best performing techniques is the recommended technique.

It has been observed that most of the studies in ensemble learning especially in education domain is focusing in several main issues but not in scholarship award decision model. There also been observed that in scholarship award decision model, only single classification was applied.

Therefore, this paper is focusing in propose the best ensemble learning model for scholarship award in higher learning to help the scholarship provider turn to new and systematic way on selecting the best candidate to receive a scholarship while in their program of study.

II. Related Work

A. Data Analytic Approach in the Education Sector

The study of data analytic approach in the Education sector includes higher learning programs has expanded extensively. These studies contributed in providing solutions to addressing problems in the education sector in a country using classification and prediction methods [1][32]. The useful information obtained through this method may provide a solution to help the decision-making process in education insight in the future. The rapid growth in the size of the educational data also shows that the classification and prediction method applied requires a more efficient set of algorithms and techniques in giving the best results [1]. From the literature studies found that, there are two main aspects of the education which are often being focused in the research namely student group issues and higher learning institution issue as shown in Fig. 1.

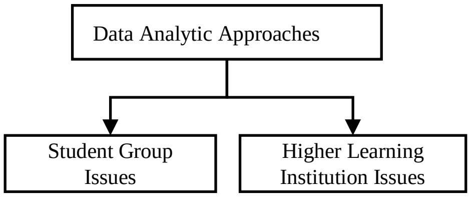


Fig. 1. Data Analytic Applied Studies in Educational Sector.

For the student’s group issues, there are many studies involved such as a classification and prediction model in monitoring student academic performance based on existing student achievement records [2][3][4], selecting best student who are eligible to be offered certain university program and selecting the best study programme or courses for student to register.

While for the higher learning organization many studies involved such as establishing a list of programme and courses to be offered by higher learning institution, selecting the best recipients for research grant award[3] and determine the best student intake for certain program ensuring a high pass rate [5]. Beside that, studies on a scholarship offers also been done using this approach to support decision making process in selecting best recipients.

B. Ensemble Learning Method

There are studies on producing the scholarship award decisions using data analytic approach which used a single classification method in producing the classification and prediction model [6][7][8][9]. However, the disadvantages of a single classification method may result in improperly constructed models and the quality of the decision made will be affected. A single classification method is found to reduce the chances of exploring knowledge using other techniques that cause the loss of information that is potentially important for the prediction process [10]. Additionally, the increasingly large data size factors required more effective technique to be used rather than single algorithm approach.

Therefore, the technique of ensemble learning is uses to combine more than two algorithms to produce the best learning model. Many studies as shown in [11][10][12][31] have proven that using this method may overcome the weaknesses of single classification method and a more robust final model can be developed. This is because using this method, the final model is capable to combine the characteristics of a single classifier used with either the same or different function [10] and gives a better result than a single classifier [10][13][12][11].

This technique has two main objectives: 1) to increase the accuracy of overall forecasts compared to a single classifier; and 2) improve the rate of generalization better because of its specific measurement measurements. As a result, the final classifier can solve unresolved issues with a single predictive model [13]. The model performance on examples not seen during training demonstrate the actual capabilities of the model [14].

- Ensemble Techniques

Boosting is the basic technique used in combined classification to enhance the capability of a classification model. This technique focuses on weak classifier by re-testing classified items inaccurate so as to be accurately classed in the next iteration. It was introduced by Schapire in 1990 and has been widely used to date. The welding process combines this technique as in Packing Techniques by sampling data on each iteration using the majority draw technique. One of the commonly used algorithms is AdaBoost.[15]

The AdaBoost algorithm was introduced by Freund and Schapire in 1997 [15]. It generates a set of hypotheses, and combines them through parliamentary majority parallel to the class achieved by each hypothesis. Hypotheses are generated by training weak classifier, using the records or samples obtained during the iteration process using regular training data. This cycle ensures that the classification of the wrong records classified by the previous classifier is more likely to be included in the subsequent iteration training data. Hence, the training data being constructed is increasingly difficult to classify but the accuracy of Success is classified accordingly is highly dependent on the previous welding pattern. It also seeks to improve welding accuracy and is commonly used in many fields. It is easy to implement and is not exposed to over-fitting problems [16]. It also produces a basic model sequence with different weight distribution over a set of exercises [10].

Bagging is an easy technique used in ensemble learning. It may combine base learners call classifiers and produce final ensemble model or use an algorithm with multiple set of testing data as ensemble basis. Based on the review by [17] the single-core tree algorithm and decision tree will produce different tree outputs. Different trees are formed when there is a change in the starting point of the training data that results in a decrease in stability. It can be used for any other classification method in the data mining approach. This technique is also suitable to be used in finding suitable models for large-sized data as the classification becomes easier as the training data sets are broken down according to certain techniques [15] (Deng 2007). The output of this technique is the final output with the majority draw technique which selects the best weld results among good models.

Voting is also a popular joining welding technique [18][19][20][20]. It combines many classmates and conducts a lottery process during welding. Many methods are used to select the best classifier vote among existing classmates. Average, minimum and maximum techniques are among those used. Classification using the majority technique is often chosen because it produces the output of the best model that is fair and balanced. For example, a set of data that has two class labels with the best three classifiers will determine the majority of classmates who label the most.

Table I shows the comparison on advantages and disadvantages between these three-ensemble techniques that leads to the experimental design.

