International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 680-687.
www.theijire.com ***ISSN No: 2582-8746***

Examining Successful Attributes for Undergraduate Using Machine Learning Techniques

S.Uma [1], Arul Prashath R[2], Bhavan Ramana E[3], Hemanth Kumar Reddy P[4], Chandru P[5]
[1] Associate professor, Department of Computer Science and Engineering, paavai Engineering College, Namakkal,TN, India.
[2,3,4,5] UG Student, Department of Computer Science and Engineering, paavai Engineering College, Namakkal,TN, India.

How to cite this paper:
S.Uma [1], Arul Prashath R[2], Bhavan Ramana E[3], Hemanth Kumar Reddy P[4], Chandru P[5] , "Examining Successful Attributes for Undergraduate Using Machine Learning Techniques", IJIRE-V4I02-680-687.

https://www.doi.org/10.59256/ijire.2023040244

Abstract: *This study utilizes both supervised and unsupervised machine learning techniques to identify the key attributes that are oftendemonstrated by successful learners in a computer course. Learning an introduction to computers course can be challenging for students. This study aims to explore how successful students regulate their learning in this course. By answering these questions, teachers can gain valuable insights into how students learn and which strategies are most effective for their success. To compare the accuracy, precision, and sensitivity levels of classifiers, this study employed seven supervised machine learning algorithms and ensembles. Additionally, association rule and clustering techniques were utilized to identify the key attributes for successfulstudents. However, it is important to note that the use of a convenience sample in this study may have limited the number of students in each cluster.*

Key Word: *Association rules, Bayesian network (BN), clustering,decision trees (DTs), K-nearest neighbour (KNN), multilayer perceptron (MLP), Naïve Bayes (NB), support vector machines (SVMs)*

I.INTRODUCTION

The use of machine learning (ML) algorithms to gain insights into learners' learning patterns has become popularin the educational community. Many studies employ supervised learning techniques to build learning models that can predict students' performance or identify students who are at risk of falling behind. For instance, Ahadi et al. collected data on students' gender, major, grade average, age, and programming experience, and used Bayesian classifiers, Naïve Bayes (NB), and decision tree (DT)-based classifiers to identify high- and low-performing students in a programming course. Ahadi et al discovered that teachers can easily identify struggling and high- performing students after the first week using Bayesian classifiers, Naïve Bayes (NB), and decision tree (DT)- based classifiers. Similarly, Quille and Bergin developed an ML model to predict student success in an introductory programming course. They trained individual classifiers using ML algorithms such as NB, logistic regression (LR), backpropagation (BP), support vector machine (SVM), DT, and K-nearest neighbour (KNN), and achieved a classification accuracy rate of 80% for their proposed model. In addition, previous research comparing the performance of five supervised learning models in predicting students' final performance was also published. In contrast to supervised learning algorithms, unsupervised learning techniques such as association rule mining have been utilized by some researchers to discover interesting correlations, frequent patterns, associations, and causal relationships among large sets of items in a given dataset. For instance, Hung and Zhang applied association rule mining to investigate the daily learning behaviours and activity patterns of 98 online undergraduate students based on their log files in the learning management systems (LMS). The study revealed that more than 50% of the students' online learning activities involved only reading or accessing course materials.However, once they accessed the course materials, 40.28% of the students would post messages on the discussionboard, and there was a more than 70% probability that they would post again on the same day. Clustering is a valuable technique when the most common attributes within a dataset are not known in advance. For example, a study used clustering to identify at-risk online students. The study employed five variables, including total frequency of accessing course materials, total number of messages posted, total number of messages read, total number of messages replied, and the final grade to classify the characteristics of 509 online students. The clustering results showed that based on different levels of participation and academic performance, a teacher could identify at-risk students from week 10. After reviewing the literature, it was found that some studies used both supervised and unsupervised learning techniques to investigate learners' learning in a single study. Romero et al.aimed to investigate the impact of online discussion forums on the final performance of 114 university students in a computer course. Initially, several classification algorithms were employed to compare accuracy and F- measure values, and J48 and Jrip algorithms demonstrated the best performances. Then, clustering techniques andassociation rule mining were applied to better understand the results. The study revealed that the majority of PASS students composed forum posts with a length exceeding 285 words, while the FAIL students only contributed posts of fewer than 18 words on two separate days throughout the course. The authors claimed that utilizing both supervised and unsupervised learning techniques produced noticeably superior outcomes compared to utilizing just one machine learning approach. Asif et al. utilized pre-university scores and scores from first- and second- year courses to anticipate students' academic performance at the end of a four-year degree program. By implementing a decision tree algorithm, the scores of four courses were utilized as crucial factors to categorize students into distinct groups. After examining the data within each cluster, the authors found that students tendedto exhibit comparable levels of scores (either low, intermediate, or high) across

