



A novel color labeled student modeling approach using e-learning activities for data mining

Selim Buyrukoğlu¹

Accepted: 7 June 2022 / Published online: 27 June 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Student modeling approaches are important to identify students' needs, learning styles, and to monitor their improvements for individual modules. Lecturers may incorrectly identify the students' needs and learning styles based on solely an exam grade or performance in the class. In doing so, students need to be classified using more parameters such as e-learning activities, attendance to virtual live class (for theory and practice) and submission time of the assignment, etc. This study proposes a novel color-labeled student modeling/classification approach using e-learning activities to identify students' learning styles and to monitor students' weekly improvements for individual modules. A novel Student Classification Rate (SCR) formula was created by combining three stages including pre-study stage, virtual_class stage, and virtual_LAB_class stage. In the evaluation part of the SCR, Artificial Neural Network and Random Forest algorithms were employed based on two different feature sets for an Object-Oriented Programming Module. Feature set 1 consisted of a combination of e-learning and regular data while the feature set 2 was referred as the combination of the SCR and the regular data. Random Forest yielded the lowest MAE (0.7) by using feature set 2. Also, the majority of the students' (81%) learning styles referred to attending the live virtual class. Students' weekly learning progress was also monitored successfully since the Pearson correlation was measured as 0.78 with the 95% confidence interval between the mean of SCR and lab grades. Additionally, SCR used for two more different modules yielded convincing results in the determination of students' learning styles. The obtained results reveal that the proposed SCR approach has significant potential to correctly classify students, identify students' learning styles, and help the lecturer to monitor the students' weekly progress. Finally, it seems that SCR would have a significant impact on improvement of students learning.

Keywords Student modeling · Student classification rate · E-learning · Data mining · Random forest · Learning style

1 Introduction

Learning types in education are generally classified into Face-to-Face (F2F) and distance learning [1]. The efficiency of the learning types regarding students' performance can be different based on the subjects and modules. There is a big challenge in F2F learning, such as that lecturers should analyze the students in the classroom properly and encourage each student to improve their learning skills about the related subjects [2]. In this sense, the lecturer should determine

appropriate features to analyze the students' needs and be consistent among students in the analysis process. However, the profile analysis of students for individual modules based on F2F education is not an easy task because the lecturers should observe students in this learning type. The lecturer may miss the important factors in the analysis process of the students during the observation.

The popularity of distance learning has increased with the start of the COVID-19 pandemic, especially in higher education [3]. Various distance learning systems (Zoom, Perculus ALMS, Google meet, etc.) have been used in different classes such as math, computer programming, sociology, chemistry, etc. A significant aspect of distance learning systems is that information belonging to students is automatically recorded as the number of the attempted quiz, the number of submitted assignment, the assignment submission time, attendance time of students in a virtual live class (for

✉ Selim Buyrukoğlu
sbuyrukoglu@karatekin.edu.tr;
selimbuyrukoglu@hotmail.com

¹ Department of Computer Engineering, Faculty of Engineering, Çankırı Karatekin University, 18100 Çankırı, Turkey

theory and practice), and attendance of students to virtual class repetition (for theory and practice), etc. Students may have the opportunity to learn properly with virtual live class and/or repetition of the class. In addition to this, students may like to study in the daytime, evening, or nighttime [4]. In this case, these types of recorded data can be used for student classification/modeling, to detect learning style (virtual live class and/or repetition of virtual class), and to monitor students' weekly/monthly progress based on the individual module.

This study focuses on the classification of students based on e-learning activities. Thus, a novel classification approach called Student Classification Rate (SCR) is proposed. Three feature categories based on e-learning activities (lecture notes, virtual class for theory, virtual class for practice-LAB) are taken into account for classifying students' learning styles and revealing the efficiency of the e-learning platform on the students' learning process. The proposed approach also advocates monitoring students' learning progress. In this sense, a lecturer can monitor students' progress weekly, and also students can be managed better compared to F2F learning through the proposed SCR approach. Also, the efficiency of the lecturers in managing students' learning progress can be increased through the SCR. Lecturers may not manage students well through F2F learning because manually capturing the three feature categories based on e-learning activities is not an easy task in F2F learning.

Learning style, affecting students' learning progress, (virtual live class and/or repetition of virtual class) is also important. A student can prefer to attend a virtual live class while a different student can prefer to attend the repetition of a virtual class. The proposed SCR approach also supports the detection of students learning styles. Hence, a lecturer can also lead students based on their learning styles.

The proposed novel SCR approach is designed and developed based on e-learning technologies (please see Sect. 3 for detailed information). Then, the impact of this novel approach is evaluated in education based on different modules. The proposed SCR approach may also contribute to the efficiency of distance learning compared to the traditional F2F learning style in the classification of students. In other words, lecturers can primarily focus on using e-learning platforms rather than traditional F2F learning through the novel SCR approach, which is the theme of this study. Thus, the learning styles of students can be gradually changed by using the designed and developed new SCR approach.

A review of prominent student classification approaches is given in Sect. 2, which reveals that existing approaches have focused on general students' information (age, gender, etc.) rather than feature categories based on e-learning activities (lecture notes, the virtual class for theory, the virtual class for practice-LAB, etc.). Thus, this study intends to answer the following research questions:

- Can a student modeling approach be developed using e-learning activities to detect students learning styles?
- Can students be classified by using SCR based on the e-learning activities?
- Can students' progress be monitored by using SCR for individual modules?
- Can SCR be effective for different modules?

The structure of the paper is as follows: related work on student classification is given in Sect. 2. Section 3 introduces the novel approach (Student Classification Rate—SCR) that can be listed under four sub-headings: data gathering, Student Classification Rate, student classification level, and data mining algorithms; a rationale for the selection of data mining algorithms to evaluate the efficiency of the SCR approach is also presented. Section 4 presents the results and discussion on the efficiency of SCR. Finally, the conclusion and future direction are outlined.