In conclusion, the disadvantages of ensemble learning model is depends on some factors such as bias, noise and variants [21][22][12]. These weaknesses can be overcome by using the rules-based ensemble model. This technique also known as association classification (AC) whereby it combine rules generated from different rules-based model to perform as final ensemble model. Many studies have shown that AC is capable to enhance the ensemble classification capabilities with higher accuracy rather than common ensemble techniques. [23][24][25]. Therefore, in this study use two rules-based classification algorithm, namely, Apriori algorithm and J48 from decision tree algorithm to develop an ensemble rules-based model to determine higher learning scholarship award decision.

TABLE I. COMPARISON ON ENSEMBLE LEARNING TECHNIQUE

Technique	Advantages	Disadvantages
boosting	Improve the ability of a weak model by repeatedly conducting exercises using different data samples. Focusing on exemplary examples in the model iteration further results in higher model accuracy	Longer and higher cost modeling time Classification model becomes more complex and difficult to implement using real-time platforms Less suitable to use if the data has high noise levels and outliers
Bagging	Improve accuracy and reduce variance, thereby reducing overfitting problems Take samples of different data samples to improve model accuracy The easiest method compared to the upgrade technique Suitable for data sets with even high noise levels	Improve the computational complexity of the model Effectively only for examples of records with high frequency
Voting	Performance improvements for a large number of classes. Can be used on all types of data	Needing more good classifier is better because the classification is based on the majority.

The association rules is a very popular and extensive method that can be used to discover interesting relationships between variables in a very large database [26]. This technique seeks to find interesting and frequent patterns, knowing the correlation between data sets in a data repository that will provide useful information in support of future decision-making. In this study, the use of apriori algorithm techniques used to look at the attributes of attributes in Malaysia's higher education scholarship packages. This attribute represents the frequency of candidate's pattern of success or is not offered a higher education scholarship offer.

Apriori is a simple algorithm used to find frequent set of records in a data set. The output of this algorithm is a set of sequential tips with the frequency set of the data item. The generated tips are dependent on several filtering metrics such as minimum support and confidence level [21]. This means that with this algorithm and using existing data sets, it is able to provide frequent data set patterns and help in identifying useful information to support future decision-making processes. There are two main processes in this technique, namely the process of obtaining the maximum set of items by relying on the minimum support parameters and the rules of generating rule rules based on the minimum confidence. Both of these processes are interconnected in determining the characteristics and number of rule rules that are generated [27].

In this study, a base learners under decision tree category used was J48. This algorithm was chosen because the decision tree 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 illustrates that the technique successfully classified well. The decision tree can handle and handle large quantities of input data such as text with numerical or nominal numerical data [26]. It is a supervised learning approach that has the ability to extract information from a large amount of data based on rules or decision rules [26][28].

In this paper, we present the comparison of ensemble learning model for scholarship award decision based on the real scholarship reciepient’s database using five different ensemble learning techniques in order to propose the best technique in scholarship award decision making. Eventhough we used the existing techniques, however, the same research strategy has not been implemented on the real scholarship data in Malaysia which requires specific and systematic reviews. Hence, the results of this study will proposed an ensemble model for shcolarship award decisions which can improve the scholarship decision making process in the future.

III. RESEARCH METHODOLOGY

This section explains the process that involved in the proposed ensemble learning framework for scholarship award decision. The study of classification model for scholarship recipients has been done widely by implementing the data mining approaches using a single classification technique. Therefore, in this study we present a better technique in a classification model using ensemble learning classification approaches to determine scholarship award decision by using common ensemble classification method and a rules-based ensemble technique. The actual scholarship data that been used was obtained from the government scholarship provider. Fig. 2 shows the proposed ensemble learning framework based on the literature review conducted.

A. Scholarship Data Collection and Pre-Processing

After a preliminary study of the data, the data preparation process was conducted to provide a complete set of data for modeling purposes. The two main processes undertaken in this stage are data integration and cleaning and features selection and transformation.

1) Data intergration and cleaning: The raw data obtained was analyzed to see the type and size of the attributes and their importance. A total of 87, 000 data was retrieved with more than 40 attributes recorded. After a preliminary study of the data on several steps has been taken to ensure clean and complete data is provided. Data noise exists due to a poorly managed database. Records that do not store the correct values for the attribute type and size are deleted. Likely, actual databases are sometimes used to carry out system entry tests that receive meaningful test data.

For a record that has no value for the final decision of a committee it 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 which indicates the applicant is not committed to applying for a scholarship.

An important attribute that determines the candidate's decision is a decent value of an office that stores a value other than 1 or 2 is also eliminated. This odd value may be due to a system testing process or an applicant who does not resubmit an incomplete document during the application process. This data record is eliminated because it is considered not important in the welding modeling process.

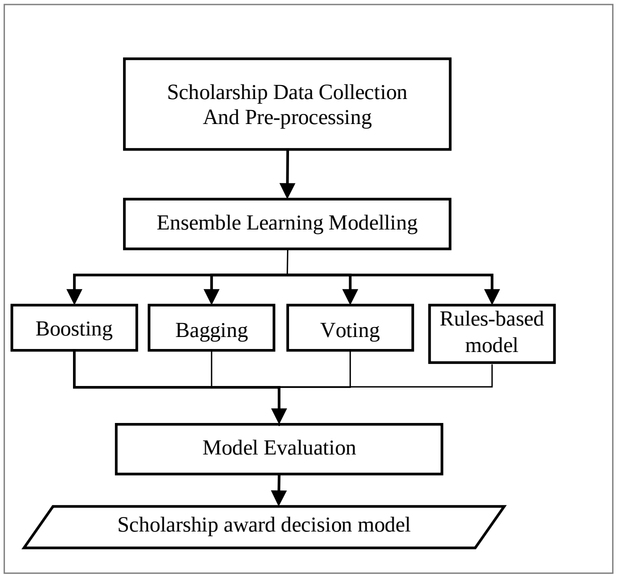

Scholarship Data Collection And Pre-processing
Ensemble Learning Modelling
Boosting
Bagging
Voting
Rules-based model
Model Evaluation
Scholarship award decision model

Fig. 2. Proposed Ensemble Learning Framework to Determine Scholarship Award Decision.