all courses. The authors asserted that their proposed model had the potential to identify and assist low-achieving students promptly. Amershi and Conati combined supervised and unsupervised machine learning techniques to create a user model that would becost-effective to develop. They utilized data from thirty-six students in a computer-based learning environment, including pre-test and post-test scores on mathematical functions, 3783 interface actions, and gaze data obtained from an eye tracker. To make sense of the vast amount of eye-tracking data, a clustering method was employed during the data pre-processing phase to identify meaningful behavioural patterns. The results of the supervised classification demonstrated that the proposed model had an accuracy of 86.3%. To summarize, the three studies mentioned above aim to differentiate between successful and unsuccessful students using criteria such as performance on discussion forums, academic scores, and eye-tracking data. In contrast, the present study focuseson students' subjective perceptions and learning behaviours to identify the characteristics of successful learners. While self-report scales are commonly used to assess learners' cognitive perceptions and behaviours during learning, there is still much uncertainty surrounding their effectiveness, and more work is needed to refine them. Furthermore, this type of research frequently employs conventional statistical methods to understand students' learning outcomes. To identify predictors of student success, a subscale of a questionnaire, such as self-efficacy, is more valuable than a single item. However, it is essential to determine which element within the subscale is the most important. Traditional statistical analyses have limitations in addressing this issue. In contrast, machine learning techniques can overcome this problem and extract critical attributes from subscales more effectively. The present study employs self-regulation theory as a framework and employs both supervised and unsupervised learning techniques to investigate which attributes are essential for student learning. This approach aims to obtain meaningful insights into student learning and enhance the interpretation of the findings. Section II provides a review of the relevant self-regulation literature related to this study. Section III presents the machine learning algorithms utilized in this research. Section IV outlines the research questions, participant selection procedure, and data gathering process. Additionally, the questionnaire content is briefly described in this section. Section V provides answers to the research questions. Section VI summarizes and discusses the research findings, along with its limitations. Finally, Section VII presents the conclusions of the study.

II.REVIEW OF SELF-REGULATION ATTRIBUTES

The goal of this study is to identify key attributes that contribute to a student's success. Four potential attributes that are believed to be important for successful learners are self-efficacy, metacognitive self- regulation, time and study environment management, and computer self-efficacy. These attributes willbe briefly described in the study.

A. Self-Efficacy

The concept of self-efficacy was originally introduced by Bandura and it refers to an individual's beliefin their ability to perform a specific action to achieve a desired outcome. Studies have shown that students with high levels of self-efficacy are generally more proactive and utilize more effective self- regulatory strategies to achieve their goals, compared to those with lower self-efficacy

B. Metacognitive Self-Regulation

Metacognition refers to an individual's ability to regulate their task and performance by utilizing self- instructions, while cognition is the means by which these self-instructions are carried out. The concept of metacognition essentially involves understanding which learning strategies are utilized during the process of learning. Important components of metacognition include self-monitoring, self-assessment,and self-evaluation.

C. Time and Study Environment Management

Time and study environment management refer to the strategies used by students to organize their studyenvironments and manage their time efficiently for effective learning. In a study conducted by Kitsantaset al., the MSLQ scale was utilized to predict the academic performance of 198 first-year students. The findings of the study indicated that only time and study environment management skills were able to predict the students' GPA one year later.

