Data Exploration and Analysis of Alternative Learning System Accreditation and Equivalency Test Result Using Data Mining

J A Talingdan¹, J T Trinidad, Jr.² and T D Palaoag³

¹ Abra State Institute of Sciences and Technology, Lagangilang, Abra
² Abra State Institute of Sciences and Technology, Lagangilang, Abra
³ University of the Cordilleras, Baguio City, Philippines

janelyn.ambre@gmail.com, jun_trinidad@ymail.com, tpalaoag@gmail.com

Abstract. Alternative Learning System (ALS) is a subsystem of Depatment of Education (DepEd) that serves as an option of learners who cannot afford to go in a formal education. The research focuses on the data exploration and analysis of ALS accreditation and equivalency test result using data mining. The ALS 2014 to 2016 A & E test results in the secondary level were used as data sets in the study. The A & E test results revealed that the passing rate is doubled per year. The results were clustered using k-means clustering algorithm and they were grouped into good, medium, and low standard learners to identify students need exceptional stuff for enhancement. From the clustered data, it was found out that the strand they are weak in is strand 4 which is the Development of Self and a Sense of Community with a general average of 84.23. It also revealed that the essay type of exam got the lowest score with a general average of 2.14 compared to the multiple type of exam that covers the five learning strands. Furthermore, decision tree and naive bayes were also employed in the study to predict the performance of the learners in the A & E test and determine which is better to use for prediction. It was concluded that naive bayes performs better because the accuracy rate is higher than the decision tree algorithm.

1. Introduction

Data mining in education is called educational data mining. Educational Data Mining (EDM) is the development of methods for the exploration of data in the educational environment. It is an interesting area of research that provides better understanding on the performance of the students to improve quality of education. This can be processed through different data mining techniques such as clustering and classification [1].

Clustering is one of the data mining techniques to group the data into different clusters. It can be a data modeling technique to summarize big data. Clustering is very helpful to any disciplines and has an important role in numerous applications [2]. On the other hand, classification is a supervised learning technique to create a model to classify data according to a predefined class label. The aim of classification is for prediction to help in decision making. It can be used to predict students’ performances [3].

There are studies that use educational data mining particularly the clustering algorithm to assess the performance of students and group them to determine what kind of learning strategies and techniques will be given according to the level of their understanding. This is necessary for the improvement of
educational learning process. Classification is also used to predict the performance of students. This is essential to educators because it will help them in the decision making to make their students succeed.

From the studies mentioned above, it is clear and understandable that educational data mining has a great impact in educational institutions. The results serve as the basis in decision making of the institution to improve the quality of education rendered to the students. Likewise in the informal basic form of education, learners’ academic performance is also their main goal. It is their goal to improve the performance of learners in the competency assessment conducted every year.

In the Philippines, there is the program of the government to bring education to all Filipinos who do not have a chance to attend the formal basic education due to many reasons such as drop outs from school, no school in the communities and poverty. The program is called Alternative Learning System (ALS). The Alternative Learning System is a parallel learning system in the Philippines that is implemented by the Department of Education. It is an option of learners who cannot afford to go in a formal education. This kind of education happens outside the classroom, in community-based organizations or any areas where learning process is possible. The learners can take the Alternative Learning System Accreditation and Equivalency Test after graduation. The test covers the five learning strands namely, Communication Skills, Problem-Solving and Critical Thinking, Sustainable Use of Resources/Productivity, Development of Self, and a Sense of Community and Expanding One’s World Vision. It is a paper-and-pencil test designed to measure the competencies of the ALS learners. The test is divided into two parts, the first part is the multiple choice type of exam that covers the five learning strands and the second part is the composition or the essay. Passers are given a certificate/diploma, certifying their competencies as comparable graduates of the formal school system. Passers are qualified to enroll in secondary and post-secondary schools.

In the 2016 A & E test result in the Cordillera Administrative Region, the Department of Education – Alternative Learning System Baguio Division again registered the highest passing rate in the region with an average of 82.78 percent for the elementary and secondary levels. Ifugao was second with a passing rate of 62.44 percent. Mountain Province followed third with 58.92 percent. Abra and Benguet followed closely with 57.58 percent and 57.16 percent, respectively, then Apayao at 43.05 percent, and Kalinga at 27.25 percent. It has been observed that the percentage of passers in the province of Abra is average compared to the passers of Baguio City [7]. Therefore, it is a challenge for every ALS facilitator to improve the performance of the learners in the A & E test in the province.

