SEUSL

Sri Lankan Journal of Technology

Research

SLJoT
2022 Sp Issue 55 - 61
ISSN 2773-6970

A Study for Determining Students' Academic Performance Based on Their Activities Using Clustering Approaches

M.P.R.I.R. Silva[1],R.A.H.M. Rupasingha[2*] and B.T.G.S. Kumara[3]

[1]Department of Information Technology, Sabaragamuwa University of Sri Lanka, Sri Lanka
[2]Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Sri Lanka
[3]Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, Sri Lanka

[*]Corresponding Author: hmrupasingha@gmail.com || ORCID: 0000-0003-3922-4290

Received: 10-10-2022 * Accepted: 15-11-2022 * Published Online: 30-11-2022

***Abstract*—In a country, education depends on the success of educational institutions as well as the abilities of the students. In higher education institutions, the grade point average is used to evaluate students' performance. On the other hand, students participate in extracurricular and academic activities. It is important to determine the relationship between students' final performance and their other activities. As a solution, we tested this on students who graduated from the Sabaragamuwa University of Sri Lanka. By collecting data using a Google form and preprocessing the data set, the results are generated using unsupervised machine learning approaches as three clustering algorithms, namely Hierarchical clustering, Simple k-means, and Expectation Maximum technique. The Simple k-means approach effectively classified data with high accuracy, precision, recall, and f-measure values using Waikato Environment for Knowledge Analysis (WEKA) data mining tool. Furthermore, error values represent the lowest error value when using the Simple k-means method. Based on the students' final results, the data set is divided into five categories. The research findings show that the Simple k-means clustering technique performed than the other two algorithms. The study's future plans include continuing to involve students from schools and comparing results by various classification methods.**

***Keywords*—**Clustering, Machine learning, Academic activities, Non-academic activities, Students' performance

I. Introduction

Today Grade Point Average (GPA) is used in the university system to determine the students' academic performance. Students are faced with lots of exams during their university period. They must complete the continuous assessments, quizzes, practical sessions, and final tests and determine their GPA based on the results of these exams and the credit given to specific subjects. In addition to academic activities, students engage in various extracurricular activities, and each student's behavior regarding academic matters and their and hobbies, non-academic activities are different. Final results are affected not only by exam results but also by the different activities students participate in. Student performance is an important consideration for all educational institutions. As a result, students' accomplishment evaluations must be supported for the institutions to get a higher grade. This might be accomplished by data mining into a database of educational data on current student performance. Universities place a high value on predicting and grouping student performance. It supports students and instructors in doing their activities more efficiently and successfully. Data mining is a phase in the Knowledge Discovery Databases process. It is done to extract useful information from the data that has been collected. If it is equipped with a variety of efficient data mining techniques, data mining may discover many different types of information from acquired data. Findings from data mining might also uncover sensitive information (Peng, Chen and Zhou, 2009). Higher education organization uses Educational Data Mining to make forecasts and predict student success. Institutions can focus on what to educate and how to teach to fulfill the aims and goals of both the institution and the students. Educators can also carefully observe students individually or in groups. Students can also enhance their learning activities and behaviors (Shahiri, Husain and Rashid, 2015). To solve this, we proposed a clustering approach for clustering the students based on their final results and academic and non-academic activities. We gather information on graduate students' extracurricular activities, behaviors, hobbies, and results. Then under the data cleaning process, there were several attributes were removed. Because those attributes a common low effect on the clustering process. After finishing the data preprocessing, the data is clustered using the unsupervised machine learning algorithms; Expectation-Maximization (EM) clustering, hierarchical clustering, and

Simple k-means clustering. A better clustering performance is shown by giving the high precision, recall, f-measure, and accuracy values and lower Mean Squared Error (MAE) and Root Mean Squared Error (RMAE) error values on the Simple k-means clustering. The rest of this paper is organized as follows. In Section II discusses the materials and method. Section III includes the Results and Discussion, and Section V includes the Conclusion of the paper and discussion of future implications.