2 Related work

This section discusses the student classification approaches, and also provides information on the reason for using e-learning activities in the classification of students. In literature, a variety of student modeling approaches have been proposed based on different purposes [5]. The overlay model is one of the popular student modeling approaches invented by Stansfield et al. [6] to identify students' levels. A different student modeling approach was also proposed to classify students using students' frequent characteristics named Stereotypes [7]. Also, machine learning techniques are well-known for student modeling approaches [8]. We used machine learning techniques to reveal the efficiency of our student modeling approach in the present study (please see Sect. 3 for detailed information).

Baker [9] proposed a model based on machine learning to automatically detect students who do not get involved in the learning tasks (off-task). In a different study, fuzzy logic was used as a constructed student modeling to assess students' homework and assignment [10]. GIAS is an educational system developed based on machine learning techniques for course selection according to students' learning styles and knowledge levels [11]. In a different student modeling approach, Bayesian networks and machine learning techniques were combined to monitor students' responses while using a Learning Management Systems (LMS) [12]. Also, Baker et al. [13] proposed a similar student modeling approach [12] to combine the Bayesian networks and machine learning techniques. This model focused on the prediction of the students' knowledge according to students' learned skills. Cetintas et al. [14] presented a student classification model applying machine learning techniques for

automatic off-task detection in LSMs. In another study, students learning performance is predicted based on a student model using machine learning techniques [5].

In a different student modeling approach, academic records (quizzes, exams, and assignments) are used as input features to predict students' performances by employing machine learning techniques [15]. Zacharis [16] applied machine learning techniques to investigate the impact of online activities on students' academic performance in Moodle. Ayub et al. [17] also used features obtained from an LMS about students' academic records including assessment and exam results to improve students' engagement in a module. Another study used enrollment data and academic records to predict students' performance [18]. The efficiency of different feature sets (students' behaviors, academic records, etc.) was examined in the prediction of students' performances [19]. Moreover, the influence of different feature sets on academic, behavioral, demographic, student attendance, and additional features about students' families, etc., was examined in the prediction of students' academic performances [20, 21]. Various machine learning algorithms have been used to predict students' performances such as Random Forest (RF) [22–24], Decision Tree (DT) [25], Multilayer Perceptron (MLP) [26], and Logistic Regression (LR) [27].

Even if the previous studies regarding student modeling focused on the impact of the different features, few studies have examined the impact of academic records obtained from the LMS systems. However, to the best of our knowledge, no previous study has investigated the impact of academic features obtained from e-learning activities to extract the students' learning styles and to monitor students' progress, weekly. Thus, this research aims to create a student modeling approach to determine students' learning styles and to monitor students' progress weekly/monthly. The recording of e-learning activity features is difficult in F2F education for lecturers. Thus, we believe that e-learning activities can

be used in student classification. In this sense, a novel model named Student Classification Rate (SCR) using the e-learning academic records is proposed. The weekly progress of each student will be labeled based on SCR using different colors while the student's learning style is identified. Thus, the lecturer can monitor progress based on individual modules during the semester. The main objective of the study is to reveal the efficiency of e-learning platforms compared to the F2F learning in student modeling. The following section presents the proposed Student Classification Rate approach (SCR) developed, based on e-learning activities.

3 A novel student modeling approach: student classification rate

This section provides detailed information about the general methodology of the study. Figure 1 presents the four main phases including data gathering, student classification rate (SCR), student classification level, and evaluation of SCR based on data mining models. SCR aims to classify students, detect students' learning styles, and monitor their learning progress in certain time intervals (such as weekly or monthly) for different modules using e-learning feature categories.

The rest of this section provides information about these phases.

3.1 Data gathering

Two types of data related to e-learning activities and students' general information were collected. Data about e-learning activities were collected to classify the students. Perculus ALMS distance learning system (<https://karatekin.almscloud.com/>) was used to obtain data about students' e-learning activities (please see Table 1). The e-learning activities of 80 students were recorded (6 weeks

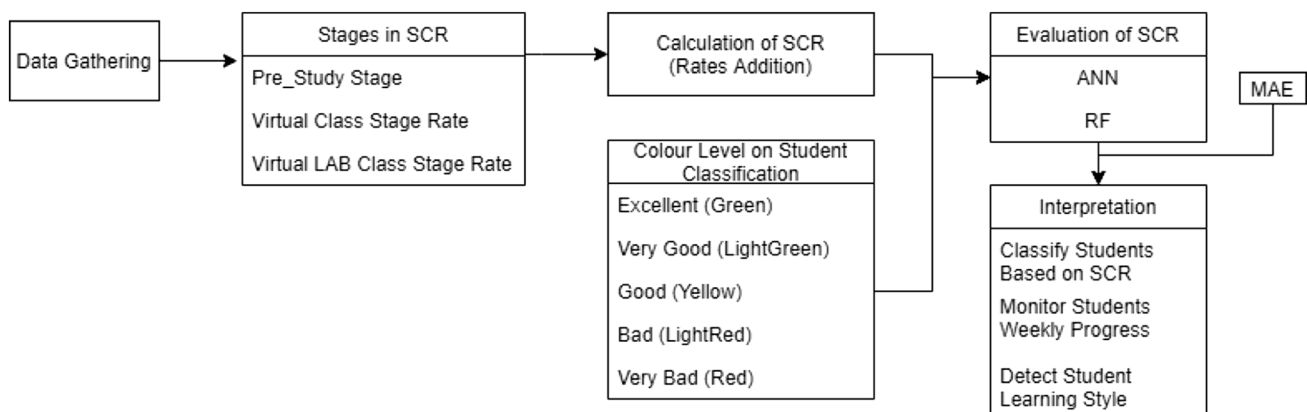


Fig. 1 Flowchart of student classification rate (SCR)