Inconsistent data records are also found during the data analysis process. For example, the applicant's age exceeds 64 years, the applicant's age is less than 22 years, the non-existent identity card number, or a record exceeding one application is also eliminated.

As a result of the above process, a total of 57,000 net records were successfully provided. Attributes can be categorized into three categories, attributes that are eliminated i.e. over 29 attributes, attributes are retained by 11 attributes, new attributes created to replace attributes that are deleted by 4 attributes. Table II shows an example of a list that has been deleted after the data clearing and integration process.

2) Features selection and transformation: 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 modeling. This new set of data is assessed using specific feature selection techniques and transformed into appropriate forms to meet modeling algorithm requirements. Table III depicts the selected attributes for ensemble knowledge model.

B. Ensemble Learning Modelling

The preprocessed scholarship data set was spilt into two set of training and testing data. This distribution used two techniques namely strata distribution and k-folds distribution. While using these two techniques, five base learner algorithm were chosen as a base learner, namely, J48, Support Vector Machine (SVM), Naïve Bayes (NB), Artificial Neuron Network (ANN) and Random Tree (RT). These five bases learner will be used in three common ensemble techniques namely Boosting, Bagging and Voting. Each model produced by these technique will be evaluated using common evaluation metrics. To develop the ensemble models, there are two type of ensemble learning modelling framework used namely Common ensemble learning model and Rules-based ensemble model.

TABLE II. TABLE SCHOLARSHIP DATASET FINAL ATTRIBUTES

Attribute name	**Type**
Age	Varchar
Gender	Varchar
marital status	Varchar
disabilities	Varchar
sponsorship status	Varchar
work status	Varchar
approval status	Integer
field of study	Integer
program structure	Integer
graduated years of study	Varchar
Level of study	Integer
University	varchar

TABLE III. ATTRIBUTES TRANSFORMATION CODING

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
University codes	Special codes for each university

For common ensemble learning model, we used three common ensemble techniques namely boosting, bagging and voting. From these techniques, three ensemble learning experiment were designed in order to obtain the best ensemble model for higher learning schorlaship award decision using ensemble learning technique. For boosting and bagging experiment design, the five base learners were used as an individual classifier while in voting experimental design, the five base learners will be taken as one ensemble classifiers as shown in Fig. 3.

For the last experiment, the rules-based ensemble model was developed using the set of rules generated previously by the decision tree rules based model, J48 and an Apriori rules based model as shown in Fig. 4. Each rules generated from this experiment will be analyze and measured in order to select the best rules. The ensemble rules then combined in an ensemble rules-based model and tested with using testing data prepared earlier.

The rules-based ensemble method is the techniques that can improve ensemble learning capabilities with high accuracy rather than single ensemble model. In this study, the rules obtained from the single-layer rules-based model were excavated and evaluated to be combined with rules obtained from the rules-based classification model using the Apriori algorithm. The combination of these rules will determine whether the model's ability to classify Eligible or Not qualified candidates is better than common ensemble techniques used.

In this framework, two sets of training data are used as training data used in previous experiment and new training data sets only sample candidates are not eligible. This is aimed at getting more tips for “Not Eligible” candidate cases. These rules are then combined with the selected rules from the J48 model in a new set of rules that will be used in the rules-based ensemble approach experiment. To determine the correct classification and incorrect classification rate, the majority algorithm is shown as Fig. 5.

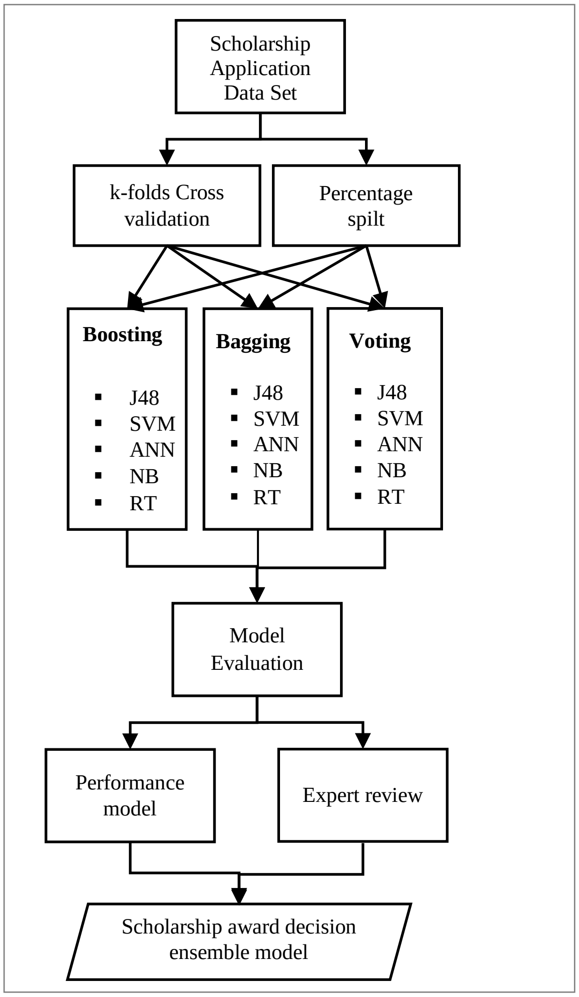


Fig. 3. Common Ensemble Learning Model to Determine Scholarship Award Decision.