D. Computer Self-Efficacy

Computer self-efficacy refers to an individual's level of confidence and belief in their abilities and knowledge of computer skills, similar to the concept of self-efficacy. However, computer self-efficacy specifically focuses on learners' perceptions and capabilities related to computer technology.

III.MACHINE LEARNING TECHNIQUES

The study employs a variety of machine learning (ML) techniques, including supervised and unsupervised learning, to classify learners into pass or fail groups and identify attributes associated with successful learning. Supervised learning algorithms used include decision tree (DT), Bayesian network(BN), logistic regression (LR), K-nearest neighbours (KNN), Naive Bayes (NB), support vector machine (SVM), and multilayer perceptron (MLP). In addition, association rule mining and clusteringtechniques, which are unsupervised learning algorithms, are also used.

The study utilizes the Waikato environment for knowledge (WEKA) software to construct the learningmodels and assess their performance. The algorithms used are explained below.

A. Decision Tree

The decision tree (DT) algorithm has two stages: construction and classification. In the construction stage, the

algorithm calculates entropies for all relevant variables to determine split values for dividing samples into groups. These variables are chosen in sequence based on their entropies, using a top-downapproach to construct the DT classifier. One well-known algorithm for building DTs is C4.5.

B. Bayesian Network

In this study, the researchers employ a Bayesian network (BN), which is a type of graphical model. TheBN is used to represent a set of conditional probability variables. Each variable is depicted as a node inthe graph, with links connecting nodes to show the conditional relationships between the variables.

C. Naïve Bayes

The Naive Bayes (NB) classifier is a simple version of the Bayesian network (BN) model. Its advantage is a fast training time, assuming all features are independent of one another. However, this assumption is unrealistic in real-world scenarios. To achieve independence between features, a useful approach is to employ feature selection strategies.

D. Support Vector Machine

The Support Vector Machine (SVM) algorithm aims to build a hyperplane that can separate two classesas much as possible by adopting a small number of crucial boundary instances, known as support vectors. Each side of the hyperplane corresponds to a different class. Using an SVM model, it is possible to predict the class of a new instance by determining which side of the hyperplane it falls on.

E. Multilayer Perceptron

The Multilayer Perceptron (MLP) algorithm consists of an activation function and three types of units:input units, output units, and hidden units. Input units receive information to be processed, output unitsdisplay the learning outcomes, and hidden units act as filters to identify real patterns.MLP is a type of feedforward artificial neural network (ANN) that allows signals to pass from input to output. The Backpropagation (BP) algorithm is a commonly used technique to train MLP.

F. Logistic Regression

The Logistic Regression (LR) algorithm creates a discriminative classifier to distinguish the outcome value into one of two classes. The prediction probability of a logistic regression model ranges between 0 and 1.

G. K-Nearest Neighbour

The K-Nearest Neighbors (KNN) algorithm is a widely used learning technique that classifies an instance based on the instances that are similar to it. It is straightforward to implement and has a short calculation time.

H. Association Rule

Association rule mining is a technique used to discover relationships among variables in a dataset, withthe aim of finding If-Then rules. However, not all rules may be useful in comprehending the dataset. Toaddress this, minimum support and confidence values are set as thresholds to discard uninteresting or useless rules.

I. Clustering

K-means clustering is a widely used unsupervised learning technique that groups similar data points into clusters. The centre of each cluster is assigned by taking the mean of all data points within that cluster.

IV.METHODOLOGY

The research questions, the study settings, subject description, procedure of data pre-processing, instrumentation introduction, and examination for internal consistency of reliability are showed as follows.

A. Research Questions

What is the most suitable supervised learning algorithm for predicting students' final performance?Which attributes are critical to a student's success in the course?

B. Subjects

A total of 215 first-year university students participated in this study. The university is located in aMidwest city of Taiwan. All first-year students must take the “Introduction to Computers” course.

