International Journal of Engineering & Technology, 7 (4.31) (2018) 285-289

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Case Study of UPNM Students Performance Classification Algorithms

Syarifah B. Rahayu[1]*, Nur D. Kamarudin[2], Zuraini Zainol[3]

[1]*Cyber Security Centre, National Defence University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia*
[2]*Computer Science Department, Faculty of Science and Defence Technology, National Defence University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia*
**Corresponding author E-mail: syarifahbahiyah@upnm.edu.my*

Abstract

Most students have a problem to keep track on their learning performance. Some lecturers with high teaching hours and burden of administration jobs may have difficulty to identify weak and low performance students. In this study, three classification techniques are applied on educational datasets to predict the students' performance based on coursework assessments. Thus, this prediction results may help lecturers and students to improve their teaching and learning process. The objective of study is to predict students' performance based on coursework assessments using classification algorithms. The selected classification algorithms applied in this study such as J48 Decision Tree, Naïve Bayes and kNN. WEKA is used as an experimental tool. The selected algorithms are applied on a data of student database of Data Mining subject. Findings shows Naïve Bayes outperforms other classification algorithms with above 80% prediction rate. Thus, the students' performance for Data Mining Subject is improved. As a conclusion, the classification algorithms can predict students' performance on a particular subject based on coursework assessments.

Keywords: *Prediction; Comparative Analysis; Educational Data Mining*

1. Introduction

Educational Data Mining (EDM) researches use data mining tools to process large quantities of data to discover meaningful patterns in order to predict students' performances to enhance teaching and learning outcomes. These researches can also be used as a platform to alert student on the risk of failure and to provide recommendations for student improvement in their learning process. One of the criteria for a high quality university is based on its excellent record of academic achievement [1]. Therefore, student performance is a crucial part in higher learning institution. A student performance is often measured based on the subject work assessments and final exam. The proposed methodology is to analyze students' performance of a particular subject. The findings are used for predicting their performance before they are taking a final exam. Thus, it will assist the lecturers or educators to identify students who need supports to perform well in the final exam. Besides, students can improve their learning process in order to pass the subject [2]. The objective of this study is to predict the students' performance based on Malaysia Grading System. These performances are predicted using three different classification algorithms, for example, J48 Decision Tree, Naïve Bayes and kNN.

The rest of this paper is organized as follows. Section 2 presents the background and related work to this study. In Section 3, we described the framework of our proposed research. Section 4 discusses the experiment and results. Finally, we conclude this paper with future work in section 5.

2. Background and Related Works

In this section, some related topics on data mining, knowledge discovery in databases, classification algorithms and reviews on related work are discussed.

2.1. Data Mining and Knowledge Discovery in Database

Data Mining (DM) and Knowledge Discovery in Databases (KDD) are two terms that are often used interchangeably. KDD can be defined as a process of finding useful information and patterns in data [3]. In KDD, DM is placed in the fourth steps of the KDD process. Technically, the KDD process consists of five main steps such as selection, pre-processing, transformation, data mining, and interpretation or evaluation (see Fig 1).

According to [3], DM is often applied to extract hidden information and useful patterns using algorithms from massive amounts of data which is derived by the KDD process. Such valuable information and patterns may assist the top level managers in decision making. DM has been applied in various application areas such as market based analysis, healthcare [5], smart homes [6], business, text documents [7-10], environmental studies [11, 12], flood detection [13], crime investigation, fraud detection, geology, food microbiology, astronomy, etc. Researchers [14] summarized some common data mining tasks and techniques (see Table 1). These tasks and techniques can be applied individually or they can be combined together to perform more sophisticated processes.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

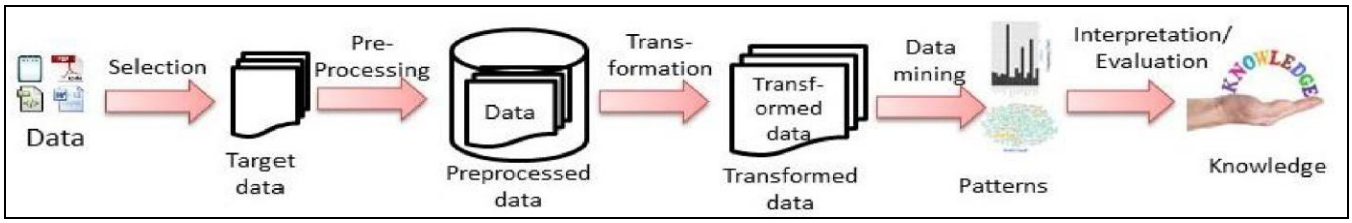

Fig. 1: Knowledge Discovery in Databases (KDD) process adopted from [4]

