ISSN 1870-4069

Analyzing Students' Performance in a Mathematics Course Sequence using Educational Data Mining

Beatriz González Beltrán, Silvia González Brambila, Lourdes Sánchez Guerrero, Josué Figueroa González

[1]Universidad Autónoma Metropolitana, Unidad Azcapotzalco, Mexico City, Mexico

{bgonzalez,sgb,lsg,jfgo}@azc.uam.mx

Abstract. One of the main concerns in institutions of higher education is the time their students need to finish their studies. Students tend to invest more time than the established in approving a particular course or a group of them. This problem is more evident when there exists a sequence of related courses that are related through a prerequisite schema, because it is not clear whether the knowledge acquired in a previous course is appropriate for the next one. This work considers this problem and presents the use of Educational data mining for determining if there is an influence of some academic performance aspects in previous courses over the tries needed for approving a later one in a sequence of mathematics subjects. The performance of different measures for the predictive generated models using decision trees was not very high, and results shown that considering only the immediate previous course before the analyzed had little better results than considering all the ones in the sequence. Analyzing all the course sequence, many of the courses had as essential variables the performance not in the immediate, but two or three previous ones.

Keywords. Courses performance influence, educational data mining, information processing, prerequisite courses' influence, sequence of course analysis.

1 Introduction

Educational institutions generate a lot of information about several aspects of their operation that can be analyzed for finding patterns that lead to solve problems or making decisions about a certain aspects. One of the principal concerns in any educational institution is related to the performance of students across their studies. This is an aspect that concerns different authorities which are responsible of academic topics. Academic performance in a superior education institution can be measured in different ways like the percentage of graduated students, time invested in finishing a bachelor's degree and so on.

Also, the academic performance of students can be affected by several reasons: personal, labor, and academic which can be related to low preparation, study habits, even how students take their courses can be an influential factor for having a good or bad performance.

Discovering knowledge from educational data is an area that has been applied for more than two decades. The main goal of this area, known as Educational Data Mining (EDM) [4], is improving academic performance, reducing the failure rate in the studies, a particular topic or even in a single exam [3]. A way for achieving this is related with predicting student performance based on several personal or academic characteristics.

A concern for superior educational institutions is the time invested by students in finishing their university careers. Many students need twice or more the time for finishing, this represents both an operational and financial problem for institutions. One of the reasons for this is related to the time needed for approving one or more topics in a study plan because students invest more time than the established for approving a group of topics, primarily when these topics are related through a prerequisite schema, where students can not take a course if they have not approved the previous one.

A prerequisite schema is established considering that in a chain of courses, the previous ones will prepare students to take the next ones succesfully. It is considered that having a good performance in a previous course will imply having a successful result in the later one, however, this is not always true, there could be other factors that influence in the results obtained in later courses and not only the obtained mark in the previous ones.

Also, the results in, for example, a first course could have a certain impact not only in the immediate next one, but in two or more courses forward.

In the university where this study was performed, it is common that students invest more than twice the time needed for approving a group of topics, especially in the mathematical chain of courses which is composed by five with a prerequisite schema. Also, the way in which students can register their topics allows them not taking the next course in the following scholar period even if they have approved the previous one. This causes that many students do not take a course the next scholar period after approving the previous one, this could also be a factor that affects their performance in the following course.

It is necessary an analysis that considers several aspects of students' performance in a course for determining if these have an impact in the results obtained in later ones.

The goal of this work is to determine a relationship between the students' performance and the math courses sequence in engineering programs. Considered aspects involve:

- Approving mark.
- Amount of tries for approving.
- Number of scholar periods needed for approving.
- Time passed between approving a previous course and taking for first time the next one.

The performance measure is related to the number of tries to approve each course in the sequence more than the obtained mark. Former variables are used for predicting the number of tries that students will need for approving a later course after approved a previous one.

The results of this work could help to guide the student in choosing the moment of taking a course considering the performance over the previous ones. Also, it could help to determine if academic aspects have a student impact on the performance and even analyze the pertinence of the prerequisite schema in a study plan.

The structure of this paper is as follows: Section 2 reports some related works about predicting the academic performance of students considering different aspects. Section 3 presents the application of a data mining process for analyzing the data. Section 4 presents the analysis of the obtained results. Finally, section 5 offers the conclusions and future works.