A. Clustering

All of the previous research data mining approaches have one thing in common. It is the automatic discovery of new relationships and dependencies in data collection. Clustering is an example of unsupervised learning, a sort of exploratory data analysis in which no labeled data is supplied. Clustering's primary function is to partition an unlabeled data set into a restricted range of natural and hidden data structures (Gera and Goel, 2015). A cluster is a collection of things that are similar among themselves but different from the objects in another cluster. The unsupervised categorization of patterns into groups is known as clustering (Verma *et al*., 2012). Clustering is a Machine Learning technique that includes grouping together data elements. Clustering categorizes data into groups based on their values, characteristics, similarities, and differences (Thamarai Selvi and Sridevi, 2019). According to (D. Karthikeyan, C.G. Saravanan, 2013), there are three clustering techniques: hierarchical approaches, partitioning methods, and density-based methods. Clustering is a technique for grouping unlabeled data items so that objects from one cluster are not comparable to those from another. It is the most fundamental and critical unsupervised learning approach in Data Mining. There are different clustering algorithms such as Simple k-means, EM, and hierarchical cluster algorithms. The most well-known and often used clustering technique is the k-means algorithm. The k-means technique in pattern recognition and unsupervised machine learning approach to grouping data (Sinaga and Yang, 2020). Simple k-means which uses an explicit distance measure to partition the data set into clusters is the most popular used clustering technique. It is built using the partitioning process. It divides n data items into k groups, where k is the number of clusters the user provides. Simple k-means algorithm uses the Euclidean distance measurement. Generally, Simple k-means clustering forms the clusters to a special extent compared to other clustering techniques (Patil, Deshmukh and Rajeswari, 2015). The EM clustering technique allows for the creation of clusters of various forms. Clustering techniques are applied in several sectors such as finance, agriculture, and medicine. A massive amount of data is created. As a result, manually processing such massive amounts of data without the help of computers is extremely difficult and time-consuming (Hamoud, Hashim and Awadh, 2018). A hierarchical algorithm merges or divides existing groups, resulting in a hierarchical structure that mirrors the order in which groups are merged or divided. It is classified into two types, agglomerative and divisive (Verma *et al*., 2012).

B. Clustering Approach to Measuring the Students' Academic Performance

There are some existing researches relevant to the proposed approach. In (Srivastava, 2014) clustering approach is used to assess pupils' academic achievement. It is advised that all of this associated material be sent to the class teacher prior to the completion of the final test. This includes a variety of elements such as internal class marks, GPA, mid and final exams, assignments, and lab–work. The research proposed an ideal technique based on the Simple k-means Clustering algorithm. It used the WEKA tool that allows academics to improve the educational quality of their students. This study (Moubayed *et al*., 2020) suggested using the Simple k-means algorithm to group students into various levels of engagement. This is a small portion of a bigger investigation into the use of predictive data mining models to identify weak students as early as possible in an effort to increase their performance. In this study (Xu, Moon and Van Der Schaar, 2017), researchers presented a novel strategy for estimating students' degree program achievement in the future based on their previous and present performance. To find pertinent courses for building base predictors, a latent component model-based course clustering method was created. They demonstrate that the proposed strategy outperforms benchmark approaches through in-depth simulations on a dataset of undergraduate student data gathered over three years at UCLA. The paper (Durairaj and Vijitha, 2014) offered a method for predicting student academic performance using the simple k-means clustering algorithm. The capacity to track the advancement of students' academic achievement is a key concern for the academic community of higher learning. The study (Islam and Haque, 2012) applied data mining to predict students' activities in a database using Simple k-means clustering. They used only a sample of 70 instances to obtain the results. They expect that the information provided from the data mining and data clustering techniques will be useful to both the instructor and the students. In (Satyanarayana and Ravichandran, 2016) to discover groups of students with similar features, employ Simple k-means clustering with bootstrap averaging. They demonstrated in the clustering area that student data may be grouped into well-defined groups based on their behavioral learning patterns. In the study, (Darcan and Badur, 2012) investigated various student segments for Boaziçi University by doing cluster analysis on several variables of academic ability. (Oyelade, Oladipupo and Obagbuwa, 2010), showed the use of the Simple k-means clustering method in combination with the deterministic model on a data set of private school outcomes with nine courses available for that semester for each student, for a total of 79 individuals. This clustering approach is a suitable standard for tracking the development of students' performance in higher education institutions. It also helps academic planners make better decisions by monitoring

students' performance semester by semester and improving future academic results in future academic sessions.