2.2. Classification Algorithms

Classification is a supervised learning where the classes are often determined before data can be mined [3]. Technically, classification will assign the data into several predefined classes. Classification technique is often applied for predicting or describing dataset or nominal categories. Each classification technique (see Table 1) will apply a learning algorithm to identify a model which is best fitted the relationship between the set of attributes and the class label (predefined class) of the input data. The model that has been produced by a learning algorithm should be able to fit the input data and predict the class label of the records correctly [15].

Table 1: Data Mining Tasks and Techniques Adopted From [14]

DM Tasks	DM Techniques
Classification	Decision Tree Induction, Bayesian Classification, Fuzzy Logic, Support Vector Machines (SVM), k-Nearest Neighbors (K-NN), Rough Set Approach, Genetic Algorithm (GA), etc.
Clustering	Partitioning Methods, Hierarchical Methods, Density-based Methods, Grid-based Methods, etc.
Association Rules	Frequent Item set Mining Methods (e.g., Apriori, FP-Growth)

Some examples of classification technique are detecting spam email messages based on the message header and content, categorizing cells as malignant or benign based on the result of MRI, identifying credit risks based on bank loan, predicting students' performance, etc.

In [16-19], the Decision Tree (J48), Bayesian Classifier and k-Nearest Neighbor (kNN) classifiers have been implemented to evaluate students' performances based on several observational attributes such as accumulated exam grades, percentages or classes (i.e distinction, fail etc). Based on comparative analysis of classifier in [16], the Bayesian classifier outperformed the decision tree and kNN classifier on predicting students' performances via average True Positive (TP) rate. However, in analyzing the TP rate for each classes (Distinction, First, Second, Third and Fail); it has been observed that, the prediction rates are not uniform among classes. Hence, the gap of prediction rate among classes is varied almost 90% in some cases. This might be due to the insufficient data of certain classes especially in distinction and fail classes.

To discover the optimal classification model for decision tree, research in [20] did the comparison of different algorithms comprises of J48, ID3, C4.5, REPTree, Random Tree and Random Forest. Out of six decision tree algorithms, the highest percentage is achieved using the model relying on the algorithm J48. Based on 161 questionnaires, two researchers from University of Basrah have analyzed and assisted academic achievers in higher education using Bayesian Classification Method [21]. For attribute selection, questions with high correlation averages have been adopted to enhance the accuracy of classification.

Recent work has been done to demonstrate the efficiency of Semi-Supervised Learning (SSL) methods for the performance prediction of high school students using their final examination assessment percentage [22]. In this work, various SSL algorithms such as Self-training, Co-training, Democratic Co-learning, Tri-training, De-Tri training and RASCO are implemented in KEEL Software tool. In addition, Friedman Aligned Ranks nonparametric test is used to measure the performances of these algorithms. Moreover, in second phase of experiments, the performance of SSL classifiers have been compared with supervised method, Naïve Bayes. From the observation, it can be concluded that SSL algorithm are comparatively better than the respective supervised algorithm, Naïve Bayes based on both measurement; the accuracy and Friedman Aligned Rank.

A comprehensive survey is then carried out by the Indian Researchers to discuss about the current approaches and potential areas in EDM [23]. This paper reported the details of researches done in the area of education in tabular form describing methodologies and findings of each research and identifies potential research areas for future scope. Similar research is conducted by the researcher in [24] where the new potential domains of EDM have been proposed. According to this paper, EDM data is not limited to predict the student's performance but can also be utilized in other domains of education sector (i.e. optimization of resources or human resource purposes).

In comparison of correlation among pre and post enrollment factors and employability using data mining tools, many of today's graduates are lacking interpersonal communication skills, creative and critical thinking, problem solving, analytical skills, and team work [25]. It has been concluded that cognitive factors such as set of behaviors, skills and attitudes play a significant role in prediction of student's marketability after graduation. Another work presented by Research Group for Work, Organizational, and Personnel Psychology, Department of Psychology, KU Leuven, Leuven, Belgium stated that employability is in strong correlation with competences and dispositions [26].