2 Related Works

Predicting student's performance in a course or through its studies is one of the most studied topics in EDM. Many works focus in predicting the results of students considering personal, social or academic factors.

In [1], the authors use data mining to predict secondary school student performance in Mathematics and Portuguese courses. The database was build from school reports (grades and number of absences) and questionnaires with closed questions related to demographic, social, and school-related variables. They had three different DM algorithms: binary, 5-level classification, and regression. They applied a Naive Predictor (NV) as a baseline comparison and four DM methods: Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM). They obtained a high predictive accuracy when the first and second school period grades are known (91.9% with NV and 93% with DT on Mathematics and Portuguese, respectively). Moreover, the number of past failures is the most important factor when no student scores are available. There are other relevant factors, such as the number of absences, extra school support, travel time, the mother's job, going out with friends, and alcohol consumption.

In [6], is proposed a model to monitor student progress in e-learning systems to predict performance, progress, and potential. A student is represented by a Student Attribute Matrix, with attributes for performance and non-performance. Was considered learning styles, learning models, and Bloom's Taxonomy. To analyze the potential of a student was proposed a descriptor with higher-level attributes of performance and non-performance attributes. For the student performance estimator and the attribute causal relationship indicator was used a Back Propagation Neural Network. Results show correct and accurate student progress in high school students. Using this tool is possible to generate feedback indicators to understand and improve the performance of the students.

A comparative study on the effectiveness of educational data mining techniques to early predict student failure rate in introductory programming courses

is presented in [2]. The authors tried to reduce the failure rate identifying problems at early stages and performing data preprocessing and algorithms fine-tuning. The comparison included Neural Networks, Decision Trees, Support Vector Machine (SVM) and Naïve Bayes, using an F score to evaluate the effectiveness. The SVM technique outperformed the other ones. The data sources were from Brazilian Public University in distance and on-campus education, with 21 and 16 characteristics, respectively.

In [5], the authors designed a classification model by utilizing data mining techniques for predicting the likelihood of a student to pass the Licensure Examination for Teachers (LET). The authors compared five data mining techniques (Neural Network, Support Vector Machine, C4.5 Decision Tree, Naïve Bayes, and Logistic Regression). This work identified the predictive variables on measuring the student's performance in the LET at the Bulacan Agricultural State College. As a result of the analysis, C4.5 Decision Tree had the highest accuracy of 73.10% followed by Neural Network with 65.67%. For that reason the authors used this algorithm for design the classification model. Using the model, the Bulacan Agricultural State College could be able to identify students who will likely fail the LET. These students will be given higher priority during their mock board review and be able to pass the board examination. Studies show that mock board performance has a significant relationship in passing the board examination.

3 Looking for Relationships between Courses

For determining if there was a relationship between the performance in a previous mathematics course and the next one, was used the data mining methodology which involves the general stages of: data exploration and preprocessing, modeling and analysis of the results. The last stage is presented in Section 4 of the document.

3.1 Data Exploration and Preprocessing

Processed data were obtained from the students' marks database, known as *kardex*, which contains information about the courses that students have taken during their studies. These data include student id, course id, obtained mark (approved or not), and scholar period when the course was taken. Were used the information of the five courses of the mathematical branch:

- Mathematics Workshop (MW),
- Essential Calculus (EC),
- Differential Calculus (DC),
- Integral Calculus (IC),
- Differential Equations (DE).

All of these courses have a prerequisite schema, meaning that for taking one, it is necessary to have approved the previous one. The relationship between these courses is shown in Figure 1.

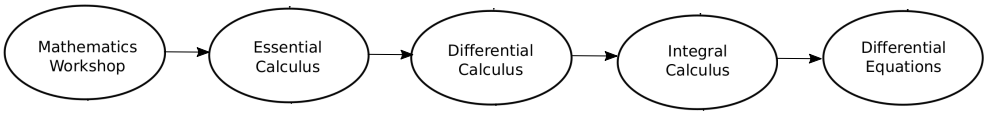

Fig. 1. Mathematics courses sequence.