Table 1 Features on e-learning

Feature category	Feature	Description	Variable
Lecture notes	Lecture_Note	Lecture note which is uploaded to the system by the lecturer. Whether the student downloaded the lecture notes before the lecture time is taken into account	numeric
Virtual class for theory	Live_Class_Time	Total virtual live class time for theory	numeric
	Attd_Time	Attendance time of students to virtual live class (for theory)	numeric
	Attd_Rpt_Time	Attendance time of students to virtual class repetition (for theory)	numeric
Virtual class for practice (LAB)	Live_Lab_Class Time	Total virtual live LAB (laboratory) class time for theory	numeric
	Attd_Lab_Time	Attendance time of students to virtual LAB live class (for practice)	numeric
	Attd_Lab_Rpt_Time	Attendance time of students to virtual LAB class repetition (for practice)	numeric
	Ass_Sub	Assignment submission time	numeric
Grade (target)	Lab_Grade	0 to 100	numeric

Table 2 Features on regular data

Feature	Description	Variable
Age	Age of the student	Numeric
Gender	Gender of the student	Binary
Study_Time	Weekly study time	Numeric
Study_Sup	Extra study support	Binary
Internet_Ava	Internet at home	Binary
Family_Sup	Family study support	Binary
Free_time	Free time after school	Binary
Go_Outside	Going out with friends	Binary
Health	Current health status	Numeric
Graduate	Wants to graduate school (MSc or Ph.D.)	Binary
Activities	Extra-curricular activities	Binary
Computer_Ava	Computer at home	Binary

separately) until the midterm exam. Students were taking semester one (2020) of the object-oriented programming module at a university in Turkey. The students downloaded lecture notes from the Perculus ALMS and then attended the virtual classes for both theory and practice lectures each week, respectively. Table 1 presents the features related to e-learning activities that were recorded during this period.

Additionally, data about students' general information were also obtained, including demographics, students' absence, and extra features about students' families, etc. (see Table 2). The aim of obtaining general information was to reveal the efficiency of the SCR approach. In the evaluation process of SCR, appropriate data mining algorithms using two feature sets were applied. The first feature set referred to all features from Table 1 (e-learning features) and Table 2 (students' general information). The second feature set consisted of SCR as a feature and all features from Table 2. Then, data mining

models' performances were compared in terms of revealing the impact of SCR on the students' performances.

3.2 Student classification rate

The Student Classification Rate (SCR) approach consisted of three main stages including Pre_study stage, Virtual_class stage, and Virtual_LAB_class stage as illustrated in Fig. 1. The sum of these three stages created the SCR. An elaborated presentation about these stages and the calculation of SCR are presented in the remaining of this section.

3.2.1 Pre_study stage

The Pre-study Stage enables lecturers to understand whether a student downloads the lecture notes from the ALMS system before the virtual class. If students study the lecture notes before the lesson, they may understand the subjects better compared to students who do not study the lecture notes before the lesson. In this sense, the students probably get a convincing LAB grade. Thus, the Pre-study Stage is created in the calculation of SCR.

In Eq. (1), the rate of the pre-study stage means that if a student downloads the lecture note from the ALMS system, the rate of the pre-study stage is accepted as 10. Otherwise, the rate is established as 0. Note that the definition of the metrics/parameters used in Eqs. 1, 2, 3, and 4 are given in Table 1.

$$\begin{aligned}
 &\text{if}(\text{Lecture_Note} = \text{YES}) \\
 &\quad \text{Pre_study_rate} = 10 \\
 &\text{else} \\
 &\quad \text{Pre_study_rate} = 0
 \end{aligned} \tag{1}$$

3.2.2 Virtual_class stage

The reason for the creation of this stage was to measure the impact of theory classes on e-learning platforms. As seen in Eq. (2), the rate of the virtual class stage equals the sum of a and b . Here, formula a shows the participation rate of a student in the virtual live class. Then, the participation rate was multiplied by 30. The same calculation was also processed to calculate the formula b . However, formula b shows the attendance time of students to virtual class repetition which is recorded by the system automatically. Then, the repetition rate is multiplied by 10. Finally, a and b were summed to obtain the virtual class rate. Note that if the b value was higher than or equal to 10, b was accepted as 10. The reason behind the choice of weights (30 and 10) used in the a and b formula is presented rest of this stage in detail. where

$$\text{Virtual_class_rate} = a + b,$$

$$\begin{aligned} a &= \left(\frac{\text{Attd_Time}}{\text{Live_Class_Time}} * 30 \right) \\ b &= \left(\frac{\text{Attd_Rpt_Time}}{\text{Live_Class_Time}} * 10 \right) \end{aligned} \quad (2)$$

Students could ask questions verbally or send a text to the lecturer from the chatbox of the learning systems during the live virtual class. Lecturers provided instant answers to students' questions during the live virtual class. These instant answers aimed to help students in terms of better understanding of the subjects. In contrast to this, students could ask questions via e-mail to the lecturer when they repeat the live virtual class. However, lecturers mostly did not give instant answers to students' questions. The reason behind this is that the lecturer can be busy or cannot take enough time to answer students' questions. In this sense, students may not learn well the subjects due to the late answers, while the students' interest in the subject may also decrease. Thus, the attendance of a virtual live class (a) was considered more efficient than the attendance of a virtual class repetition (b) in the calculation of student classification rate.

Therefore, the weight of the attendance time of students to a virtual live class (a) was accepted more than the attendance time of students to virtual class repetition (b). However, we tested the possibilities in the determination of the weights for a and b to obtain robust results in the student classification. In other words, a case study was carried out to determine the optimum weights of a and b in this study. Weights for a and b were tried as 20–20%, 10–30%, 25–15% and 30–10%, respectively. The best student classification performance was obtained by using the 30% (for a) and 10% (for b) weights in the calculation of the virtual class rate. It should be noted that the weights were also used in the calculation of the LAB virtual class rate stage.