Motivated by the previous researches, this research attempts to evaluate the performances of several students by measuring their subject work assessment percentages (Quizzes, Tutorials and Test) via predetermined classes endorsed by the university and Malaysian Grading System to predict their performance in final exam. To the best of our knowledge, this is the only research paper that discusses the student's performance prediction in Malaysia based on Malaysian Grading System apart from using the students' CGPA. We proposed this research as a preliminary assessment tool where we narrower the scope of research to cater early recognition of student who needs help in certain subject not the whole performances of student from his or her CGPA. In addition, by analyzing the distributed rank or weightage on each assessment using decision tree, this research will offer guidance to a lecturer on improving the teaching plan based on the learning outcomes in certain assessment as well as to identify weak students to improve the students' learning process prior to the final exam. Future contribution will be the automatic application on a platform that is able to read, analyze and predict the outcome of student's progress based on certain assessments in difficult or challenging university subjects for intelligent tutoring or lecturing applications.

3. Research Framework

Figure 2 illustrates the proposed framework for predicting students' performance in Data Mining subject. The first stage is data collection. The data about students related to a particular subject

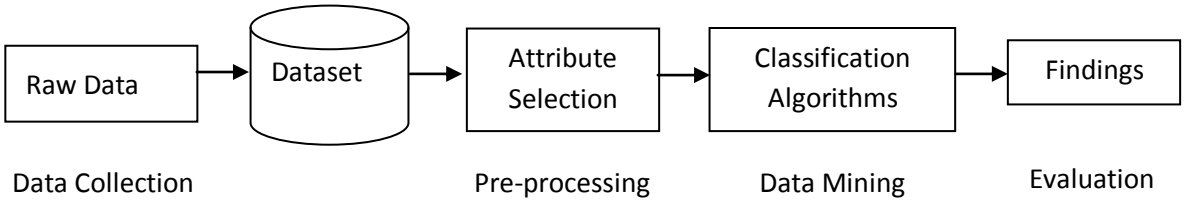

Fig. 2: A proposed framework of classification model to predict students' performance on a particular subject

are collected. In this study, the data set is obtained from last semester records of students who are registering for Data Mining subject. These students are majoring in Computer Science in Science Computer Department, National Defence University of Malaysia. For the study case, a sample of 71 students is selected in this experiment. This study will be used as starting point for a deep machine learning in the future. Thus, we are focusing on the smaller dataset before gathering large number of dataset. The Data Mining subject is consists of 60 marks coursework assessments and 40 marks final exam. The coursework assessments of this subject are quizzes, tutorials and test. Students are required to obtain at least 40 marks to pass the subject.

During the pre-processing stage, the dataset is prepared before applying the classification algorithms. Then, data attributes are identified and selected. Table 2 below is the students' related attributes. There are 3 quizzes, 3 tutorials and a test. The accumulated values of these attributes are equal to 60 marks. Thus the calculation will be the values of attributes, *t* is divided with the total values of attributes, *m* and multiplied with marks of assessment coursework, *n.*

Table 2: Students Related Attributes

Attributes	Type	Values	Grade
Quiz 1	Real	[1,10]	[A+, F]
Quiz 2	Real	[1,10]	[A+, F]
Quiz 3	Real	[1,10]	[A+, F]
Tutorial 1	Real	[1,5]	[A+, F]
Tutorial 2	Real	[1,5]	[A+, F]
Tutorial 3	Real	[1,5]	[A+, F]
Test	Real	[1,50]	[A+, F]

Each coursework assessment including the final exam is graded based on Malaysia Grading System for university level as tabulated in Table 3.