Information used for evaluating the performance in each group, as mentioned in the Introduction section, was:

- Obtained mark: the mark with the one the student approved the course.
- Number of tries needed for approving a course: amount of tries the student needs for approving the course.
- Scholar periods invested in approving a course: total of scholar periods needed for approving a course, consider that not necessarily correspond to the number of tries. A student could have approved in their second try, but invested more scholar periods.
- Number of scholar periods passed after approving one course and taking for the first time the next one: it is common that students do not take the later course the next scholar period after approving the previous one.

Table 1 presents the variables names, their description and domain.

For the study, were considered engineering students from all the study plans offered by the university which already have approved the five courses, the total of processed students was 2,119.

Kardex file only contains obtained marks and scholar periods, so it was necessary to process data for obtaining the first scholar period when the student took the course and calculate the number of periods invested in approving it. The number of tries was obtained, for each course, counting the total of non-approving marks obtained plus the one when it was approved. The time elapsed between approving one course and taking the following was measured considering the scholar period when a previous course was approved and the one in which the student took the following course for the first time.

3.2 Modeling

Were used decision trees for analyzing the information, the predicted variable was the total of tries needed for approving each course. Were considered two scenarios, using only the performance in the previous course, and considering the performance in all the previous courses according to Figure 1. In this way, were analyzed the following relationships:

- Relationship 1. TEC $\sim$ (MMW, PNMW, TMW, TPMW),
- Relationship 2. TDC $\sim$ (MEC, PNEC, TEC, TPEC),
- Relationship 3. TIC $\sim$ (MDC, PNDC, TDC, TPDC),
- Relationship 4. TDE $\sim$ (MIC, PNIC, TIC, TPIC).

Table 1. Description and domain of used variables.

Variable name	Description	Domain
MMW	Mathematics Workshop mark	Very Good, Good, Sufficient
TMW	Number of tries for approving Mathematics Workshop	1, 2 or more
PNMW	Number of periods needed for approving Mathematics Workshop	1 to 13
TPMW	Number of scholar periods passed after approving Mathematics Workshop and studying Essential Calculus	1 to 15
MEC	Essential Calculus mark	Very Good, Good, Sufficient
TEC	Number of tries for approving Essential Calculus	1, 2 or more
PNEC	Number of periods needed for approving Essential Calculus	1 to 17
TPEC	Number of scholar periods passed after approving Essential Calculus and studying Differential Calculus	1 to 13
MDC	Differential Calculus mark	Very Good, Good, Sufficient
TDC	Number of tries for approving Differential Calculus	1, 2 or more
PNDC	Number of periods needed for approving Differential Calculus	1 to 15
TPDC	Number of scholar periods passed after approving Differential Calculus and taking Integral Calculus	1 to 6
MIC	Integral Calculus mark	Very Good, Good, Sufficient
TIC	Number of tries for approving Integral Calculus	1, 2 or more
PNIC	Number of periods needed for approving Integral Calculus	1 to 16
TPIC	Number of scholar periods passed after approving Integral Calculus and taking Differential Equations	1 to 11
TDE	Number of tries for approving Differential Equations	1, 2 or more

Considering the sequence of previous topics for each course (EC only has one previous course), the analyzed relationships were:

- Relationship 5. TDC $\sim$ (MMW, PNMW, TMW, TPMW, MEC, PNEC, TEC, TPEC),
- Relationship 6. TIC $\sim$ (MMW, PNMW, TMW, TPMW, MEC, PNEC, TEC, TPEC, MDC, PNDC,TDC, TPDC),
- Relationship 7. TDE $\sim$ (MMW, PNMW, TMW, TPMW, MEC, PNEC, TEC, TPEC, MDC, PNDC,TDC, TPDC, MIC, PNIC, TIC, TPIC).

Decision trees were generated using the CART algorithm and a ten-fold cross-validation was used for each relationship. Although the number of tries has values from 1 to 5 (5 is the maximum of tries for approving a course in the university), most of the cases correspond to 1 or 2. The first results considering the five categories (one per try) have an accuracy less than 40%. For this reason, 2 to 5 tries were considered as a single category: 2 or more (2M). Besides the accuracy of prediction, also was measured the importance of each variable for each relationship.