C. Settings

The course had a total of four periods per week, consisting of two face-to-face sessions and two computer lab sessions. The face-to-face sessions covered various topics related to computers, while thecomputer lab sessions focused on teaching students how to use Microsoft Office packages.

D. Instrumentation

The survey instrument used in this study was the Student Learning Questionnaire (SLQ), which consistsof 64 items divided into five sections. The first section contains demographic items, while the second measures students' motivational beliefs and the third evaluates their metacognitive learning strategies. The fourth section assesses students' abilities in time

and study environment management, while the fifth section measures their computer motivational beliefs. These five subscales aim to evaluate various aspects of students' academic self-regulation. The self-efficacy, metacognition, and time management scales were selected from the MSLQ, while the computer self-efficacy subscale was adopted from the Computer Self-Efficacy Scale.

In addition to the SLQ, the proposed model includes the average of ten weekly assignment scores. The final score of a student is calculated based on two paper-and-pencil tests that assess their understanding of key computer concepts and one online test that examines their computer skills acquired in the computer labs.

E. Data Pre-processing

The study collected 215 questionnaires at the end of the semester before the final examination. After checking for missing values, a total of 136 subjects were included in the analysis, consisting of 60 males and 76 females. It is important to note that this study used a convenience sample, which may limit the generalizability of the findings. The students' final grades were obtained from the instructor and were used for further analysis with ML techniques.

F. Internal Consistency Reliability

The internal consistency of reliability for each subscale was assessed before conducting further analysis. The Cronbach's alphas for the subscale scores are presented in Table I, and they are all within acceptable levels, typically equal to or greater than 0.70. The reliability analyses indicated that the four measures utilized in the study were highly reliable.

V.EXPERIMENTAL RESULTS

The procedure for constructing the learning model is expressed as follows.

A. Attribute Selection

In the first section of the SLQ, three variables, namely class, gender, and computer experience, were selected to be kept for analysis as they may have an impact on the final academic outcomes. Other variables were excluded as they were outside the scope of this research. To determine which attributes should be included in the ML model, an attribute selection procedure was used. The CfsSubsetEval approach selects attributes that have a high relationship with the class but low correlation with each other. The Info Gain Attribute Eval method uses the concept of information gain to identify appropriate features.

Table I

CRONBACH'S ALPHAS FOR FOUR SUBSCALES ON THE SLQ, ^{N} = 136

Scale	Pintrich *et al.* (MSQL) 29	This study
Self-Efficacy for Learning and Performance	.93	.95
Metacognitive Self-Regulation	.79	.84
Time and Study Environment Management	.76	.72
	Cassidy & Eachus [30] (Computer Self-Efficacy Scale)	
Computer Self-Efficacy	.97	.93

Table II shows that both approaches selected the same attributes, with only a difference in their order. In later experiments, all selected attributes were input into the same ML model to compare the performance of classifiers.

Table II

Method of Feature Selection	No of Attributes Selected	Selected Attributes and Its Order
CfsSubsetEval	9	weekly assignments, sef01, sef05, sef08, mcog01, mcog09, time05, csef10, csef26
InforGainAttributeEva	9	sef05, sef01, sef08, time05, weekly assignments, csef10, csef26, mcog01, mcog09

Table III provides a description of each attribute based on the MSLQ and Computer Self-Efficacy Scale.The other two attributes, weekly assignment score and final outcome, were included as part of the factors in the learning model in this study.

B. Measurement Techniques

To assess the performance of the classifiers, accuracy, sensitivity (true positive rate), and specificity (true negative rate) were used. The confusion matrix, shown in Table IV, displays the results of the classification. True positives (TP) and true negatives (TN) indicate correct classifications. A false positive (FP) occurs when the prediction is inaccurately forecasted as true, but it is actually false. A false negative (FN) occurs when the prediction indicates false, but it is actually true.. As expressed in (1), the overall accuracy is the number of correct classifications, i.e., TP + TN, over the total number of

classification as TP and TN.