3.2.3 Virtual_LAB_class stage

The reason for the creation of this stage was to measure the impact of LAB classes on e-learning platforms. As presented in Eq. (3), the calculation of the virtual LAB class stage is the sum of a , b , and c . The calculation of a and b is similar to the a and b presented in Eq. (2). There is only one difference which is that the a and b presented in Eq. (3) are calculated by using the records about the virtual LAB live class instead of the virtual live class. Moreover, the formula c is related to the submission time of an assignment. In other words, if a student submits his/her assignment on time or before the submission deadline, the c rate is considered as 10. In contrast to this, the c rate is considered as 0. Finally, the sum of a , b , and c refers to the virtual LAB class rate. Assignments were announced to students after the virtual LAB class, and so the formula c was included in the virtual LAB class rate formula. where

$$\text{Virtual_LAB_class_rate} = a + b + c,$$

$$\begin{aligned} a &= \left(\frac{\text{Attd_Lab_Time}}{\text{Live_Lab_Class_Time}} * 30 \right) \\ b &= \left(\frac{\text{Attd_Lab_Rpt_Time}}{\text{Live_Lab_Class_Time}} * 10 \right) \\ &\quad \text{if (Ass_Sub = YES)} \\ &\quad c = 10 \\ &\quad c = \text{else} \\ &\quad c = 0 \end{aligned} \quad (3)$$

3.2.4 Student classification rate (SCR)

The sum of rates of these three stages (pre_study, virtual_class, and virtual_LAB_class) creates the SCR. The formula of SCR is given in Eq. 4.

$$\text{SCR} = \text{Pre_study_stage} + \text{Virtual_class_stage} + \text{Virtual_LAB_class_stage} \quad (4)$$

The weight of pre_study_stage, virtual_class_rate, and virtual_LAB_class_rate was determined as 10%, 40% and 50% in the calculation of SCR, respectively. It is known that the virtual_class and virtual_LAB_class stages are more important than the pre_study_stage [28]. Thus, its weight in the calculation of SCR was accepted as 10%. Before we decide to use 10% as a weight for the pre_study stage, we tried different combinations such as 15% and 20%. In the end, 10% was accepted as the weight of the pre_study stage since it provided more beneficial results for this research. On the other hand, even if

virtual_LAB_class seems more important than virtual_class, the virtual_class is also important as virtual_LAB_class. The reason is that students understand the theory of subjects in the virtual_class before the practice related to the subject in the virtual_LAB_class. Thus, the virtual_class and the virtual_LAB_class had the same importance for this research.

3.3 Student classification level

Table 3 provides information on the student classification level. As can be seen in Table 3, the difference between threshold values of student classification levels is 20 which means that there is a standard range between the levels. The definition of the levels is given below:

- **Excellent:** If a student gets more than 90 grades, the level of the student is considered Excellent;
- **Very Good:** If a student gets a grade between 70 and 89, the level of the student is considered convincing for the course/module in many countries around the World [29]. Thus, the threshold value for Very Good is set as 70;
- **Good:** If a student wants to pass any module, the student has to get a minimum of 50 grades. Hence, 50 grades is accepted as the threshold value for Good;
- **Bad:** The grade for any module less than 50 is not sufficient to pass any module. In this case, 30 is specified as a threshold value for Bad;
- **Very Bad:** A grade less than 30 is highlighted as a threshold value for the Very Bad level.

Table 3 Colored level table on student classification

Colored level	Grade	
	Min (Threshold)	Max
Excellent (Green)	90	100
Very Good (Light Green)	70	89
Good (Yellow)	50	69
Bad (Light Red)	30	49
Very Bad (Red)	0	29

Additionally, Table 3 consists of colored levels based on the students' grades. Lecturers can easily monitor students' progress through these colors.

3.4 Selection of data mining models to evaluate the efficiency of SCR

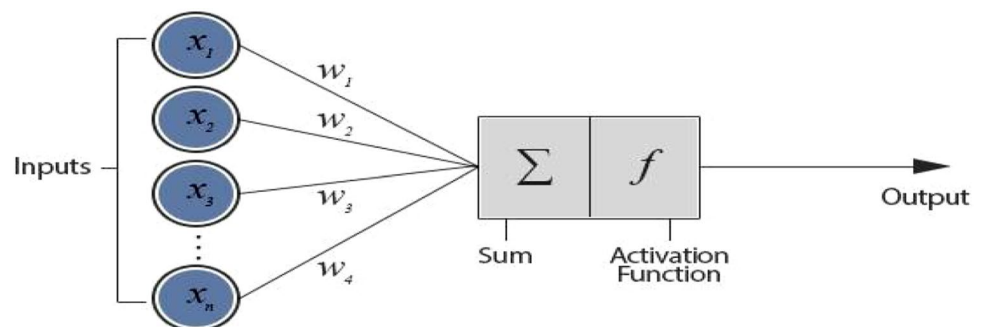
The efficiency of the SCR approach is measured by using data mining models based on two feature sets described in Sect. 3.1. These models are Artificial Neural Network (ANN) and Random Forest (RF). The reason behind the selection of these models is that they are the most preferred models in the students' classification or student performance prediction cases [30, 31].

3.4.1 Artificial neural network

ANN consists of input, hidden, and output layers [32]. The input parameters are processed in the hidden layer, and then the result is generated through the output layer. The utilization of an activation function enables the network to make a relationship between input and output. The reason behind this is that nonlinear property is introduced to the network through the activation function. Figure 2 shows the architecture of the employed neural network.