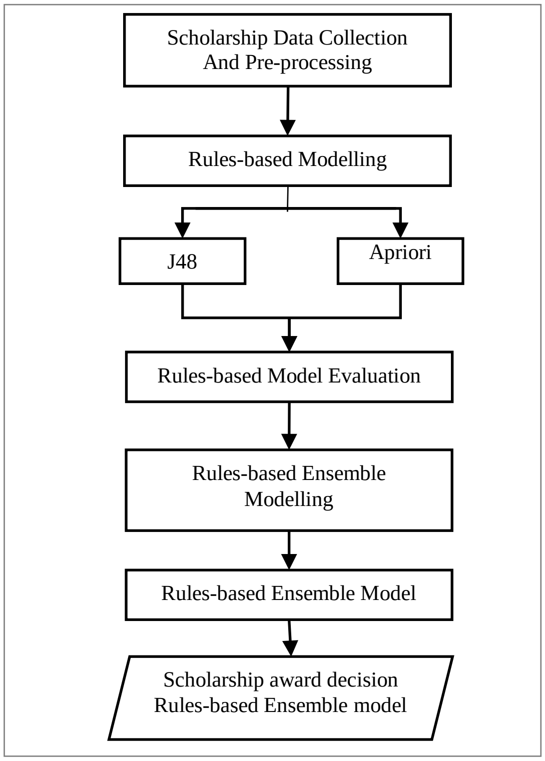
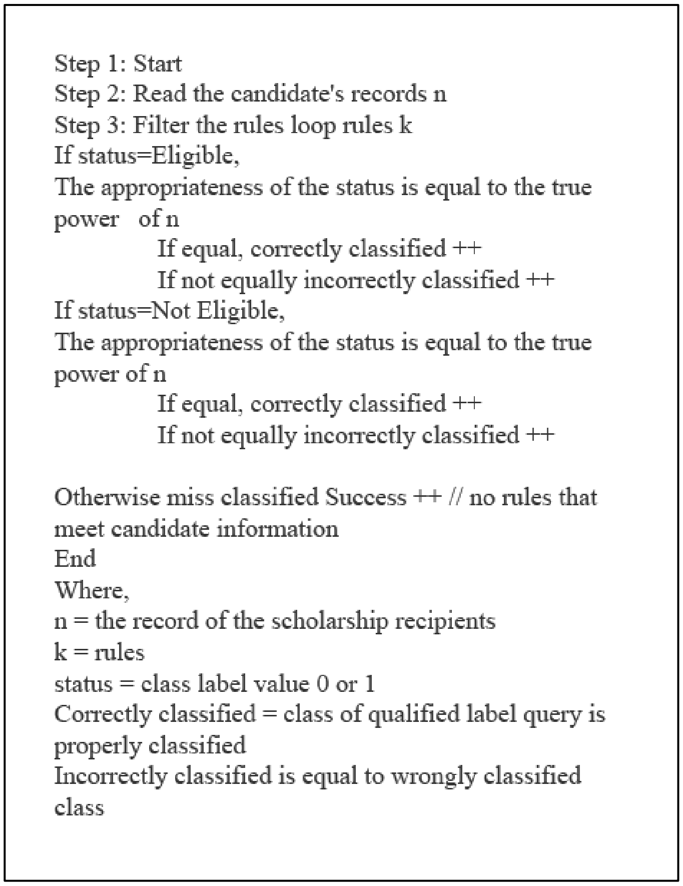

Scholarship Data Collection And Pre-processing
Rules-based Modelling
J48
Apriori
Rules-based Model Evaluation
Rules-based Ensemble Modelling
Rules-based Ensemble Model
Scholarship award decision Rules-based Ensemble model

Fig. 4. Rules-based Ensemble Model to Determine Scholarship Award.



```
Step 1: Start
Step 2: Read the candidate's records n
Step 3: Filter the rules loop rules k
If status=Eligible,
The appropriateness of the status is equal to the true
power   of n
            If equal, correctly classified ++
            If not equally incorrectly classified ++
If status=Not Eligible,
The appropriateness of the status is equal to the true
power of n
            If equal, correctly classified ++
            If not equally incorrectly classified ++

Otherwise miss classified Success ++ // no rules that
meet candidate information
End
Where,
n = the record of the scholarship recipients
k = rules
status = class label value 0 or 1
Correctly classified = class of qualified label query is
properly classified
Incorrectly classified is equal to wrongly classified
class
```

Fig. 5. Majority Algorithm.