Table 3: Malaysia Grading System for University Level

Grade	Scale	Grade Description
A+	90.00 – 100.00	Exceptional
A to B+	76.00 – 89.99	Excellent
B to C+	65.00 – 75.99	Good
C to D	40.00 – 64.99	Average
F	0.00 – 39.99	Fail

The next stage is to apply classification algorithms on the data set. This study is using WEKA (Waikato Environment for Knowledge Analysis) version 3.8.2 as an experimental tool. This tool is developed at University of Waikato, New Zealand, and known as a prominent open source data mining tools comprises of several machine learning classifiers. It has been widely used among researchers and data scientists for data pre-processing, classification, regression, clustering, association rules, and visualization. Three algorithms have been selected in this stage. The selected classification algorithms are J48 Decision Tree, Naïve Bayes and kNN. The data set is divided into 10 equally sized folds using the 10-fold cross validation procedure provided by WEKA (fig 3).

4. Results and Discussion

The objective of this paper is to predict students' performance in Data Mining subject. Three classification algorithms are selected to perform the prediction model.

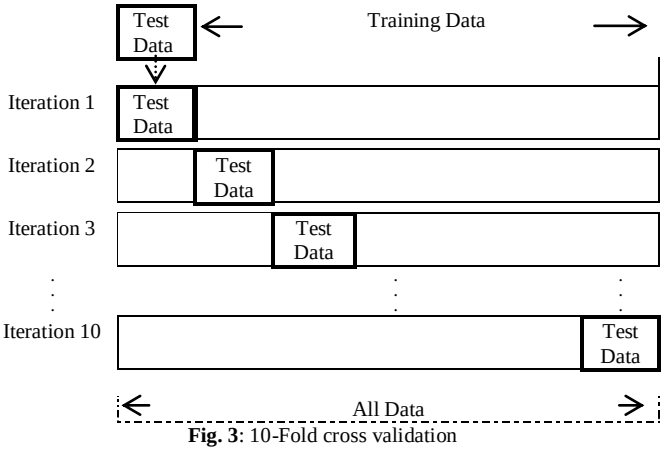

Fig. 3: 10-Fold cross validation

4.1. Results of Decision Tree

J48 classification algorithm produces a decision tree (fig 4) with the size of 15 nodes and 8 leaves. The correctly classified instances are 56 with 78.8732% for the 10-fold cross-validation testing. Below is the IF-THEN rules based on the produced decision tree:

IF Test <= 23.7 AND Tutorial 3 <= 4 THEN Class = "Good"

IF Test <= 23.7 AND Tutorial 3 > 4 AND Quiz 1 <= 1.5 THEN Class = "Good"

IF Test <= 23.7 AND Tutorial 3 > 4 AND Quiz 1 > 1.5 AND Test <= 15.9 AND Quiz 3 <= 4.5 THEN Class = "Good"

IF Test <=23.7 AND Tutorial 3 > 4 AND Quiz 1 > 1.5 AND Test <= 15.9 AND Quiz 3 > 4.5 AND Quiz 1 <=3.5 THEN Class = "Good"

IF Test <=23.7 AND Tutorial 3 > 4 AND Quiz 1 > 1.5 AND Test <= 15.9 AND Quiz 3 > 4.5 AND Quiz 1 > 3.5 THEN Class = "Excellent"

IF Test <=23.7 AND Tutorial 3 > 4 AND Quiz 1 > 1.5 AND Test> 15.9 THEN Class = "Excellent"

IF Test > 23.7 AND Quiz 2 <= 4.5 THEN Class = "Excellent"

IF Test > 23.7 AND Quiz 2 > 4.5 THEN Class = "Exceptional"

Based on the IF-ELSE rules, students who perform better in the two attributes such as Test, and Quiz 2 are likely to pass the subject with flying color. Other attributes may cause them to be at risk. Examples of students data based on J48 Decision Tree (table 4).

Table 4: Example of Student Data based on J48 Decision Tree

	Test	Quiz 2	Tutorial 3	Class
Student C	24.9	5.0	5.0	Exceptional
Student B	22.2	5.0	5.0	Excellent
Student B	14.1	5.0	4.0	Good

Table 5 below shows the classification results for the decision tree algorithm. The highest True Positive (TP) Rate is recorded in Excellent class (91.7%) while the lowest TP rate is recorded in Good class (42.9%). This maybe due to high number of students who fall in the Excellent grade for coursework assessments. However, the weighted average of TP Rate for all observational classes is 78.9%. In addition, the Precision is high for two classes: Good (85.7%) and Excellent (80%), and low for Exceptional (66.7%). The weighted average of Precision is 79.4%.