Two different ANNs were created to measure the efficiency of the SCR approach. In the first ANN, twenty features were used (see Tables 1 and 2) in the input layer, and one hidden layer with 14 neurons was used to predict the students' LAB grades. According to researchers, the number of inputs should take into account when deciding the number of neurons in the hidden layer [33, 34]. Moreover, the ReLu activation function was used in terms of making a relationship between the input and output. On the other hand, in the second ANN, thirteen features including SCR (which was calculated using e-learning features presented in Table 1) and general features (see Table 2) were used in the prediction of students' lab grades. In this case, 8 neurons were used in the hidden layer and then the ReLu activation function was also used to generate an output.

Fig. 2 General Architecture of ANN



3.4.2 Random forest

Random Forest (RF) is considered a tree-based ensemble learning algorithm that uses multiple decision trees to provide an output [35]. The output is calculated using the random subset of features in each node of the decision tree. Then, the outputs of individual trees are combined in Random Forest to generate the final output. The number of trees in the Random Forest should be determined to obtain optimum results [36]. Thus, here, a different number of trees (10, 50, 100) were tried to decide an optimum number of trees. The best performance was obtained with fifty trees.

3.4.3 Evaluation metrics

Root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE) evaluation metrics are often used in the evaluation of the performances of models. In this study, performances of ANN and RF models were evaluated based on mean absolute error (MAE). The reason for this is that MAE is a more efficient estimator compared to the RMSE/MSE based on the population mean of absolute error [37]. In MAE, the difference between the actual and predicted value is used to measure the average magnitude of the error [38].

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |P_i - \hat{P}_i| \quad (5)$$

where P_i is the actual value, $\hat{P}_i$ is the predicted value from the model and n is the number of observations. Also, fivefold cross-validation was used in the study [39].

4 Results and discussion

4.1 Importance of features for student classification

To reveal the importance of features in the student classification process, the Pearson correlation between SCR mean of six weeks and the lab grades mean of six weeks has been calculated (0.78 with the 95% confidence interval). Note that the correlation between the SCR mean and the lab grade mean is statistically significant indicating that there is a positive relationship between the mean of SCR and lab grade. The correlation also indicates that the features and the weight of each stage used in the SCR formula seem appropriate in the classification of students. It can be inferred that students will be classified correctly by using the proposed SCR approach, and the most effective

feature can be considered as the attention to the live virtual class and live virtual LAB class.

Means of students' lab grades for six weeks were compared to determine the students' learning styles. The mean lab grades for students attending the live virtual class was 64 (81% of students), whereas the mean lab grade was 16 (19% of students) for students attending the only repetition of the virtual class. This shows that most of the students' (81%) learning styles refer to attending the live virtual class. Moreover, it should be noted that attendance of live virtual class or live virtual class repetition was not compulsory for students. Also, 24% of students attended both live virtual class and live virtual class repetition. However, these students did not watch the whole video, and so they probably repeated some parts of the lecture which were not clear to them. In general, most of the students prefer to study during daytime.

Repetition of the virtual class and pre-study stage has the same weight in the proposed SCR approach. The reason is that students cannot ask any instant questions to the lecturer in the repetition of the virtual class and pre-study stage. In other words, students should learn subjects themselves if any subject is not clear for them or ask questions to lecturers via email, etc. Thus, repetition of the virtual class and pre-study stage weights is assigned the same in the SCR approach. Also, the lecturer captures information about the self-study ability of students through the pre-study stage.

On the other hand, the submission time of the assignment is also taken into account in the virtual_class_LAB_rate as 10%. This feature is important as the pre-study stage and attendance time of students to virtual class repetition (for theory/LAB). The reason behind this is that the lecturer has information about whether the student has self-disciplined or not. For example, if a student submits his/her assignment on time or before the submission deadline, these students can be considered self-disciplined students. Thus, submission time plays an important role in the calculation of SCR. Also, no additional time was given to students for assignment submission after the deadline. However, students may not submit their assignments on time due to health problems, and this health problem may affect students' learning. It should be noted that students did not provide any medical reports during the submission times in our dataset. We assumed that students did not have any health problems during the first six weeks of the term. If students had health problems, they should have been considered as outliers to obtain robust classification results.

4.2 Evaluation of the SCR approach using data mining method

The SCR approach is evaluated in this section using ANN and RF data mining models. These models were evaluated

based on MAE error metrics as explained in Sect. 3.4.3. As presented in Sect. 3.1, data consisted of 21 features. Eight of these features were obtained from e-learning activities and the rest of them were regular features. The eight e-learning features were used to obtain SCR as given in Eq. 4.

In the evaluation part of SCR, all features (20) were initially used to predict the LAB result of students based on ANN and RF. Then, SCR (1 feature) and the regular features (12) were combined to predict the LAB result of students using ANN and RF algorithms. The evaluation was carried out for the first six weeks of the term. According to the results presented in Table 4, the SCR + General feature set provided better results for each week compared to the E-learning + General feature set. The SCR approach can be considered an effective approach to student classification. The best MAE value for SCR + General and E-learning + General feature sets are 0.7 and 1.1, respectively, and they were obtained by applying the RF.

The results show that the proposed SCR approach can be used to classify students in terms of weekly progress information. Also, the proposed SCR approach can be applicable only for this kind of data.

4.3 Color-labeled student classification information using the SCR approach

Students' levels were classified according to the SCR which was described in Sect. 3.3. Table 5 provides information on the weekly number of students' classification results based on the SCR.

One of the research goals of this study is to correctly classify students through SCR. From Table 5, it can be

seen that the number of students at the Good level is more than the number of students in other levels for each week. On the other hand, the number of students at the Excellent level is the least among the groups. It is noted that a normal histogram is an expected case on the separation of the student numbers based on the classification levels or grades, etc. [40]. In our study, a normal histogram was obtained through the SCR approach because the density was at around the three levels (Very Good, Good, Bad). The number of students at the Very Bad level is reduced from Week 1 to Week 6 in general. This implies that students learn computer programming, and so the level of students usually moves up to the next or further levels. These results show that the proposed SCR approach is appropriate for the student classification.