C. Research Questions and Objectives

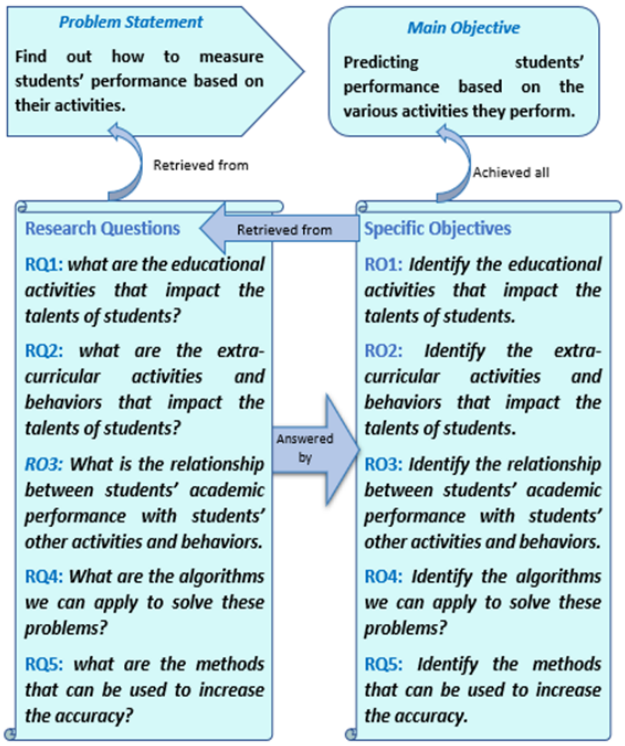

Figure 1: Mapping images of research questions and research objectives

II. Methodology

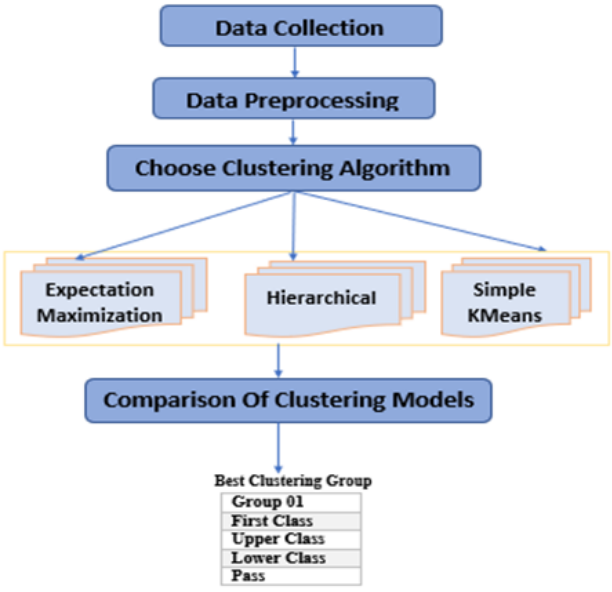

Figure 2: Steps of the proposed approach

Figure 2 explains the steps of the proposed clustering approach using three clustering techniques such as hierarchical, Simple k-means and Expectation Maximum.