The performance of the model was measured using accuracy, precision, recall and F_1 score metrics based on confusion matrices. Instead accuracy is the most intuitive measure, it is only valid when there are symmetric data sets, in this case, the amount of students with 1 try was greater that those ones with 2, 3, 4 or 5 (2 or more), so accuracy did not give a good measure of the results. other measures offers a better measurement of model performance where there is an unbalanced class distribution.

3.3 Obtained Results

Accuracy of the predictive models considering only the relationship between a course and the previous one are presented in Table 2.

Table 2. Performance measures for immediate course relationships.

Relationship	Accuracy	Precision 1 try	Precision 2M tries	Recall 1 try	Recall 2M tries	F_1 score 1 try	F_1 score 2M tries
1	0.652	0.790	0.431	0.689	0.563	0.7365	0.487
2	0.650	0.878	0.230	0.677	0.508	0.764	0.316
3	0.656	0.905	0.159	0.686	0.441	0.778	0.226
4	0.659	0.859	0.271	0.696	0.499	0.769	0.348

The importance of each variable for the relationships that only considers the immediate previous course performance are:

- For relationship 1: MMW 51.709, PNMW 33.456, TMW 10.686, and TPMW 8.453,
- For relationship 2: MEC 13.491, PNEC 24.588, TEC 14.951, and TPEC 8.116,
- For relationship 3: MDC 9.211, PNDC 32.812, TDC 22.904, and TPDC 2.806,
- For relationship 4: MIC 20.39, PNIC 30.760, TIC 25.383, and TPIC 8.742.

Figure 2 shows an evolution of the importance of the variables mark, periods needed, tries and time elapsed at the moment of predicting the total of tries needed for approving a course considering only the immediate previous one.

Table 3 shows the results of the models generated considering the influence of all the courses that compose the sequence before the predicted one.

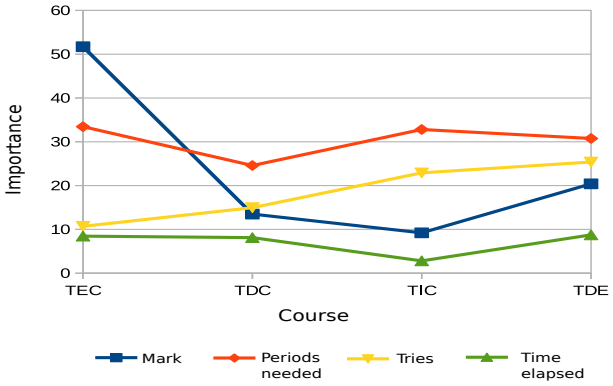

Fig. 2. Evolution of crietria importance in the prediction.

Table 3. Performance measures for all courses sequence relationships.

Relationship	Accuracy	Precision 1 try	Precision 2M tries	Recall 1 try	Recall 2M tries	F_1 score 1 try	F_1 score 2M tries
5	0.646	0.824	0.284	0.701	0.445	0.757	0.346
6	0.629	0.806	0.284	0.686	0.429	0.741	0.341
7	0.624	0.730	0.366	0.704	0.446	0.726	0.4

The importance of each variable was also measured for relationships that consider not only the previous course but also the ones that compose the sequence until the studied one. Table 4 presents the importance of variables across the complete sequence, consider that the importance of zero represents that a variable corresponds to a course that is taken after the one analyzed.

Figure 3 presents a radar chart showing the importance of each variable in the analyzed courses of the mathematical course sequence.

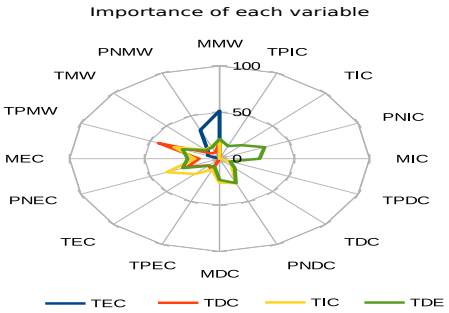

Fig. 3. Importance of all variables in the complete courses sequence.

Considering that predicted value for Relationship 2 was the same that for

Table 4. Importance of each variable through the courses sequence.