The research focused on Educational Data Mining (EDM) using the A & E test results of secondary level of the previous three years from 2014 to 2016 as data sets. The A & E test results were clustered using K-means algorithm of Rapid Miner to determine how the group of learners performed in the A & E test for three years. The research also determined what strands the learners are weak in and needs improvement. Furthermore, the research predicted the learners’ performance in the next A & E test using the decision tree and naïve bayes algorithm. Lastly, the performance of the two algorithms was tested to determine which is better in prediction.

The result of the study will be a great help to the ALS facilitators of Abra to come up with solutions to improve the passing rate in the province. It could be beneficial by providing additional support in the delivery of instruction and improving their teaching strategies to meet the standard of learners according to the group they belong to. Through this the ALS facilitators can enhance the quality of education and provide better educational services. This will also help the ALS administration to assess, evaluate, plan and decide on how they will level up their performance in the A & E test.

1.1. Conceptual Framework

The ALS 2014 to 2016 Accreditation and Equivalency Test results who enrolled in the secondary level were used as data sets. These data sets were clustered using k-means clustering algorithm. Afterwards, the learners were grouped into three major categories as good, medium and low standard learners [5]. From the clustered data it determined the performance of the learners in the 2014 to 2016 A & E test and determined the strand(s) the learners are weak in and need improvement. On the other hand, decision tree and naïve bayes were also employed in the study to predict the performance of the learners on the A & E test. It also determined which is better to use in prediction by testing their performance rate.
2. Methodology

2.1. Data Set
The data set used in this study was the result of the A & E test of ALS learners who enrolled secondary level in the province of Abra for the past three years (2014-2016). The data is composed of six attributes namely, Communication Skills that denoted by SS1, Problem-Solving and Critical Thinking denoted by SS2, Sustainable Use of Resources/Productivity denoted by SS3, Development of Self, and a Sense of Community denoted by SS4 and Expanding One’s World Vision denoted by SS5 and Essay. Table 1 presents the attributes and their values that exist in the data set as taken from the A & E test result.

|    | SS1 | SS2 | SS3 | SS4 | SS5 | Essay | Remarks |
|----|-----|-----|-----|-----|-----|-------|---------|
| 1  | 103 | 110 | 114 | 103 | 120 | 4     | Passed  |
| 2  | 100 | 110 | 110 | 102 | 120 | 3     | Passed  |
| 3  | 93  | 95  | 105 | 87  | 105 | 2     | Failed  |
| 4  | 103 | 110 | 114 | 99  | 120 | 4     | Passed  |

2.2. Data Processing Models
The clustering method applied in this study was the k-means algorithm. This algorithm was used to choose the best cluster center to be the centroid. It is one of the simplest unsupervised learning algorithms used for clustering. Initially the learners are all in one group then the researchers segmented the learners into three groups based on their standard score for each strand. The clustered groups were labeled as good, medium and poor standard learner. Through this, it is easy to determine the strands that need improvement and the group of learners that need extra effort in the teaching process to improve their A & E test result.

Decision tree algorithm and naive bayes were applied in the A & E test results to generate a model to predict learners’ performance. Through these algorithms, they predicted learners who will fail and pass the A & E test. The results obtained may be helpful to the facilitators so that appropriate actions could be done to increase the passing rate. The facilitators may prepare a remedial program or more additional works for the learners predicted to fail in order for them to excel in the next A & E test.

3. Results and Discussion
From the 2014 to 2016 A & E test results of ALS learners enrolled in the secondary level that served as data sets in the study, it is noted that the passing rate is improving every year. From 11.5% of 2014,
it increased to 26.5% in 2015 and jumped to 53.9% in 2016. The result signifies that there is a great improvement in the passing rate since it was doubled every year. However, the passing rate is not yet enough to make the province top the A & E test compared to Baguio City which topped the 2016 A & E test with a passing rate of 82.78. Table 2 shows the percentage rate of those who failed and passed.

| Remarks | 2014 | 2015 | 2016 |
|---------|------|------|------|
| Passed  | 45   | 122  | 219  |
| Failed  | 347  | 339  | 187  |
| Total   | 392  | 461  | 406  |