C. Model Evaluation

In this study, model evaluation is based on two approaches, namely technical evaluation and expert evaluation. The performance on classification model were measured using common performance measurement namely accuracy, precision, recall and F-measure. The formula to calculate these measure are shown below (1) (2) (3) and (4).

$$Accuracy = \frac{correctly\ classified\ scholarship\ result}{total\ scholarship\ data} \quad (1)$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (2)$$

$$recall\ = \frac{true\ positive}{true\ positive + false\ negative} \quad (3)$$

$$F - measure\ = \frac{precision\ \times recall}{precision + recall} \quad (4)$$

Where,

1) True positive(TP) = eligible candidate correctly classified

2) False positive(FP) = not eligible candidate incorrectly labeled as eligible candidate

3) False negative(FP) = eligible candidate incorrectly labeled as not eligible

In this study, the expert evaluation also used to review and validate the acquisition of rules from the rules-based ensemble modelling. The scholarship domain expert who has been involved in scholarship management for more than two years involved in selecting and verifying the best rules to be used in rules-based ensemble modelling. The results from all experiment, were then compared to propose the best ensemble model generated by this study.

IV. Result and Discussion

A. Ensemble Learning Modelling

The accuracy of a classification model for scholarship award decision can be measured by calculating the percentage of correctly classified number of eligible scholarship candidate and divide by the total of candidate number. In this study, the voting technique showed the highest accuracy of 86.9% compared to bagging and boosting techniques. This is followed by bagging ANN model with accuracy of 86.5% while the best model for boosting technique is the ensemble classification model using boosting SVM which achieved 86.3% of accuracy rate. While the weakest model is ensemble classification model using boosting RT algorithm which is only 81.6% of accuracy. However, overall, the ensemble technique with five base learners produced accuracy more than 80% and can be categorized as good with the best technique of voting techniques as shown in Fig. 6.

Beside accuracy, precision rate is used to show the rate of predicted positive by dividing the number of eligible candidate correctly classified and the total of the number of eligible candidate correctly classified and the number of not eligible candidate incorrectly labeled as eligible candidate. By observed this measure, we can see the ability of a model to predict the eligible candidate to be awarded. In this study, the bagging SVM technique show the highest precision rate of

0.868 followed by the model of bagging NB technique with precision rate of 0.863. While the weakest model was boosting SVM model with precision rate at 0.7712.

In this study, recall measurement indicate the ability of a model to predict the eligible candidates by dividing the number of eligible candidate correctly classified and the total of the number of eligible candidate correctly classified and the number of eligible candidate incorrectly labeled as not eligible candidate. In this study, the ensemble classification model using bagging SVM technique achieves the highest recall rate of 0.869. The overall result show that recall measurement for overall model shows good value which is more than 0.800. This demonstrates the ability of each model to classify the eligible candidate class is good.

For F-measure rates, both measurements precision and recall was combined to see the optimal blend of prediction ability for an ensemble learning model. In this study, the best F-measure rate is 0.808, using bagging SVM technique. Overall result for precision, recall and F-measure are shown in Fig. 7.

B. Rules-Based Ensemble Modelling

1) J48 model rules extraction: In the classification process besides being able to specify the class labels for data records, the algorithm is used to generate rules that determine the class labels. One of the common rules-based algorithm widely used is J48 algorithm. The rules pattern generated from J48 classification modeling process can be used to obtain hidden information that may be useful to the data owner and combine with other rules-based model to generate new rules-based model.

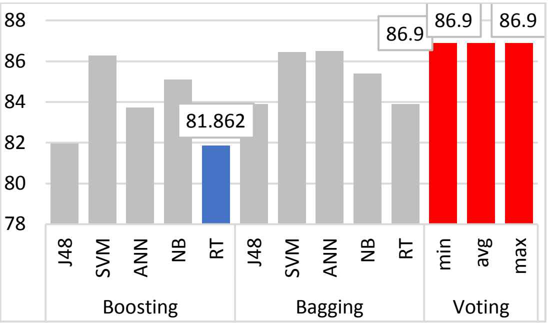


Fig. 6. Rules-based Ensemble Model to Determine Scholarship Award.

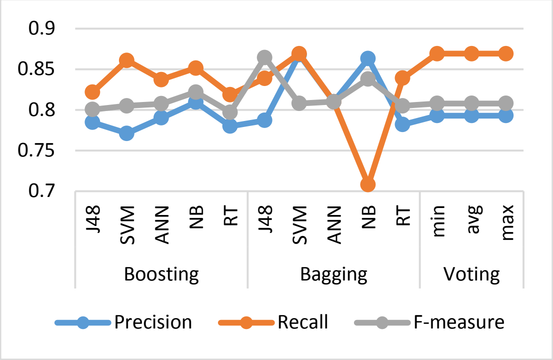


Fig. 7. Others Measures Result for Common Ensemble Learning Model.

Extracting useful information and knowledge to support decision making is one of the main goals of data mining. A knowledge analysis process enables hidden information to be obtained and verified to determine its importance to the stakeholders. In this study, knowledge analysis was conducted on the results of the rules generated by the J48 model to develop an ensemble rules-based model in determine a higher learning scholarship eligible candidate. Rules with high frequencies are taken and been verified to ensure that the rules used in ensemble rules-based model are valid and useful in selecting the most qualified candidates for awarding Higher Education scholarships based on records of successful and failed candidates.

The rules generated by this algorithm is depended by two class labels, which are rules that give the label class 1 the ‘Eligible’ and class 2 ‘Not Eligible’. The rule rules generated by this model are 2771 rules with the size of the rules 3265. These rules are verified with the help of expert domains to see their importance and validity. The process of validation this rules is important to ensure that rules are not significant can be filtered and remove. The method of filtering useful rules is run by the number of frequencies each rules and expert assistance [29]. The result of this process manages to identify two categories of rules namely the rule of determining the classification of eligible candidates and the rule of determining the classification of not eligible candidates. For the classification of qualified candidates, the discussion concurred that the rationing tips of more than 15 candidates were classified as important tips of 70 of the rules while the rules for unqualified cases considered important were 67 rules. Table IV lists examples of rules that define the Eligible label class and the knowledge acquisition gathered when the rules transformed based on the original data set.