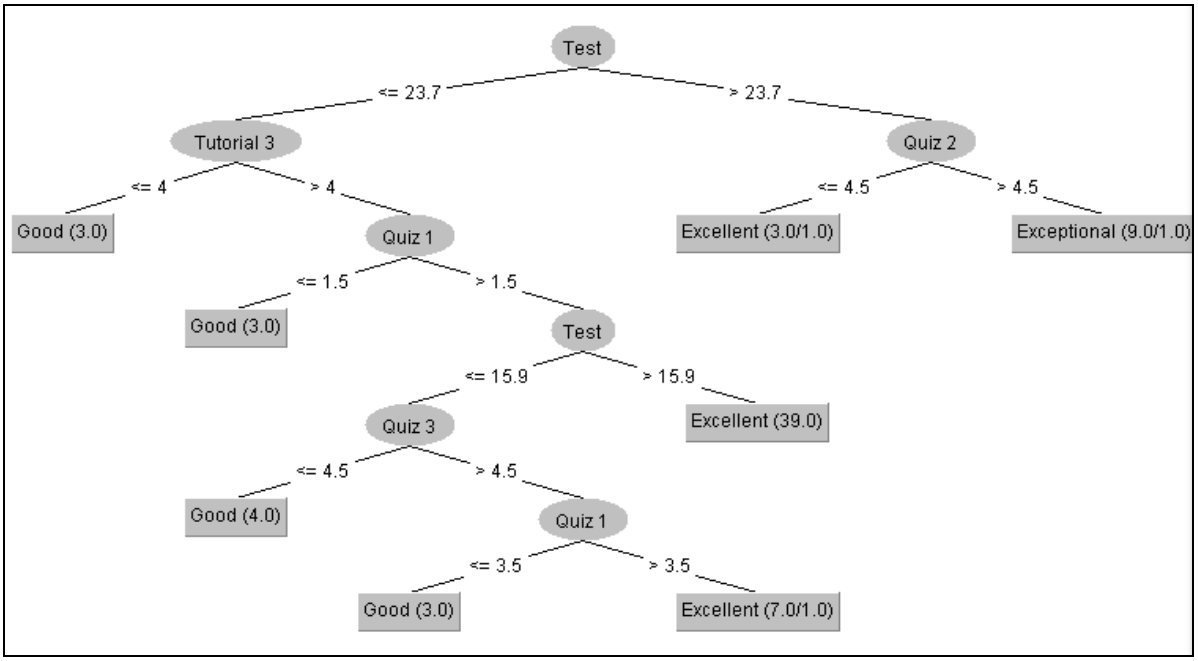

Fig. 4: J48 tree visualization

Table 5: Classification Results for the Decision Tree Algorithm (J48)

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Excellent	0.917	0.478	0.800	0.917	0.854
Exceptional	0.667	0.048	0.667	0.667	0.667
Good	0.429	0.018	0.857	0.429	0.571
Weighted Avg.	0.789	0.333	0.794	0.789	0.775

4.2. Results of Naïve Bayes

On the other hand, Naïve Bayes classifier has correctly classified 57 instances with 80.2817 % TP rate by using 10-fold cross-validation testing.

Table 6 below shows the classification results for Naïve Bayes algorithm. The True Positive (TP) Rate is the highest for the Exceptional class (100%). However, the TP Rate is lower in two classes; Excellent (79.2%) and Good (71.4%).The weighted average of TP Rate is recorded as 80.3%. The Precision is high for Excellent class (90.5%), and low for: Good (66.7%) and Exceptional (64.3%) classes. The weighted average of Precision is 82.5%.

Table 6: Classification Results for the Naïve Bayes Algorithm

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Excellent	0.792	0.174	0.905	0.792	0.844
Exceptional	1.000	0.081	0.643	1.000	0.783
Good	0.714	0.088	0.667	0.714	0.690
Weighted Avg.	0.803	0.145	0.825	0.803	0.806

4.3. Results of kNN

kNN classifier correctly classified 53 instances with 74.6479% TP rate by using 10-fold cross-validation testing.

Table 7 below shows the classification results for k-NN algorithm. The True Positive (TP) Rate is the highest for Excellent class (87.5%). The TP Rate of Exceptional Class is recorded as 77.8% and the lowest TP rate is in Good class (28.6%). However, the weighted average of TP Rate is 74.6%. Two classes have highest Precision: Excellent (77.8%) and Exceptional (77.8%). The class with low Precision is Good (50%). However the weighted average of Precision is 72.3%.