3.1. Cluster Model

From the 2014 to 2016 A & E test results, k – means clustering was applied and group the learners into three clusters. For year 2014, there were 392 learners who took the A & E test and 91 of them were grouped in cluster 0, 148 learners for cluster 1 and 153 learners to cluster 2. For year 2015, there were 221 learners grouped to cluster 0, 1 leaner to cluster 1 and 239 learners to cluster 2 to give a total of 461 learners who took the A & E test. Lastly, for year 2016 there were 139 learners grouped in cluster 0, 93 learners in cluster 1 and 174 learners in cluster 2 to make the 406 learners who took the A & E test on that year. Table 3 shows the distribution of the learners.

| Cluster | 2014 | 2015 | 2016 |
|---------|------|------|------|
| Cluster 0 | 91   | 221  | 139  |
| Cluster 1 | 148  | 1    | 93   |
| Cluster 2 | 153  | 239  | 174  |
| Total    | 392  | 461  | 406  |

3.2. Centroid Table

Table 4 is the centroid table and it is observed that from 2014 to 2016, the essay type of exam is the weakest area of the learners with a general average of 2.14 compared to the multiple choice type of exam that covers the five learning strands. From the five learning strands, strand 4 which is the Development of Self and a Sense of Community is the strand that the learners are weak in and needs improvement because it has the lowest score of a general average of 84.23. Moreover, learners should be given different teaching strategies and learning techniques based on the group they belong. Learners labeled as poor standard learners should be given exceptional stuff for enhancement compared to other groups as they got low performance in all strands. For the good standard learners, the strand that they got low score should be given extra emphasis to help them excel in the A & E test. Lastly, the average standard learners should be given more attention like giving extra work especially on the strands that they are weak in to improve their score in the A & E test. Along with this, more exercises should be given on essay type of exam to improve the test result.

From the k-means clustering centroid table of 2014 A & E test result, it can be perceived that in the multiple choice type of exam, learners belong to cluster 2 had mediate scores in all strands compared to the other two clusters. This group is called medium standard learners. They scored best in strand 5 which is the Expanding One’s World Vision but they score low in strand 4 which is the Development of Self and a Sense of Community. Learners who belong to cluster 1 had the best performance in all aspects of exam compared to other clusters. Their group can be grouped as good standard learners. On the other hand, cluster 0 got the lowest score and can be considered as poor standard learners. It can be observed also that all clusters did not excel in the essay type of exam.
Table 4. Centroid Table

| Attribute | 2014 Clusters | 2015 Clusters | 2016 Clusters | Gen. AVG. |
|-----------|--------------|--------------|--------------|----------|
|           | 0 1 2 AVG.   | 0 1 2 AVG.   | 0 1 2 AVG.   | AVG.     |
| SS1       | 78.64 93.66 87.18 86.49 93.03 83.85 91.63 98.89 83.88 91.33 89.82 |
| SS2       | 87.37 98.76 93.65 93.26 98.98 91.16 95.05 105.4 90.86 96.78 97.68 95.33 |
| SS3       | 80.18 96.89 87.65 88.24 97.3 85.09 92.46 109.1 83.93 91.53 94.86 91.85 |
| SS4       | 72.07 88.85 80.85 80.59 88.22 88 78.75 84.99 97.62 77.52 86.2 87.11 84.23 |
| SS5       | 90.53 115.7 100.9 102.4 115.3 14 97.65 75.65 118.8 96.47 113.4 109.5 95.86 |
| Essay     | 1.275 1.905 1.641 1.607 2.715 3 2.188 2.634 1.115 2.484 2.931 2.177 2.14 |

In addition, it can be also inferred in the centroid table of 2015 that in the multiple choice type of exam, learners belong to Cluster 1 had mediate scores and they excel more in strand 1 which is Communication Skills compared to other clusters but they got a very low performance in strand 5 which is Expanding One’s World Vision. This group is called average standard learners. Learners who belong to cluster 0 had the best performance in four strands except strand 1 which is lower compared to cluster 1. Learners who belong to this group are called good standard learners. On the other hand, cluster 2 represents learners who had worst performance compared to the two clusters and can be labelled as poor standard learners. It is also noted that all clusters did not perform well in the essay type of exam.