Based on the analysis of these rules, it can be concluded that the rules generated by the J48 algorithm are more focused on the case of ‘Eligible’ candidates more than the ‘Not eligible’ candidates. This is because the number of rules for determining the Ineligible candidate case is small and the number of records is small compared to the Eligible Candidate.

In order to improve the model's capabilities, the improvement in the classification of negative records, which is not feasible, should be improved. In order to address this issue, the study will then conduct experiments on one of the techniques to improve the accuracy of a classifier model, i.e. the rules-based ensemble approach. To derive more useful rules, this study developed a model from a common rules-based algorithm namely Apriori and extracted the rules generated by this model.

Furthermore, the identified rules will be combined with the rules from the J48 model as discussed earlier into a new rules-based ensemble model using the simple majority technique which will be discussed in next section. This data mining approach is solution that can improve classification and prediction task with higher accuracy. According to the study (Mittal et al., 2017), this technique provide a solution to extract a classifier that contains the simple If-Then’s rule and produces high accuracy.

TABLE IV. RULES WITH HIGH FREQUENCY FOR ELIGIBLE CANDIDATE

Rules	**Information Gathered**
age_range = 2 AND level_of_study = 8 AND university_code = M0101028 AND field_of_study = 2 AND gender = 2	Women candidates aged 25 to 29 who apply to continue their studies at the University 'M0101028' and continue their studies in Science and Technology
university_code = M0100612 AND work_status = 1 AND sponsorship_status = 0 AND field_of_study = 2 AND graduated_years_of _study = 5 AND program_structure = 1 AND marital_status = 2 AND gender = 2	Candidates applied to continue their studies at University 'M0100612' and continue their studies in Science and Technology, have been sponsored at undergraduate, graduating years between 2006 and 2010 and married status
age_range = 2 AND level_of_study = 8 AND university_code = M0100612 AND work_status = 1 AND sponsorship_status = 0 DAN field_of_study = 2 AND graduated_years_of _study = 5 AND program_structure = 1 AND marital_status = 2 AND gender = 2	Women candidates aged 25 to 29 who apply to continue their studies at the University 'M0100612' at Masters level and continue their studies in Science and Technology, have been sponsored at undergraduate and marital status
age_range = 2 AND level_of_study = 8 AND university_code = M0100612 AND work_status = 1 AND sponsorship_status = 0 AND field_of_study = 2 AND julat_thn_ijz = 5 AND program_structure = 1 AND gender =2	Women candidates aged 25 to 29 who apply to continue their studies at the University 'M0100612' at Masters level and continue their studies in Science and Technology

2) *Apriori model:* In this experiment, the input data is a set of scholarship data of 10,000 candidates as D1. This data set has gone through the process of cleansing and transformation as described in previous chapters. This data set is also the same set of data used in the process of ensemble learning modeling before. While the output of this algorithm is a set of rules with the frequency. The minimum support value used in this experiment is 0.1 with a confidence level of 0.9. The purpose of selecting these values is to produce a number of important rules.

As a result, a total of 500 rules were observed, and only 221 rules related to the class label were identified. The rules generated by this model are includes of 2, 3, 4 and 5 sequences. Examples of rules with the frequency values obtained are as in Tables V and VI. Furthermore, the rules generated by this model is evaluated by technical measurement and expert domains knowledge to find best rules to be used in rules-based ensemble model.

In the rules evaluation process, only rules with minimum confident level and minimum support was reviewed and selected. During this process, the rules was taken and analyzed using SQL to determine whether it is useful or not. Expert review also been performed to identify the best rules. From this process, we found that only 18 important rules were identified from a total of 221 rules on the corresponding label class. These rules includes rules for eligible candidate cases. As no suitable rules are in place for Not eligible cases, another rules-based model was developed using the only Not eligible candidate data set (D2) using the same value for confident level and minimum support.

The results of this experiment produced estimated a total of 200 rules. From the evaluation process, only 69 rules were associated with the class label and from that only 8 important rules identified from this experiment. The total of 26 rules were obtained from Apriori modelling will be used as rules condition for rules-based ensemble model to determine scholarship award decision. Table VII shows the result of rules selected from these two experiments.

3) *Rules-based ensemble model:* In this study, the available rules from a single classifier model (J48) experiment and the model of associative rules are combined using simple majority technique as described in previous chapters. The rules-based ensemble model was develop using PHP programming language and the same training data in previous experiment. As shown in Table VIII. A total of 137 rules were taken from the J48 model while 26 rules from the Apriori model then ensembled in the rules-based model. This model then tested using the same set of data and the accuracy of the model was observered.

As the result, Table IX shows that the ability of this model to weld properly for qualified candidates is high and this gives a good value for the accuracy and accuracy of this model of 85.94% and 0.8609.