A. Data Collection

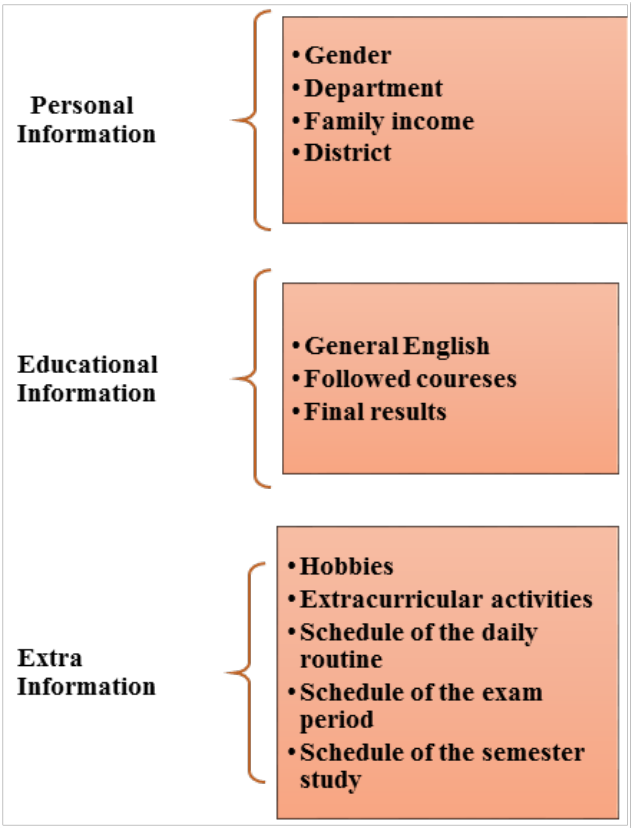

Figure 3: Structure of the Google Form

The data were collected from the graduate students of the Sabaragamuwa University of Sri Lanka. The questions covered the final educational achievements and relevant academic and non-academic activities before graduation. The Google form is created to collect data, and can be answered quickly, covering personal information, educational information, other extracurricular activities, and hobbies, as shown in Figure 3.

B. Preprocessing

Data preprocessing, which is an important part of the data mining process, is done after data collection. In this process, incomplete and irrelevant data is recognized and replaced, altered, or removed. Inconsistent data can produce numerous difficulties and lead to the rendering of false outcomes. The raw data is included low-quality values, and it may be affected by the data mining process. Therefore, data preprocessing plays a major role in the data mining process. Data cleaning, transformation, and reduction are applied to preprocess the data. Figure 4 explains the Sample graphical representation of attributes. In the beginning, the data set contained 24 attributes, but the data was reduced to

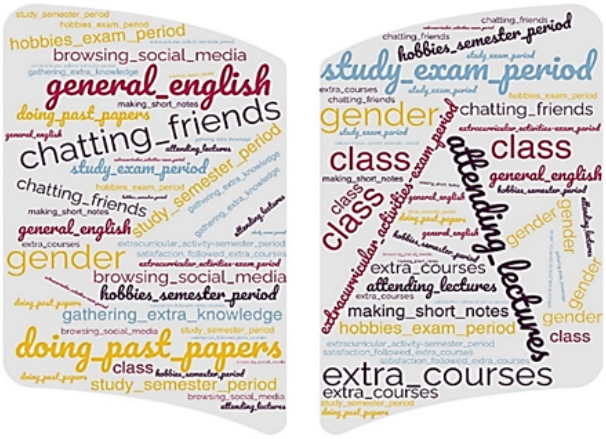

Figure 4: Sample graphical representation of attributes

17, including one dependent variable and other independent variables. Some attributes were removed because they were identified as not important to building a prediction model. The department, district, and A/L stream are removed. And also, hyper-parameter tuning was used in the WEKA data mining tool to rank attributes. Monthly family income and degree type are shown in low ranks. Therefore, the top 17 attributes are selected from the ranked list for the clustering process as the following Table1.

C. Applying Clustering Techniques

After completing the data preprocessing, the WEKA 3.8.5 data mining tool is applied for the clustering and evaluation process. The CSV file format is used to store the input data set. The pattern categorized the records in a database into separate categories. The attributes of the groups in the same group are comparable. Differences between groups should be as large as feasible, but differences within the same group should be as little as possible. This study uses Simple k-means, hierarchical cluster, and EM as clustering algorithms. It is evaluated using the Euclidean Distance function on the full training data set. The preprocessed data set is modified and used for the clustering process. Attributes are ranked, and the top ten attributes are selected for clustering. These changes are made to facilitate clustering and group data into different clusters.

III. Result and Discussion

The precision, recall, and f-measure calculations presented in (1), (2), and (3) are used to compare the evaluation results. These are used to measure the validity of the results. (tp – true positive, FP – false positive, fn – false negative).

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$F\,measure = \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

A. Evaluation Results of Precision, Recall and F-measure

As shown in the following Figure 5, Figure 6, and Figure 7, the highest precision, recall, and f-measure values are indicated in cluster 4 of Simple k-means clustering, respectively. Cluster 04 represents three cluster groups (First+Upper, Lower, Pass).