Table 7: Classification Results for the kNN Algorithm

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Excellent	0.875	0.522	0.778	0.875	0.824
Exceptional	0.778	0.032	0.778	0.778	0.778
Good	0.286	0.070	0.500	0.286	0.364
Weighted Avg.	0.746	0.371	0.723	0.746	0.727

4.4. Comparison Results for the kNN algorithm

By comparative analysis, the results of the classification algorithms reveal that the Naïve Bayes performs better than the other algorithms. Furthermore, Naïve Bayes produces the highest classification accuracy for Exceptional class. Naïve Bayes produces accuracy above 80% for all observational classes. The worst classification accuracy is recorded in J48 and kNN for Good class, which is below 50%. This is due to the number of tested data applied on each class. It has been observed that, higher number of sample data resulted in better classification accuracy. The weighted average for overall accuracy of all tested classifiers is well above 70%. Table 8 shows the performance of each classification algorithms.

Table 8: Classification Algorithm Performance

	J48 Decision Tree	Naïve Bayes	k-NN
TP Rate	0.789	0.803	0.746
Precision	0.794	0.825	0.723

This finding assists lecturers to identify weak students and help them to improve their marks. Weak students may also identify their weaknesses and change their study learning. Table 9 shows the comparison of results before and after final exam. Number of students who had grade C to D+ before the final exam was reduced from 22 students to 4 students only. Students who had grade A to B+ and B to C+ increased from 21 students to 35 students and 25 students to 30 students, respectively. Overall the results show students are getting better results after the final exam.

Table 9: Students' Study Performance

Grade	Coursework	Final Exam
A+	3*	2*
A to B+	21*	35*
B to C+	25*	30*
C to D	22*	4*
F	-	-

5. Conclusion

In this paper, three classification algorithms have been implemented on students' subject databases to identify their performance in Data Mining subject based on coursework assessments. The findings show Naïve Bayes produces the highest classification accuracy for Exceptional class. Overall weighted average for all algorithms is above 70%.

This study helps both students and lecturers to enhance their learning and teaching methods. Students who have low marks in their subject perform better during the final exam since remedial and necessary action are taken to enhance their learning process. At the same time, lecturers give better supports on certain assessment to improve their students' performance. As a result, the overall students' performance is improved. This prediction system may embed into e-learning or tutoring system, where students can measures their study performance.

Future work is to identify other attributes that may contribute students' performance. Furthermore, Next study it will involve larger educational dataset and different types of classification algorithms such as Support Vector Machine (SVM), REPTree, RandomTree and LMT to predict the students' performance during their undergraduate studies.

References

[1] Shahiri AM & Husain W, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, (2015), pp. 414-422.

[2] Dunham MH, *Data mining: Introductory and Advanced Topics*, Pearson Education India, (2006).

[3] Fayyad U, Piatetsky-Shapiro G & Smyth P, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, Vol. 39, No. 11, (1996), pp. 27-34.

[4] Azam M, Loo J, Naeem U, Khan S, Lasebae A, & Gemikonakli O, "A framework to recognise daily life activities with wireless proximity and object usage data," *23rd International Symposium on Personal, Indoor and Mobile Radio Communication*, Sydney, Australia, (2012), pp. 590-595.

[5] Jothi N, Abdul RN & Husain W, Data Mining in Healthcare – A Review, Procedia Computer Science,Vol. 72, (2015), pp. 306-313.

[6] Rashid MM, Gondal I, & Kamruzzaman J, "Mining associated patterns from wireless sensor networks," *IEEE Transactions on Computers*, Vol. 64, No. 7, (2015), pp. 1998-2011.

[7] Zainol Z, Marzukhi S, Nohuddin PNE, Noormaanshah WMU & Zakaria O, "Document Clustering in Military Explicit Knowledge: A Study on Peacekeeping Documents," in *Advances in Visual Informatics: 5th International Visual Informatics Conference, IVIC 2017, Bangi, Malaysia, November 28–30, 2017*, H. Badioze Zaman *et al.*, Eds. Cham: Springer International Publishing, (2017), pp. 175-184.