C. Training Procedure

The study used ten-fold stratified cross-validation for both Experiments 1 and 2, which means that the dataset was divided into ten subsets of equal size. Each subset was used as both a training and testing dataset when building a classifier, in order to reduce bias and increase the reliability of the classification models.

D. Experiments for supervised learning

1) *Experiment 1 (Research Question 1):* The first experiment aimed to answer Research Question 1 and used nine attributes selected by the CfsSubsetEval method to test seven ML algorithms.

Table II

$$\text{Accuracy} = \overline{\text{TP} + \text{TN}++ \text{FP} + \text{FN}}. \quad (1)$$

Sensitivity is defined in

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}). \quad (2)$$

Specificity is defined in

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}). \quad (3)$$

DESCRIPTION OF ALL ATTRIBUTES

Attribute	Attribute Content	Measure Level
sef01	I believe I will receive an excellent grade in this class	1 (strongly disagree) 2 3 4 5 (Strongly Agree)
sef05	I'm confident I can do an excellent job on the assignments and tests in this course	1 (strongly disagree) 2 3 4 5 (Strongly Agree)
sef08	Considering the difficulty of this course, the teacher, and my skills, I think I will do well in this class	1 (not at all true of me) 2 3 4 5 (very true of me)
mcog01	During class time I often miss important points because I'm thinking of other things (REVERSED)	1 (not at all true of me) 2 3 4 5 (very true of me)
mcog09	I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over when studying.	1 (not at all true of me) 2 3 4 5 (very true of me)
time05	I make sure I keep up with the weekly readings and assignments for this course	1 (not at all true of me) 2 3 4 5 (very true of me)
csef10	I often have difficulties when trying to learn how to use a new computer package (REVERSED)	1 (strongly disagree) 2 3 4 5 (strongly agree)
csef26	As far as computers go, I don't consider myself to be very competent (REVERSED)	1 (strongly disagree) 2 3 4 5 (strongly agree)
weekly assignment score	The average of 10 weekly assignments	low, middle, high
final outcome	The average of two paper-and-pencil tests and one online computer skills test	pass, fail

TABLE IV

Sample Confusion Matrix

The results indicated that the NB algorithm had the highest accuracy at 83.26%, followed by BN (82.37%), LR (81.33%), SVM (81.08%), MLP (78.68%), KNN (74.55%), and C4.5 (73.79%), as shown

in Table V. In terms of sensitivity, NB had the highest value followed by LR, BN, SVM, MLP, KNN, and C4.5. The specificity values were ranked in descending order as SVM, NB, MLP, BN, LR, KNN, and C4.5, with KNN having the lowest score at 66.53%.

1) Experiment 2 (Research Question 2): The second experiment used the same seven ML algorithms as Experiment 1 but employed nine attributes selected by the In for Gain Attribute Eval method. Surprisingly, the results showed that the accuracy, sensitivity, and specificity were the same as those presented in Table V for Experiment 1.
2) Experiment 3: After conducting basic supervised learning algorithms in Experiments 1 and 2, Experiment 3 used ensemble algorithms, including Vote, Random Forest, Bagging, and AdaBoostM1. Table VI shows the descending order of accuracies of the four classifiers: Vote (84.06%), Random Forest and Bagging at the same level (82.28%), and AdaBoostM1 (80.16%).

The sensitivities of the Vote, Random Forest, and Bagging are all 93.18%, which is the highest, while AdaBoostM1 is 79.25%. The best specificity is Vote (78.48%), followed by AdaBoostM1 (70.67%), and then Random Forest and Bagging at the same level (60.78%). The Vote ensemble inthis study combines NB, BN, and SVM classifiers, which demonstrates the best performance in terms of accuracy, sensitivity, and specificity compared to the other classifiers and ensembles.