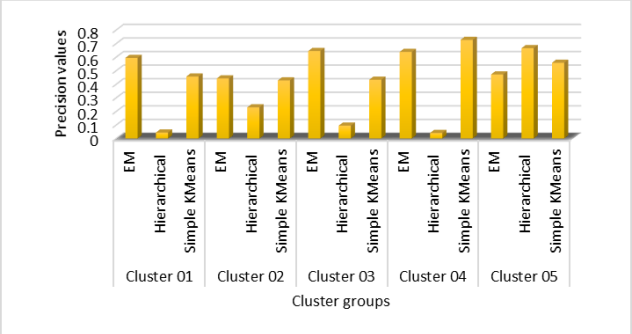

Figure 5: Evaluation results of precision

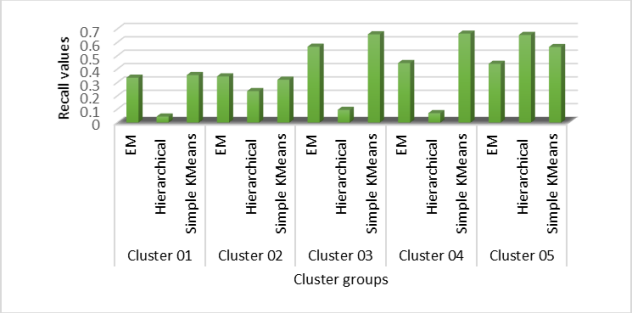

Figure 6: Evaluation results of the recall

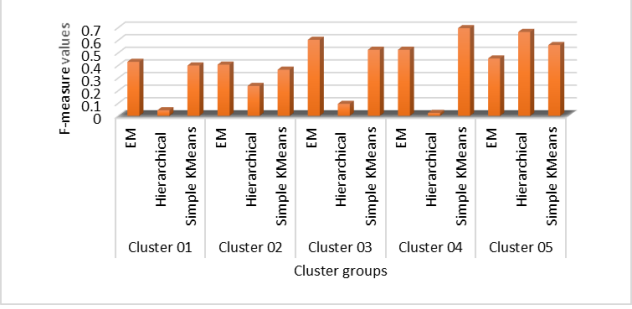

Figure 7: Evaluation results of f-measure

The lowest error increases the algorithm's accuracy based on the MAE and RMSE. These two requirements are represented in (4) and (5). Here, n denotes the sample size, Y_j is the actual value, and X_j denotes the predicted value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_j - X_j| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_j - X_j)^2} \tag{5}$$

Table I: Details of the attributes

	Attribute Name	Description		Attribute Name	Description
1	Hob_Sem	Doing hobbies during the semester (hours)	10	Short_Notes	Short notes prepared by the student in addition to the note given by the lecturer (5 levels from strongly agree to strongly disagree)
2	Extra_Know	In addition to the lectures how to gather extra knowledge (hours)	11	Study_Exam	Time spent doing academic activities during the exam period (hours)
3	Past_Ppr	Time spent doing past papers during the examination period or Semester. (hours)	12	Study_Sem	Time spent doing academic activity during the semester(hours)
4	Social_media	Time spent browsing social media (hours)	13	ExtraActivities_Exam	Time spent doing extracurricular activity during the exam period(hours)
5	Chat_Friends	Time spent doing with friends (hours)	14	Extra_courses	Additional courses followed before and after entering the university (yes, no)
6	English	Results of the GCE Advanced Level Exam for General English (A,B,C,S,W)	15	Satisfaction_Courses	Satisfaction with academic activities through the followed courses (5 levels from 0 to 5)
7	ExtraActivity_Sem	Time spent doing extracurricular activity during the semester(hours)	16	Attend_Lecture	Level of participating lectures (regularly, normal, never)
8	Hob_Exam	Time spent doing hobbies during the exam period (hours)	17	Final_Class (Target variabl)e	First Class, Second Upper Class, Second Lower Class and Pass
9	Gender	Female or Male			

IV. Evaluation Results of Error Values

According to Figure 8 and Figure 9, MAE and RMSE values indicate the lowest value with the Simple k-means algorithm in cluster number 04 data set.