Table 6 illustrates the weekly color-labeled progress of five different students. As seen in Table 6, the progress of Student 1 slightly decreases from the first week to last week. One can state that this student had a high grade from the first lab exam and then student 1 may have reduced studying or not focused on subjects. On the other hand, progress during the first three weeks of Student 4 remained stable and slightly increased from week 4 to week 6. The rest of the students' (Students 2, 3, and 5) progress varied, as can be seen in Table 6.

Even if the proposed approach answered the research questions presented in Sect. 1, there are some outliers. For instance, some students attended both live virtual class and repeated the virtual class regularly but their weekly progress is stable and does not increase. On the other hand, some of the students' weekly progress slightly increases even if they did not repeat the virtual class. This means that the

Table 4 Result of the SCR evaluation

Feature Set	Data mining model	Week No					
		W1	W2	W3	W4	W5	W6
E-learning + General	ANN	2.7	3.3	2.9	4.6	3.5	3
	RF	1.2	1.4	1.1	1.7	1.7	1.2
SCR + General	ANN	2	2.9	2.6	3.1	2.9	2.8
	RF	0.7	1.1	0.8	1.4	1.2	0.9

The bold values show the best MAE rate for each feature set

Table 5 Weekly number of student classification level results based on SCR

Week No	Excellent	Very good	Good	Bad	Very bad	Total
1	5	11	28	14	22	80
2	7	19	33	10	11	80
3	7	16	25	13	19	80
4	5	20	24	22	9	80
5	4	27	31	13	5	80
6	5	24	29	15	7	80

Table 6 Students' weekly color labeled progress information

Students	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Colour Definition
S1 (repeat of the lesson)	Excellent	Excellent	Very Good	Very Good	Good	Good	Excellent
S2 (live virtual lesson)	Bad	Good	Good	Very Good	Good	Very Good	Very Good
S3 (live virtual lesson)	Bad	Bad	Bad	Bad	Good	Bad	Bad
S4 (live virtual lesson)	Good	Good	Good	Very Good	Very Good	Excellent	Excellent
S5 (repeat of the lesson + live virtual lesson)	Very Good	Good	Very Good	Very Good	Good	Very Good	Very Good

live virtual class can be considered more important than the repetition of the virtual class in the student classification. Moreover, the weekly improvement process of some of the students who did not attend the live class reduces. The reason behind this is that the students cannot ask any questions to the lecturer while they are watching the repetition of the virtual class, and so students cannot improve their skills. However, it is noted that lecturers can easily manage students based on their weekly color-labeled progress information (see Table 6). Also, the lecturers may accelerate the improvement progress of the students through the SCR. Additionally, a lecturer can monitor students' weekly progress not only for a module but also for more than one module during the term. That is, students' overall weekly progress for each module can efficiently be monitored through SCR by a lecturer.

4.4 Applying SCR to different datasets

The proposed student modeling approach was also used for different datasets behind the object-oriented programming module. Two different data sets were also collected using the Perculus ALMS platform. Data of 132 students taking Java programming module and data of 28 students taking Deep Learning module in Python were recorded until the midterm exam. Table 7 shows the performance results of the algorithms using these data.

The SCR + General feature set provided better results for each week compared to the E-learning + General feature set based on Java programming and Deep Learning data sets. In the Java programming module, the best MAE value for SCR + General and E-learning + General feature sets obtained by application of the RF are 0.8 and 1.2, respectively. On the other hand, RF provided better MAE values compared to ANN based on the Deep Learning data set (the best MAE value for SCR + General and E-learning + General feature sets are 4.9 and 5.2, respectively).

It should be highlighted that the SCR approach is not the best in terms of student classifying. However, it can be considered an efficient and sufficient approach. The reason behind this is that SCR provided desired results with different datasets (Java programming and Deep Learning modules) besides data set about object-oriented programming modules.

5 Summary of contributions

The proposed, novel, SCR approach achieved to classify students, which is the main contribution of this study. Thus, a lecturer can reliably prefer to use e-learning platforms for teaching. Then, the lecturers can monitor students learning progress at weekly (or any other frequency) intervals, which is another contribution of this study (see Table 6). In this sense, lecturers may manage their students based on weekly

Table 7 Results of SCR evaluation for different courses

Data set	Feature set	Data mining model	Week No					
			W1	W2	W3	W4	W5	W6
Java Prog	E-learning + General	ANN	2.1	2.8	4.3	1.6	3	3.5
		RF	1.7	2.1	4	1.2	2.4	2.9
	SCR + General	ANN	1.8	2.3	3.7	1.4	2.2	3.1
		RF	1.2	1.5	3.3	0.8	1.8	2.7
Deep Learning	E-learning + General	ANN	6.6	7.1	5.7	8.3	6.2	6.4
		RF	6.1	6.5	5.2	7.4	5.7	6
	SCR + General	ANN	6.4	6.7	5.5	7.7	5.8	5.9
		RF	5.8	6.1	4.9	6.9	5.1	5.3

The bold values show the best MAE rate for each feature set

learning progress. For instance, a student's level can be bad in the first week of the semester. Then, the lecturer aims to improve the student's level from bad to good. It could be inferred from this example that the proposed novel SCR approach enables to increase the quality of education level in terms of both students and lecturers. The proposed approach was also used for different modules (Java Programming in BSc and Deep Learning at MSc level) besides the Object-Oriented Programming module at BSc level. The reason behind the use of the SCR approach for different modules is to reveal the efficiency and robustness of this approach in different modules which is another contribution of this study. Moreover, even if F2F learning has significant advantages for the educational process, obtaining the features through e-learning activities is difficult using F2F learning. Moreover, the use of e-learning platforms has become widespread with the COVID-19 pandemic. In this sense, it can be considered that the use of e-learning platforms in the student classification can be more efficient than traditional F2F learning because the academic features are automatically recorded through the e-learning platforms. Highlighting the efficiency of the e-learning platform in student classification using academic records against the F2F learning is another contribution of this study.