Variable	TEC	TDC	TIC	TDE
MMW	51.709	17.222	19.520	21.352
PNMW	33.465	7.190	14.349	13.102
TMW	10.686	8.267	13.439	12.921
TPMW	8.453	44.012	32.664	26.652
MEC	0	13.710	16.553	21.271
PNEC	0	23.262	37.900	26.525
TEC	0	10.093	23.360	10.983
TPEC	0	13.394	12.673	9.025
MDC	0	0	25.704	22.910
PNDC	0	0	28.670	28.406
TDC	0	0	11.729	14.363
TPDC	0	0	8.123	7.738
MIC	0	0	0	26.693
PNIC	0	0	0	32.471
TIC	0	0	0	20.604
TPIC	0	0	0	14.729

Relationship 5 (TDC), similar for Relationship 3 and Relationship 6 (TIC) and for Relationships 4 and 7 (TDE), a comparison of performance measures obtained using both schema of relationship was performed, results are shown in Table 6.

In the same way, Table 5 presents the most important variable in the immediate course and in the complete course sequence relationship. Also the place that the most important variable in the immediate course relationship occupied considering all the courses sequence is presented in the last column. Consider that the TEC relationship is discarded because it is the same for immediate and full sequence relationship.

Table 5. Most important variable comparison for both schemes.

Variable	Immediate relationship	Full sequence relationship
TDC	PNEC	TDC
TIC	PNDC	PNEC
TDE	PNIC	PNIC

4 Results' Analysis

Results' analysis involved the analysis of the values of used performance measures of the obtained models for both relationships, the immediate course and the full course sequence. Also the importance of each variable in the prediction was analyzed.

Table 6. Comparison of immediate and full course sequence performance measures.

Measure	Rel. 2	Rel. 5	Rel. 3	Rel. 6	Rel. 4	Rel. 7
Accuracy	0.650	0.646	0.656	0.629	0.659	0.624
Precision 1 try	0.878	0.824	0.905	0.806	0.859	.730
Precision 2M tries	0.230	0.284	0.159	0.284	0.271	0.366
Recall 1 try	0.677	0.701	0.686	0.686	0.696	0.704
Recall 2M tries	0.508	0.445	0.441	0.429	0.499	0.446
F1 score 1 try	0.764	0.757	0.778	0.741	0.769	0.702
F1 score 2 tries	0.316	0.346	0.226	0.341	0.348	0.4

4.1 Immediate Course Relationship

From the relationships that involved only the immediate course, were obtained a low accuracy (65% as maximum), however this could be due to an unbalanced distribution in the classes, where most of them had the value of 1 try. Other measures had a better performance, precision which shows how many classes labeled as one try actually correspond to that value had the biggest value with 90%, which corresponds to the third relationship, the number of tries needed for approving Integral Calculus (TIC). Other relationships also had acceptable values with a minimum of 79%. However, the precision for classes labeled with 2M class was very poor with a maximum of 43% and a minimum of 15%. This shows that the generated model fails at the moment of classifying those students which need 2 or more tries for approving a course.

Recall value indicates the classes that correspond to a specific value and were labeled in the same way. Cases that correspond to the one try class did not obtain very good results, with values a little higher than the ones of the accuracy, a maximum of 69% and a minimum of 67%. For the classes labeled as two or more tries, results were also very low, with a maximum of 56% and a minimum of 44%.

For the F_1 score, which is a weighted average of recall and precision, the results can be considered barely appropriate, with a maximum of 77% and a minimum of 73% for the ones labeled as 1 try. Again, cases of 2 or more class, obtained poor results with a minimum of 22% and a maximum of 48%.

About the importance of each variable in the prediction of the total of tries needed for approving a course, as Figure 2 shows, the most important variable for the second course of the sequence, EC, was the mark obtained in MW. However, the obtained mark importance decreased through the sequence of courses.

Two variables shew a continual behavior, the number of scholar periods needed for approving and the number of tries. The number of scholar periods had values between 25 and 35 through all the sequence. Number of tries shew an increase across the course sequence.

The variable that represents the number of scholar periods passed after approving a course and taking the later one always had the least importance in the prediction.

4.2 Full Course Sequence Relationship

Results considering all the courses in the sequence had a similar behavior that the ones obtained with immediate course relationship, even, including all the courses in the sequence before the predicted one, reduced a little the value of the performance measures.