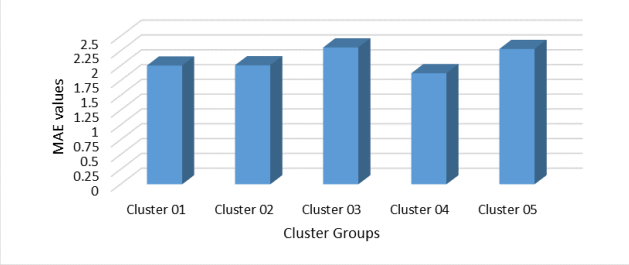

Figure 8: Evaluation results of MAE

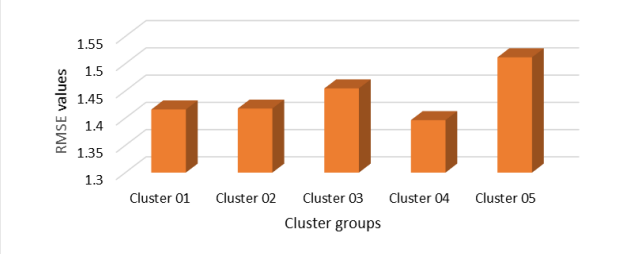

Figure 9: Evaluation results of RMSE

According to the following Table 2, the hierarchical cluster method shows the highest accuracy. When considering the Simple k-means clustering algorithms, the highest accuracy is in the cluster 04 data set.

A. *Evaluation of Accuracy*

The evaluation results were obtained using Microsoft Windows 10 on a PC with an Intel® Core (TM) i3-5005U CPU @ 2.00GHz and 4.0GB of RAM. The WEKA 3.8.5 data mining tool is used for the data set. The 200 data points acquired via a Google form from graduates in the Sabaragamuwa

Table II: Accuracy of the result

Data set	EM	Hierarchical	Simple k-means
Cluster 01	54%	78.5%	52%
Cluster 02	38%	51.5%	41%
Cluster 03	53.5%	55%	66%
Cluster 04	53.5%	44%	66.5%
Cluster 05	58%	66.5%	56.5%

University of Sri Lanka were employed as a data set for the evaluation procedure. The data set is modified into five groups based on the final results (class) obtained by the students. They are cluster 01 (First, Upper, Lower, Pass), cluster 02 (First, Upper, Lower+Pass), cluster 03 (First, Upper+Lower, Pass), cluster 04 (First+Upper, Lower, Pass), and cluster 05 (First+Upper, Lower+Pass). The “classes to cluster evaluation” is used to evaluate the output. Three clustering algorithms were used for the comparison: simple k-mean, EM, and hierarchical clustering. The proposed approach first identified the students’ academic activities, extracurricular activities, daily routine schedule and behaviors etc. The time duration, levels and frequency which they spend on the above activities were also obtained. It is a one of the objectives of this study. For that the information was obtained from the graduated students of the Sabaragamuwa University of Sri Lanka. The information required for the study has been collected using Google Form, which is a way that can easily reach the respondent. Another main objective is identifying the relationship between academic performance and their other activities. For that we proposed a clustering approach. After the preprocessing, we applied different clustering algorithms to find a best method. Hierarchical clustering, Simple k-means, and Expectation Maximum are applied for comparison in this study. The Simple k-means clustering of cluster-04 (First+Upper, Lower, Pass) dataset indicates a 66.5% accuracy out of five Simple k-means data sets. According to the entire data set, the Hierarchical cluster is shown at 78.5% as the highest value. The overall accuracy performance of the

Simple k-means cluster algorithm is better than the other two algorithms. Furthermore, minimum MSE and RMSE values are indicated by the Simple k-means cluster in the cluster number 04 data set. When considering the precision, recall, and f-measure, the Simple k-means of cluster 04 group shows the best grouping results. Precision shows 0.665; Recall is 0.727, and F-measure represents 0.695. According to the results, the Simple k-means clustering algorithm shows better results than the other two algorithms in terms of accuracy, precision, recall, f-measure MSE, and RMSE values. Cluster number 04, divided as First+Upper, Lower, and Pass, indicate the best clustering results. As a result of this study, higher education institutions will be able to detect student performance, identify weak students, and take the necessary actions to encourage them to achieve a good final result with the highest final class. This may be applied not only in the universities but also in schools and tuition classes. In future work, this study is planned to evaluate with more clustering algorithms and supervised learning algorithms techniques such as classification. And also planned to enhance the data set and the factors that affected students' future goals.