6 Conclusion and future directions

A novel color-labeled student modeling approach has been presented in this paper. E-learning activities were used to create the Student Classification Rate (SCR) in terms of detecting students learning styles and monitoring students' weekly learning progress. The findings show that the proposed student modeling approach correctly detects students' learning styles and monitors students' weekly progress during the term. Thus, lecturers can easily contribute to the improvement of students' learning levels. The lecturer can also provide personalized feedback for each student based on weekly color-labeled progress information. Artificial neural networks and random forest algorithms were applied based on two different feature sets in the evaluation part of the SCR. The findings show that random forest provided the lowest MAE value (0.7) using the SCR and general data. Additionally, the proposed SCR approach was also used for two different modules, namely Java Programming and Deep Learning modules. The random forest was again provided the lowest MAE values (0.8 and 4.9, respectively) using the SCR and general data.

Case-Based Reasoning (CBR) can be used to accelerate the detection of student learning levels and the personalized feedback process. Thus, the feedback process can be automated as much as possible. This will enable lecturers to

use the SCR approach more efficiently. Moreover, students can be classified separately based on three feature categories presented in Table 1 namely lecture notes, the virtual class for theory, and the virtual class for practice (LAB). In this sense, the lecturer may manage students based on separated feature categories.

Funding None.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The author declares that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Vázquez-Colina, M.D., Russo, M.R., Lieberman, M., Morris, J.D.: A case study of using peer feedback in face-to-face and distance learning classes among pre-service teachers. *J. Furth. High. Educ.* **41**(4), 504–515 (2017). <https://doi.org/10.1080/0309877X.2015.1135884>
2. Binmohsen, S.A., Abrahams, I.: Science teachers' continuing professional development: online vs face-to-face. *Res. Sci. Technol. Educ.* **00**(00), 1–29 (2020). <https://doi.org/10.1080/02635143.2020.1785857>
3. Dumford, A.D., Miller, A.L.: Online learning in higher education: exploring advantages and disadvantages for engagement. *J. Comput. High. Educ.* **30**(3), 452–465 (2018). <https://doi.org/10.1007/s12528-018-9179-z>
4. Johnston, J., Killion, J., Oomen, J.: Student satisfaction in the virtual classroom. *Internet J. Allied Health Sci. Pract.* **3**(2), 1–7 (2005)
5. Li, N., Cohen W.W., Koedinger R.K., Matsuda, N.: A machine learning approach for automatic student model discovery. In *Proceeding 390p* (Issue December). Edm. (2011)
6. Stansfield, J., Carr, B., Coldstein, I.: *Wumpus Advisor* (p. 67). (1976)
7. Rich, E.: User modeling via stereotypes. *Cogn. Sci.* **3**(4), 329–354 (1979). [https://doi.org/10.1016/S0364-0213\(79\)80012-9](https://doi.org/10.1016/S0364-0213(79)80012-9)
8. Webb, G.I., Pazzani, M.J., Billsus, D.: Machine learning for user modeling. *User Model. User-Adap. Inter.* **11**(1–2), 19–29 (2001). <https://doi.org/10.1023/A:101117102175>
9. Baker, R.S.J.D.: Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Conf. Human Factors Comput. Syst. Proc.* (2007). <https://doi.org/10.1145/1240624.1240785>
10. Jurado, F., Santos, O. C., Redondo, M. A., Boticario, J. G., Ortega, M. Providing dynamic instructional adaptation in programming