E. Experiment for Unsupervised learning

In Experiment 5, the minimum support was set to 0.1, minimum confidence was 0.5, and the number of rules was limited to 150. The top ten rules generated from the association rule mining are presentedin Table VII. Out of 136 students, 82 were labelled as "pass" because their final grades were 60 or higher, while the remaining 54 were labelled as "fail" due to their final grades being lower than 60.

Table VII

Rule #	Explanation	No of Students
1	if a student has good ability to keep up with the weekly schedule (time05 = 4) and receives at least "middle" level of the grade on weekly assignments, he/she will pass the course	26
2	If a student has a middle level of belief that he/she will receive an excellent grade in this class (sef01 = 3) and has good ability to keep up with the weekly schedule (time05 = 4), he/she will pass.	24
3	If a student has a middle level of belief that he/she will receive an excellent grade in this class (sef01 = 3), with a middle level of belief that he/she will do well in this class (sef08 = 3), and has good ability to keep up with the weekly schedule (time05 = 4), he/she will pass.	22

Explanations Of Top Three Pass Rules Table VII shows that time05 appears in almost every top ten rule except rule 4. Self-efficacy beliefs such as sef01 are present in six of the top ten rules (rules 2, 3, 4, 5, 8, and 9), sef08 appears in three rules (rules 3, 6, and 9), and sef05 appears once (rule 8). In total, the relevant self-efficacy beliefs appear 10 times in the top ten rules. Moreover, time05 achieves "high"standards (= 4) in all the rules it appears in, while sef01 presents a high degree (= 4) in rule 4. The otherself-efficacy beliefs are all at the "middle" level.

TOP TEN RULES OF PASS STUDENTS

Rule #	Rule Description	Confidence Value	No of Students
1	weekly assignments = middle , time05 = 4 → pass	1	26
2	sef01 = 3, time05 = 4 → pass	1	24
3	sef01 = 3, sef08 = 3, time05 = 4 → pass	1	22
4	sef01 = 4 → pass	1	21
5	weekly assignments = middle, sef01 = 3, time05 = 4 → pass	1	18
6	sef08 = 3, time05 = 4 → pass	1	21
7	mcog01 = 3, time05 = 4 → pass	1	19
8	sef01 = 3, sef05 = 3, time05 = 4 → pass	1	15
9	sef01 = 3, sef08 = 3, time05 = 4 → pass	1	15
10	time05 = 4, csef26 = 4 → pass	1	10

TABLE VIII

Table VIII explains the top three rules based on the contents of Table III. The confidence of these top three rules is 1, which means that when students possess these specified features, there is a 100% probability of them passing the course.

In conclusion, the analysis suggests that two key factors for student success are "keeping up with the progress of the class" (time05) and self-efficacy beliefs (sef01, sef08, and sef05). Additionally, out of the 150 rules generated in Experiment 5, only five rules explain the attributes of "fail" students. The common features of the first and second "fail" rules include a "lack of confidence" (sef05 = 2) and a "low" level of weekly assignments. These "fail" students also had a "middle" or lower level of "keepingup with the progress of the class".

Next, the clustering technique is used to divide students as "pass" or "fail" groups to answer some unclear points.

2) Experiment 6-The results of experimental 6 show the use of clustering technique to divide students into "pass" or "fail" groups, as presented in Tables X and XI. Clusters 1, 2, 3, and 4 belong to the "pass" groups, whereas clusters 5 and 6 are categorized as "fail" groups. The analysis indicates that students in the "pass" groups, except for cluster 3, have a good ability to "keep up with the progress of the class"(time05 = 4), which is also reflected in the top ten "pass" rules presented in Table VII. On the other hand, students in both "fail" clusters (clusters 5 and 6) are comfortable in learning computer packages (csef10 = 4), but they feel incompetent with computers (csef26 = 2).