REFERENCES

D. Karthikeyan, C.G. Saravanan, E. J. G. (2013) 'P Erformance a Nalysis of T Errestrial', *International Journal of Advances in Engineering, 5*(6), pp. 55–64.

Darcan, O. and Badur, B. (2012) 'Student Profiling on Academic Performance Using Cluster Analysis', *Journal of e-Learning Higher Education, 2012*, pp. 1–8. doi: 10.5171/2012.622480.

Durairaj, M. and Vijitha, C. (2014) 'Educational Data mining for Prediction of Student Performance Using Clustering Algorithms', *International Journal of Computer Science and Information Technologies, 5*(4), pp. 5987–5991.

Gera, M. and Goel, S. (2015) 'Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity', *International Journal of Computer Applications, 113*(18), pp. 22–29. doi: 10.5120/19926-2042.

Hamoud, A. K., Hashim, A. S. and Awadh, W. A. (2018) 'Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis', *International Journal of Interactive Multimedia and Artificial Intelligence, 5*(2), p. 26. doi: 10.9781/ijimai.2018.02.004.

Islam, H. and Haque, M. (2012) 'An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree', *International Journal of Advanced Computer Science and Applications, 3*(8), pp. 146–149. doi: 10.14569/ijacsa.2012.030824.

Moubayed, A. *et al*. (2020) 'Student Engagement Level in an e-Learning Environment: Clustering Using K-means', *American Journal of Distance Education, 34*(2), pp. 137–156. doi: 10.1080/08923647.2020.1696140.

Oyelade, O. J., Oladipupo, O. O. and Obagbuwa, I. C. (2010) 'Application of k Means Clustering algorithm for prediction of Students Academic Performance', 7, pp. 292–295. Available at: http://arxiv.org/abs/1002.2425.

Patil, R., Deshmukh, S. and Rajeswari, K. (2015) 'Analysis of SimpleKMeans with Multiple Dimensions using WEKA', *International Journal of Computer Applications, 110*(1), pp. 14–17. doi: 10.5120/19280-0694.

Peng, W., Chen, J. and Zhou, H. (2009) 'An Implementation of IDE3 Decision Tree Learning Algorithm', *Project of Comp 9417: Machine Learning, 1*, pp. 1–20. Available at: http://cis.k.hosei.ac.jp/ rhuang/Miccl/AI-2/L10-src/DecisionTree2.pdf.

Satyanarayana, A. and Ravichandran, G. (2016) 'Mining student data by ensemble classification and clustering for profiling and prediction of student academic performance', American Society for Engineering .... Available at: https://www.hofstra.edu/pdf/academics/colleges/seas/asee-fall-2016/asee-midatlantic-f2016-satyanarayana.pdf.

Shahiri, A. M., Husain, W. and Rashid, N. A. (2015) 'A Review on Predicting Student's Performance Using Data Mining Techniques', *Procedia Computer Science, 72*, pp. 414–422. doi: 10.1016/j.procs.2015.12.157.

Sinaga, K. P. and Yang, M. (2020) 'Unsupervised K-Means Clustering Algorithm', 8. doi: 10.1109/ACCESS.2020.2988796.

Srivastava, S. (2014) 'Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining', *International Journal of Computer Applications, 88*(10), pp. 26–29. doi: 10.5120/15389-3809.

Thamarai Selvi, K. and Sridevi, R. (2019) 'Clustering Techniques based Student Performance Analysis', *The International journal of analytical and experimental modal analysis, 11*(8), pp. 1415–1422.

Verma, M. *et al*. (2012) 'A Comparative Study of Various Clustering Algorithms in Data Mining Manish Verma , Mauly Srivastava , Neha Chack , Atul Kumar Diswar , Nidhi Gupta', 2(3), pp. 1379–1384. Xu, J., Moon, K. H. and Van Der Schaar, M. (2017) 'A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs', *IEEE Journal on Selected Topics in Signal Processing, 11*(5), pp. 742–753. doi: 10.1109/JSTSP.2017.2692560.