- learning. In: International Workshop on Hybrid Artificial Intelligence Systems (pp. 329–336). Springer, Berlin, Heidelberg. (2008)
11. Castillo, G., Gama, J., Breda, A.M.: An adaptive predictive model for student modeling. *Adv. Web-Based Educ. Pers. Learn. Environ.* (2005). <https://doi.org/10.4018/978-1-59140-690-7.ch004>
 12. Suarez, M.: Predicting student's appraisal of feedback in an its using previous affective states and continuous affect labels from eeg data. *Res. Pract. Technol. Enhanc. Learn.* (2011)
 13. Baker, R.S.J.D., Goldstein, A.B., Heffernan, N.T.: Detecting the moment of learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6094 LNCS(PART 1), 25–34. (2010). https://doi.org/10.1007/978-3-642-13388-6_7
 14. Cetintas, S., Si, L., Xin, Y.P., Hord, C.: Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Trans. Learn. Technol.* **3**(3), 228–236 (2010). <https://doi.org/10.1109/TLT.2009.44>
 15. Mueen, A., Zafar, B., Manzoor, U.: Modeling and predicting students' academic performance using data mining techniques. *Int. J. Modern Educ. Comput. Sci.* **8**(11), 36–42 (2016). <https://doi.org/10.5815/ijmecs.2016.11.05>
 16. Zacharis, N.Z.: Classification and regression trees (CART) for predictive modeling in blended learning. *Int. J. Intell. Syst. Appl.* **10**(3), 1–9 (2018). <https://doi.org/10.5815/ijisa.2018.03.01>
 17. Ayub, M., Toba, H., Wijanto, M.C., Yong, S.: Modelling online assessment in management subjects through educational data mining. *Proceedings of 2017 International Conference on Data and Software Engineering, ICODSE 2017, 2018-January*, 1–6. (2018). <https://doi.org/10.1109/ICODSE.2017.8285881>
 18. Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D.J., Long, Q.: Predicting academic performance by considering student heterogeneity. *Knowl.-Based Syst.* **161**(December 2017), 134–146 (2018). <https://doi.org/10.1016/j.knosys.2018.07.042>
 19. Conijn, R., Kleingeld, A.: Influence of course characteristics , student characteristics , and behavior in learning management systems on student performance. *Nips.* (2016)
 20. Francis, B.K., Babu, S.S.: Predicting academic performance of students using a hybrid data mining approach. *J. Med. Syst.* (2019). <https://doi.org/10.1007/s10916-019-1295-4>
 21. Hasibur Rahman, M., Rabiul Islam, M.: Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques. 2nd International Conference on Electrical and Electronic Engineering, ICEEE 2017, December, 1–4. (2018). <https://doi.org/10.1109/ICEEE.2017.8412892>
 22. Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., Sarker, K. U.: Student Academic Performance Prediction by using Decision Tree Algorithm. In: 4th International Conference on Computer and Information Sciences: Revolutionising Digital Landscape for Sustainable Smart Society, ICCOINS 2018 - Proceedings, 1–5, (2018). <https://doi.org/10.1109/ICCOINS.2018.8510600>
 23. Hussain, S., Dahan, N.A., Ba-Alwib, F.M., Ribata, N.: Educational data mining and analysis of students' academic performance using WEKA. *Indones. J. Elect. Eng. Comput. Sci.* **9**(2), 447–459 (2018). <https://doi.org/10.11591/ijeecs.v9.i2.pp447-459>
 24. Sanvitha Kasthuriarachchi, K.T., Liyanage, S.R., Bhatt, C.M.: A data mining approach to identify the factors affecting the academic success of tertiary students in sri lanka. *Lect. Notes Data Eng. Commun. Technol.* **11**, 179–197 (2018). https://doi.org/10.1007/978-3-319-68318-8_9
 25. Al-Saleem, M., Al-Kathiry, N., Al-Osimi, S., Badr, G.: Mining educational data to predict students' academic performance. *Lect. Notes Comput. Sci. (Include. Subseries Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **9166**, 403–414 (2015). https://doi.org/10.1007/978-3-319-21024-7_28
 26. Widayahastuti, F., Tjhin, V.U.: Predicting students performance in final examination using linear regression and multilayer perceptron. In: *Proceedings - 2017 10th International Conference on Human System Interactions, HSI 2017*, 188–192, (2017). <https://doi.org/10.1109/HSI.2017.8005026>
 27. Bucos, M., Drăgulescu, B.: Predicting student success using data generated in traditional educational environments. *TEM J.* **7**(3), 617–625 (2018). <https://doi.org/10.18421/TEM73-19>
 28. Spickard, A., Alrajeh, N., Cordray, D., Gigante, J.: Learning about screening using an online or live lecture. *J. Gen. Intern. Med.* **17**(7), 540–545 (2002)
 29. Miguéis, V.L., Freitas, A., Garcia, P.J.V., Silva, A.: Early segmentation of students according to their academic performance: a predictive modelling approach. *Decis. Support Syst.* **115**(August), 36–51 (2018). <https://doi.org/10.1016/j.dss.2018.09.001>
 30. Huang, A.Y.Q., Lu, O.H.T., Huang, J.C.H., Yin, C.J., Yang, S.J.H.: Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interact. Learn. Environ.* **28**(2), 206–230 (2020). <https://doi.org/10.1080/10494820.2019.1636086>
 31. Sokkhey, P., Navy, S., Tong, L., Okazaki, T.: Multi-models of educational data mining for predicting student performance in mathematics: a case study on high schools in cambodia. *IEIE Trans. Smart Process. Comput.* **9**(3), 217–229 (2020). <https://doi.org/10.5573/IEIESPC.2020.9.3.185>
 32. Zacharis, N.Z.: Predicting student academic performance in blended learning using artificial neural networks. *Int. J. Artif. Intell. Appl.* **7**(5), 17–29 (2016). <https://doi.org/10.5121/ijaia.2016.7502>
 33. Aydoğdu, Ş: Predicting student final performance using artificial neural networks in online learning environments. *Educ. Inf. Technol.* **25**(3), 1913–1927 (2020). <https://doi.org/10.1007/s10639-019-10053-x>
 34. Vujičić, T., Matijević, T.: Comparative Analysis of Methods for Determining Number of Hidden Neurons in Artificial Neural Network. In: *Central European Conference on Information and Intelligent Systems* (pp. 219–223), (2016)
 35. Rakhmanov, O.: ScienceDirect ScienceDirect a comparative study on vectorization and classification techniques in sentiment analysis to classify student-lecturer comments. *Procedia Comput. Sci.* **178**, 194–204 (2020). <https://doi.org/10.1016/j.procs.2020.11.021>
 36. Hamoud, A.K., Hashim, A.S., Awadh, W.A.: Predicting student performance in higher education institutions using decision tree analysis. *Int. J. Interact. Multimed. Artif. Intell.* **5**(2), 26 (2018). <https://doi.org/10.9781/ijimai.2018.02.004>
 37. Davydenko, A., Fildes, R.: Forecast Error Measures : Critical Review and Practical Recommendations. *Business Forecasting: Practical Problems and Solutions*, January, 1–12, (2016). <https://www.researchgate.net/publication/284947381>
 38. Dien, T.T., Luu, S.H., Thanh-Hai, N., Thai-Nghe, N.: Deep learning with data transformation and factor analysis for student performance prediction. *Int. J. Adv. Comput. Sci. Appl.* **11**(8), 711–721 (2020)
 39. Grimm, K.J., Mazza, G.L., Davoudzadeh, P.: Model selection in finite mixture models: A k-fold cross-validation approach. *Struct. Equ. Model.* **24**(2), 246–256 (2017). <https://doi.org/10.1080/10705511.2016.1250638>
 40. West, R. P., Wride, M.A.: Examining large student cohorts—A question of questions. 1, 406–411, (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.