Table XI
Results Of Six Clusters (The Two Fail Clusters)

Item	Cluster5 (N = 18)	Cluster6 (N = 14)	Overall (N = 136)
weekly assignments	low	high	middle
sef01	1	3	3
sef05	2	3	3
sef08	2	3	3
mcog01	3	4	3
mcog09	3	4	3
time05	3	2	3
csef10	4	4	3
csef26	2	2	3
pass / fail	fail	fail	pass

This suggests that even if students do not face difficulty in learning new computer packages, without adequate confidence, they may still fail the course. The "fail" rule 55 in Table IX also supports this finding, which shows that students with a "high" level of csef10 may still fail due to their low self- efficacy in computers. Therefore, the analysis indicates that students' self-efficacies in computers play a critical role in their learning outcomes.

VI.DISCUSSION

The current study collected data from 136 undergraduate students to investigate the characteristics of successful learners in a computer course.

A. Summary of findings

In this study, data was collected from 136 undergraduate students using a self-report questionnaire thatcontained 64 items or factors. Out of these factors, eight were selected for running machine learning (ML) algorithms, including three self-efficacy elements, two metacognitive factors, one time-related attribute, and two computer self-efficacy characteristics. The score of weekly assignments was also usedas a part of the learning model. Seven supervised ML algorithms were used, including DT, BN, LR, NB, KNN, SVM, and MLP, to compare the performance of all classifiers. The results showed that NBwas the best model for predicting students' final performance, with an accuracy of 83.26% and sensitivity of 92.88%. The Vote ensemble, which combined NB, BN, and SVM, had the best accuracy(84.06%), sensitivity (93.18%), and specificity (78.48%) compared to other classifiers and ensembles.

To answer the second research question, unsupervised ML techniques were conducted using associationrules mining. The results showed that the two key factors for students to succeed in their class were demonstrating a "high" level of "keeping up with the weekly progress of the class" (the first factor) andachieving a "middle" level of self-efficacy (the second factor). On the other hand, most of the failed students received a "low" grade on weekly assignments or presented a "low" level of self-efficacy.

To verify the answer to the second research question, the clustering technique was used to comprehend the characteristics of each cluster. The results showed that the majority of students in the four "pass" clusters were those with a "middle" or "high" level on all the attributes, except for cluster 2. Additionally, three of the four "pass" clusters were "high" level on "keeping up with the weekly progressof the class" (time05). This finding is consistent with prior research, which suggests that the ability of time and study environment management significantly contributes to predicting academic outcomes. Incontrast, most of the students in the "fail" groups expressed "low" on both weekly assignments and self-efficacy beliefs. Self-efficacy has been identified in numerous studies as a strong predictor of student learning, and without such beliefs, it is challenging for students to succeed in the class.

B. Discussion of findings

The findings obtained through the association rule and clustering technique are consistent, suggesting that successful students perform better in terms of "keeping up with the progress in the class" (time05)and self-efficacy beliefs (i.e., sef01 and sef08). However, there were some inconsistencies regarding weekly assignments. For example, a "fail" group (cluster 5) had "low" scores on weekly assignments, while another "fail" group (cluster 6) had "high" scores. Similarly, a "pass" group (cluster 2) had "high"scores on weekly assignments, while another "pass" group (cluster 3) had "low" scores. This suggests that individual attributes and learning strategies have a more significant impact on learning outcomes than weekly assignments.

C. Limitations

The study reveals some ambiguous points that require further investigation. For instance, cluster 2, which consisted of 20 students, demonstrated a "low" level of related self-efficacy (sef01, sef05, and sef08 = 2) but was still classified as a "pass" group. On the other hand, cluster 6, which consisted of 14 students, had a "middle" level of the same attributes (sef01, sef05, and sef08 = 3) but was considered a "fail" group. One possible explanation for this discrepancy could be the small size of the sample. In future work, it would be meaningful to collect more participants to enhance the model's performance. Additionally, including more valuable parameters such as individual students' learning styles in the model would be worth exploring.

VII.CONCLUSION

In summary, this study provides a useful prediction model and identifies critical attributes for success in the class, which can guide teachers to provide appropriate learning environments and support studentsto become successful learners. The proposed methodology can also be applied to other courses or different grade levels, with adjustments made to the composition of the final grades based on the natureof the course. However, researchers must consider the impact of various factors such as computer resources, Internet speed, and learning context on student achievement before building their learning models. Overall, this study demonstrates the effectiveness of using ML algorithms to discover the critical attributes of successful learners, which can contribute to improving the quality of education.