Furthermore, from the centroid table of A & E test result of 2016 it shows that learners who belong to cluster 0 had the highest score and can be labelled as good standard learners. Cluster 1 is the least performing group and can be labelled as poor standard learners and the learners who belong to cluster 2 can be labelled as average standard learners. It can also be observed that in the multiple choice type of exam all clusters excelled in strand 5 which is the Sense of Community and Expanding One’s World Vision but they all got low score in strand 4 which is the Development of Self. Essay type of exam is still the hardest part of the exam as all clusters got low score compared to the multiple type of exam that covers the five learning strands.

3.3. Prediction

Naïve Bayes and Decision tree algorithms were used to predict the performance of learners on the 2016 A & E test. The 2014 and 2015 A & E test results were used as the training data set in the study. The performance rate of the two supervised learning techniques was tested to determine which is better to use in predicting the performance of learners. It has been noted that the Naïve Bayes is more accurate than the decision tree as the accuracy rate is 94.83% compared to 87.44% accuracy rate of decision tree. When it comes to precision, decision tree is more precise with 97.73% compared to 91.25% precision rate of Naïve Bayes but in the recall it can be observed that Naïve Bayes got a perfect rate of 100% compared to the 78.54% of decision tree. From this performance test, it shows that naïve bayes is better to use when it comes to the prediction of performance of the ALS learners in the A & E test compared to the decision tree. Prediction of performance of learners is necessary in order for the facilitators to identify the learning techniques and teaching strategies they will use to improve the learners’ performance. This is also to help the facilitators to determine learners that need more attention for them to improve.

The result of the study supported the research entitled Educational Data Mining to Reduce Student Dropout Rate by Using Classification wherein the aimed was to evaluate student performance by using different decision tree and bayes algorithms. This help them to identify the weak students having enrollment status at risk and students needing further help. It was found out that naïve bayes is the best algorithm as it gets the highest accuracy rate which is above ninety percent compared to other algorithms used in the study. Table 5 below is the performance rate of the two algorithms [8].

Table 5. Performance Rate

|             | Naïve Bayes   | Decision Tree |
|-------------|---------------|---------------|
| Accuracy    | 94.83%        | 87.44%        |
| Precision   | 91.25%        | 97.73%        |
| Recall      | 100%          | 78.54%        |
4. Conclusion
The percentage of passers significantly increased every year as a new set of A & E takers (ALS learners) takes the exam from 2014 to 2016. This is a reflection of some form of intervention by the ALS facilitators because teacher factor cannot be undervalued. Considering that the results on essay being consistently the attribute in which most learners fail every year, this could be associated with their weak foundation in reading and writing. In the same manner that strand 4 or Development of Self and a Sense of Community is also consistent in being the strand in which the learners performed the least among the strands alone excluding essay. This may be associated to their lack of self and social awareness and participation. Naïve Bayes being the better supervised learning technique for prediction as tested in this study may be linked with the fact that it is based on the theory of probability and not on the divide and conquer technique as being used in the decision tree. Thus, the exploration of ALS A & E test results is very crucial in determining the performance of ALS learner on the exam so proper application of data mining techniques shall be used to achieve enhancement in the exam.

References
[1] Romero C and Ventura S 2013 Data mining in education Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3(1) pp 12-27
[2] Berkhin P 2006 A Survey of clustering data mining techniques In Grouping Multidimensional Data (Heidelberg: Springer Berlin) pp 25-71.
[3] Al-Barrak M A and Al-Razgan M S 2015 Predicting students’ performance through classification: a case study J. of Theoretical & Applied Information Technol. 75(2).
[4] Oyelade O J, Oladipupo O O and Obagbuwa I C 2010 Application of k means clustering algorithm for prediction of students academic performance arXiv (preprint arXiv:1002.2425).
[5] Shovon M, Islam H and Haque M 2012 An approach of improving students academic performance by using k means clustering algorithm and decision tree arXiv (preprint arXiv:1211.6340).
[6] Baradwaj B K and Pal S 2012 Mining educational data to analyze students’ performance arXiv (preprint arXiv:1201.3417).
[7] Dar C 2016 Baguio City posts 82% passing rate in 2016 A & E exam.
[8] Shakeel K and Butt N A Educational data mining to reduce student dropout rate by using classification.