TABLE V. 2 AND 3 SEQUENCE RULES EXAMPLE

2 sequences	**3 sequences**
age_range =2 5174 ==> gender=1 5169	age_range =2 class_label=1 4665 ==> gender=1 4661
marital_status =1 4984 ==> gender=1 4979	class_label =1 field_of_study =1 4582 ==> gender=1 4578

TABLE VI. 4 AND 5 SEQUENCE RULES EXAMPLE

4 sequences	**5 sequences**
age_range =2 disabilities =1 class_label =1 4653 ==> gender=1 4649	disabilities =1 sponsorship_status =0 level_of_study =8 program_structure =1 4620 ==> gender=1 4614
disabilities =1 class_label =1 field_of_study =1 4573 ==> gender=1 4569	gender=1 disabilities =1 sponsorship_status =0 class_label =1 4937 ==> program_structure =1 4930
disabilities =1 sponsorship_status =0 work_status =1 class_label =1 5716 ==> gender=1 5708	gender=1 sponsorship_status =0 level_of_study =8 program_structure =1 4621 ==> disabilities =1 4614

TABLE VII. NUMBER OF RULES OBTAINED FROM APRIORI MODEL

	Number of rules gathered	**Number of rules includes class label**	**Number of selected rules**
Apriori D1	500	221	18
Apriori D2	200	69	8
Total	700	290	26

TABLE VIII. NUMBER OF RULES FOR RULES-BASED ENSEMBLE MODELLING

	Number of rules
J48	137
Apriori	26
Total	163

TABLE IX. RULES-BASED ENSEMBLE MODEL RESULT

Measures	**Value**
TP	8585
FP	1387
TN	2
FN	0
Accuracy	85.94%
Precision	0.8609
Recall	1
F-measure	0.9252

From the experiment result shows that True positive (TP) value is height and this gave high accuracy rate (85.95%). For the FP value, there are 1,387 record of not eligible candidate were incorrectly labeled as eligible candidate. However, compared to other ensemble model experiment, the F-measure for rules-based ensemble model is higher 0.9252.

V. CONCLUSION

In this study we present a better technique for a classification model for higher learning scholarship award decision using ensemble learning approach by using common ensemble classification method and a rules-based ensemble technique. All the model, then evaluated by technical and domain experts to obtain the best model.

Additionally, the preprocessing data managed to provide a complete dataset to be used as training and testing dataset. The preparation of this dataset also managed to propose the suitable type and transformation codes for some attributes which can be applied at the beginning of input data process in scholarship application form in the future.

Overall result shows that each of ensemble learning techniques used, perform a good result on the classification and prediction performance as shown Table X with the best model was voting technique with highest accuracy of 86.9% which is identical to the experimental results of [30]. In addition, this study also explores the study of rules-based ensemble model for higher learning scholarship award. This study found that the accuracy of rules-based ensemble are slightly different than the best ensemble model (85.94%) but acceptable. This study also explores the rules obtained from the rules-based model J48 and Apriori and managed to select the best rules to enhance to develop an ensemble rules-based models which is improved the study for classification model for scholarship award in [8][6][9].

This means that when we ensemble five base learners J48, SVM, ANN, NB and RT, the ensemble model performs significantly better than ensemble it alone using boosting and bagging techniques.

The experiments also suggest the experiment using different dataset with different scholarship program ensemble techniques can be adopted to build better learning and prediction model to predict scholarship award for higher learning students.

TABLE X. OVERALL RESULT SHOWS THAT EACH OF ENSEMBLE LEARNING TECHNIQUES USED

Ensemble	Boosting					Bagging					Voting		
Techniques	J48	SVM	ANN	NB	RT	J48	SVM	ANN	NB	RT	min	avg	max
Accuracy (%)	81.97	86.299	83.7325	85.11034	81.862	83.9	86.45	86.5	85.4	83.9	86.9	86.9	86.9
Precision	0.7848	0.7712	0.79025	0.8094	0.78	0.869	0.868	0.81	0.863	0.782	0.793	0.793	0.793
Recall	0.8216	0.8608	0.83725	0.8514	0.8188	0.869	0.869	0.81	0.708	0.839	0.869	0.869	0.869
F-measure	0.8004	0.805	0.8075	0.8218	0.797	0.869	0.808	0.81	0.838	0.805	0.808	0.808	0.808

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

[1] Dutt A, Ismail MA, Herawan T. A Systematic Review on Educational Data Mining. IEEE Access. 2017;5:15991–6005.

[2] Adejo OW, Connolly T. Predicting student academic performance using multi-model heterogeneous ensemble approach. J Appl Res High Educ. 2018;10:61–75.

[3] Mesarić J, Šebalj D. Decision trees for predicting the academic success of students. Croat Oper Res Rev [Internet]. 2016;7:367–88. Available from: http://hrcak.srce.hr/174215?lang=en.

[4] Shahiri AM, Husain W, Rashid NA. A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Comput Sci [Internet]. Elsevier Masson SAS; 2015;72:414–22. Available from: http://dx.doi.org/10.1016/j.procs.2015.12.157.

[5] Yu CH, Digangi S, Jannasch-Pennell A, Kaprolet C. A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. J Data Sci. 2010;8:307–25.

[6] Raharja YP. Rancang Bangun Sistem Rekomendasi Beasiswa Menggunakan ALgoritma Klasifikasi C4.5 pada Universitas Dian Nuswantoro. Undinus [Internet]. 2014;1–4. Available from: http://eprints.dinus.ac.id/13408/.

[7] Tun KT, Aye AM. Selection of Appropriate Candidates for Scholarship Application Form using KNN Algorithm. Int J Sci Eng Technol Res. 2014;3:1019–26.

[8] Alhassan JK, Lawal SA. Using Data Mining Technique for Scholarship Disbursement. 2015;9:1511–4.