References

[1] *R. L. Ahadi, H. Haapala, and A. Vihavainen, "Exploring machine learning methods to automatically identify students in need of assistance," in Proc. 11th Annu. Int. Conf. Int. Comput. Educ. Res., 2015, pp. 121–130.*

[2] *K. Quille and S. Bergin, "Programming: Further factors that influence success," in Psychology of Programming Interest Group (PPIG). Cambridge, U.K.: Univ. Cambridge, 2016.*

[3] *C. Y. Ko and F. Y. Leu, "Analyzing attributes of successful learners by using machine learning in an undergraduate computer course," in Proc. 32nd IEEE Int. Conf. Adv. Inf. Netw. Appl. (AINA-2018), Krakow, Poland, 2018, pp. 801–806.*

[4] *S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," Int. Trans. Comput. Sci. Eng., vol. 32, no. 1, pp. 71–82, 2006.*

[5] *J.-L. Hung and K. Zhang, "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching," J. Online Learn. Teach., vol. 4, no. 4, pp. 426–436, 2008.*

[6] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach," in Proc. 5th Int. Conf. Learn. Anal. Knowl., 2015, pp. 146–150.

[7] J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, "Identifying at-risk students for early interventions—A time-series clustering approach," IEEE Trans. Emerg. Topics Comput., vol. 5, no. 1, pp. 45–55, Jan.–Mar. 2017.

[8] C. Romero, M.-I. López, J.-M. Luna, S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," Comput. Educ. vol. 68, pp. 458–472, Oct. 2013.

[9] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, Oct. 2017.

[10] S. Amershi and C. Conati, "Unsupervised and supervised machine learning in user modeling for intelligent learning environments," in Proc. 12th Int. Conf. Intell. User Interfaces, 2007, pp. 72–81.

[11] B. J. Zimmerman, "Attaining of self-regulation: A social cognitive perspective," in Handbook of Self-Regulation, Research, and Applications, M. Boekaerts, P. Pintrich, and M. Zeidner, Eds. Orlando, FL, USA: Academic, 2000, pp. 13–39.

[12] B. J. Zimmerman, "Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects," Amer. Educ. Res. J., vol. 45, no. 1, pp. 166–183, 2008.

[13] A. Bandura, Social Learning Theory. Oxford, U.K.: Prentice-Hall, 1977.

[14] A. Bandura, Self-Efficacy: The Exercise of Control. New York, NY, USA: Freeman, 1997.

[15] R. Lynch and M. Dembo, "The relationship between self-regulation and online learning in a blended learning context," Int. Rev. Res. Open Distance Learn., vol. 5, no. 2, pp. 1–16, 2004.

[16] M. V. J. Veenman, B. H. A. M. Van Hout-Wolters, and P. Afflerbach, "Metacognition and learning: Conceptual and methodological considerations," Metacogn. Learn., vol. 1, pp. 3–14, Mar. 2006.

[17] P. R. Pintrich, "The role of goal orientation in self-regulated learning," in Handbook of Self-Regulation, M. Boekaerts, P. R. Pintrich, and M. Zeidner, Eds. San Diego, CA, USA: Academic, 2000, pp. 451–502.

[18] D. H. Schunk, "Self-regulated learning: The educational legacy of Paul R. Pintrich," Educ. Psychol., vol. 40, no. 2, pp. 85–94, 2005.

[19] A. Kitsantas, A. Winsler, and F. Huie, "Self-regulation and ability predictors of academic success during college: A predictive validity study," J. Adv. Acad., vol. 20, no. 1, pp. 42–68, 2008.

[20] D. Compeau and C. Higgins, "Computer self-efficacy: Development of a measure and initial test," MIS Quart., vol. 19, no. 2, pp. 189–211, 1995.