[8] Zainol Z, Nohuddin PNE, Mohd TAT & Zakaria O, "Text analytics of unstructured textual data: a study on military peacekeeping document using R text mining package". *Proceeding of the 6th International Conference on Computing and Informatics (ICOCI17)*, (2017), pp. 1–7.

[9] Nohuddin PNE, Zainol Z, Chao FC, James MT & Nordin A, "Keyword based Clustering Technique for Collections of Hadith Chapters," *International Journal on Islamic Applications in Computer Science And Technologies – IJASAT*, Vol. 4, No. 3, (2015), pp. 11-18.

[10] Zainol Z, Nohuddin PNE, Jaymes MTH & Marzukhi S, "Discovering "interesting" Keyword Patterns in Hadith Chapter Documents," in *Information and Communication Technology (ICICTM16), International Conference on*, Kuala Lumpur, Malaysia, (2016), pp. 104-108.

[11] Rashid RAA, Nohuddin PNE & Zainol Z, "Association Rule Mining Using Time Series Data for Malaysia Climate Variability Prediction," in *Advances in Visual Informatics: 5th International Visual Informatics Conference, IVIC 2017, Bangi, Malaysia, November 28–30, 2017, Proceedings*, H. Badioze Zaman *et al.*, Eds. Cham: Springer International Publishing, (2017), pp. 120-130.

[12] Nohuddin PNE, Zainol Z & Ubaidah MAB, "Trend Cluster Analysis of Wave Data for Renewable Energy," *Advanced Science Letters*, Vol. 24, No. 2, (2018), pp. 951-955.

[13] Harun NA, Makhtar M, Aziz AA, Zakaria ZA, Abdullah FS & Jusoh JA, "The Application of Apriori Algorithm in Predicting Flood Areas," *International Journal on Advanced Science, Engineering and Information Technology*, Vol. 7, No. 3, (2017).

[14] Nohuddin PNE, Z. Zainol, Lee ASH, Nordin AI & Yusoff Z, "A case study in knowledge acquisition for logistic cargo distribution data mining framework," *International Journal of Advanced and Applied Sciences*, Vol. 5, No. 1, (2017), pp. 8-14.

[15] Tan PN, *Introduction to data mining*, Pearson Education India, (2006).

[16] Anuradha C & Velmurugan T, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance," *Indian Journal of Science and Technology*, Vol. 8, No. 15, (2015).

[17] Goga M, Kuyoro S & Goga N, "A Recommender for Improving the Student Academic Performance," *Procedia - Social and Behavioral Sciences*, Vol. 180, (2015), pp. 1481-1488.

[18] Kabakchieva D, "Predicting Student Performance by Using Data Mining Methods for Classification," in *Cybernetics and Information Technologie*, Vol. 13(2013), pp. 61.

[19] Hussain S, Abdulaziz DN, Ba-Alwib FM & Najoua R, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 9, No. 2, (2018) pp. 447-459.

[20] Mesaric J & Šebalj D, "Decision trees for predicting the academic success of students," *Croatian Operational Research Review*, Vol. 9, No. 2, (2016), pp. 367-388.

[21] Khalaf A, Humadi AM, Akeel W & Hashim AS, "Students' Success Prediction based on Bayes Algorithms," *SSRN elibrary*, Elsevier, (2017), pp. 6-12.

[22] Kostopoulos G, Livieris I, Kotsiantis S & Tampakas V, "Enhancing high school students' performance based on semi-supervised methods," *in 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, (2017), pp. 1-6.

[23] Thakar P, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," *International Journal of Computer Applications*, Vol. 110, No. 15, (2015), pp. 60-68.

[24] Kumar DV & Anupama C, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education*," International Journal of Advanced Computer Science and Applications*, Vol.3, No. 2, (2011), pp. 81-84.

[25] Bakar NAA, Mustapha A & Nasir KM, "Clustering analysis for empowering skills in graduate employability model," *Australian Journal of Basic and Applied Sciences*, Vol. 7, No. 14, (2013), pp. 21-24.

[26] Dorien V, Nele DC, Ellen P & Hans DW, "Defining perceived employability: a psychological approach," *Personnel Review*, Vol. 43, No. 4, (2014), pp. 592-605.