[9] Azuraliza, Arshad A. Rough Set and Decision Tree Model for Determining Scholarship Award Qualification. 2013;12:65–70.

[10] Oza NC, Tumer K. Classifier ensembles: Select real-world applications. Inf Fusion. 2008;9:4–20.

[11] Dietterich TG. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees. Mach Learn [Internet]. 2000;40:139–57. Available from: http://en.scientificcommons.org/42637098%5Cnuuid/7906280C-AEF8-405A-9A94-6BAA1DDAED1E

[12] Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33:1–39.

[13] Stapel M, Zheng Z, Pinkwart N. An Ensemble Method to Predict Student Performance in an Online Math Learning Environment. Proc 9th Int Conf Educ Data Min. 2016;231–8.

[14] Schapire RE. The boosting approach to machine learning: an overview. Nonlinear Estim Classif [Internet]. 2003;171:149–71. Available from: http://www.ams.org/mathscinet/search/publications.html?pg1=MR&s1=MR2005788%5Cnpapers2://publication/uuid/13F19159-186B-4FAA-9433-AE2BE3D153D8.

[15] Deng H. A Brief Introduction to Adaboost. 2007;1–35.

[16] Palaniappan S, Rajinikanth T V, Govardhan A. Emerging Trends in Electrical, Communications and Information Technologies. 2017;394:165–74. Available from: http://link.springer.com/10.1007/978-981-10-1540-3.

[17] Goyal A, Thakur S, Chowdhury R. Using Ensemble Learning and Association Rules to Help Car Buyers Make Informed Choices. Proc Int Conf Big Data Adv Wirel Technol - BDAW '16 [Internet]. 2016;1–5. Available from: http://dl.acm.org/citation.cfm?doid=3010089.3010093

[18] Ahmed M, Rasool AG, Afzal H, Siddiqi I. Improving handwriting based gender classification using ensemble classifiers. Expert Syst Appl [Internet]. Elsevier Ltd; 2017;85:158–68. Available from: http://dx.doi.org/10.1016/j.eswa.2017.05.033.

[19] Hamsagayathri P, Sampath P. Decision Tree Classifiers for Classification of Breast Cancer. Int J Curr Pharm Res [Internet]. 2017;9:31. Available from: http://innovareacademics.in/journals/index.php/ijcpr/article/view/17377.

[20] Pham BT, Tien Bui D, Prakash I. Landslide Susceptibility Assessment Using Bagging Ensemble Based Alternating Decision Trees, Logistic Regression and J48 Decision Trees Methods: A Comparative Study. Geotech Geol Eng. Springer International Publishing; 2017;35:2597–611.

[21] Maniar H. A Predictive Student Performance Analytics Scheme using Auto-Adjust Apriori Algorithm. 2017;157:4–7.

[22] Polikar R. Ensemble based systems in decision making. Circuits Syst Mag IEEE. 2006;6:21–45.

[23] Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2:916–54.

[24] Trandafili E, Allkoçi A, Kajo E, Xhuvani A. Discovery and evaluation of student's profiles with machine learning. Proc Fifth Balk Conf Informatics - BCI '12 [Internet]. 2012;174. Available from: http://dl.acm.org/citation.cfm?doid=2371316.2371350.

[25] Wook M, Yahaya YH, Wahab N, Isa MRM, Awang NF, Seong HY. Predicting NDUM Student's Academic Performance Using Data Mining Techniques. 2009 Second Int Conf Comput Electr Eng [Internet]. 2009;357–61. Available from: http://ieeexplore.ieee.org/document/5380417/.

[26] Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR Models - Strategies and Importance. Int J Drug Des Disocovery. 2011;2:511–9.

[27] Mittal K, Aggarwal G, Mahajan P. a Comparative Study of Association Rule Mining Techniques and Predictive Mining Approaches for Association Classification. Int J Adv Res Comput Sci [Internet]. 2017;8:365–72. Available from: http://ijarcs.info/index.php/Ijarcs/article/viewFile/4984/4331.

[28] Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. J Intell Learn Syst Appl [Internet]. 2017;9:1–16. Available from: http://www.scirp.org/journal/doi.aspx?DOI= 10.4236/jilsa.2017.91001.

[29] Azuraliza, Idris N, Hamdan AR, Othman Z, Ahmad Nazari MZ, Zainudin S. Classification models for outbreak detection in oil and gas pollution area. Proc 2011 Int Conf Electr Eng Informatics, ICEEI 2011. 2011;0–5.

[30] Manaf SA, Mustapha N, Sulaiman N, Husin NA, Shah Zainuddin MN, Mohd Shafri HZ. Majority voting of ensemble classifiers to improve shoreline extraction of medium resolution satellite images. J Theor Appl Inf Technol. 2017;95:4394–405.

[31] Hamood Ali Alshalabi, Sabrina Tiun, Nazlia Omar. A Comparative Study Of The Ensemble And BaseClassifiers Performance In Malay Text Categorization. Asia-Pacific Journal of Information Technology and Multimedia. Vol. 6 No. 2, December 2017; 53 - 64.

[32] Azhar Mohd Khairy, Afzan Adam, Mohd Ridzwan Yaakub. Data Analytics In Malaysian Education System:Revealing The Success Of Sijil Pelajaran Malaysia From Ujian Aptitud Sekolah Rendah. Asia-Pacific Journal of Information Technology and Multimedia. Vol. 7 No. 2, December 2018; 29 - 45.