Accuracy of models had values from 62% to 64% which is considered low, again the best results were given by precision with a minimum of 73% and a maximum of 82% in the classification on cases labeled with one try. Similar to the former analysis, cases labeled as two or more tries had a low value. Recall measure also had better results for one trial class with a minimum of 68% and a maximum of 70% meanwhile cases labeled as of two or more tries had a minimum of 42% and a maximum of 44%.

F_1 score also shew a similar behavior than for immediate course relationship with a minimum of 72% and a maximum of 75% for one try class and 34% as minimum and a maximum of 40% for two or more classes.

As Figure 3 shows, for EC, the most important variable at the moment of prediction was MMW; for DC, the most important variable was not one of the previous course (EC), but the time passed after approving MW, followed by the scholar periods for approving the previous course EC and then, the mark in MW. This represents that two of the three most important variables were not from the immediate previous course, but from the previous one of this.

For TIC, only one of the three most important variables corresponded to the immediate previous course (PNDC), and the other two to different courses (MW and EC). Notice than the most important variable (TPMW) was from the first course of the sequence.

Finally, for TDE, two of the three most important variables corresponded to the previous course (IC), including the most important (PNIC), then one of DC and finally, another from the previous course (MIC). It is interesting that the fourth in importance, corresponded again to the first course of the sequence (TPMW).

4.3 Comparing Both Immediate and Full Sequence Courses Schemes

Table 5 shows that from the three relationships, two of them change their most important variable in the prediction. However, analyzing Table 4, was found that:

- PNEC, the most important variable in immediate relationship for TDC, in the full sequence relationship occupied the second place.
- PNDC, which was the most important for predicting TIC in the immediate relationship, had the third place in the full sequence one.
- PNDC, remains as the most important variable for both relationships in the prediction of TDE.

It is also of interest notice that in the immediate next course, three of four courses has as the most important variable the periods needed for approving the previous one. Also, in the three full course sequence relationships, the periods needed for approving a course (immediate previous or not) was the most important in all of them.

As values of Table 6 show, better results for almost all the performance measures were obtained considering only the previous course, meaning that adding more variables corresponding to the students' performance in all the previous courses of the sequence did not really help to improve the prediction of the number of tries.

5 Conclusions

The goal of this work was determining if there exists an influence of the academic performance in a course over the number of tries needed for approving the next one. For this, were considered several performance measures and tested two relationships, one considering only the immediate previous course and another which considered the complete sequence of previous courses. This, for a sequence of five mathematics courses.

According to the results of the different performance measures used for evaluating the predictive model generated using decision trees, the results did not show a convincing evidence that analyzed criteria had an important relationship with the amount of periods. Best results were obtained using precision as performance measure, but only for those students that approve in their first try.

Comparing the results of both relationships, considering only the immediate course obtained a little better results for almost all the performance measures.

Analyzing the importance of each variable, it is of interest noticing that the obtained mark in a previous course had not a significant relevance over approving the next course at the first try, only in one of the four relationships, this variable was the most important.

The variables which appeared more times as the most important ones were related to the number of scholar periods needed for approving a previous course, from the seven studied relationships, a variable related to this aspect, appeared in six.

Something remarkable is related with the time passed after approving a course and taking the next one, it was expected that at greater this time, the number of tries tend to rise, however, this variable always appeared as the least important in all the relationships.

A similar analysis for the complete courses sequence shows that it is common that variables related to previous courses more than the immediate one, had certain importance in the amount of tries needed for approving one. Interestingly, that all of the courses have a variable considered important which correspond to the Mathematics Workbench course, the first of the sequence, which at the university is a leveling course.

It can be said that considering only the performance in a immediate or all the courses of a sequence did not offer a definitive idea about how would be the performance of students in later ones.

Future works related to this study will be directed in obtaining better results for the performance measures of the models. A first option is using different predictive algorithms, however, it is considered that a bigger set of data could offer better results. Adding other characteristics like personal ones, or the behavior during the courses including for example, use of different platforms, amount of study hours, number of exercises, level of interaction with professors and other could help to model in a better way the behavior of the student, so that the models could generate better results.

References

1. Cortez, P., Silva, A.: Using Data Mining to Predict Secondary School Student Performance. In: Brito, A., Teixeira, J. (eds.) Proceedings of the 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008). pp. 5–12. EUROSIS, Porto, Portugal (Apr 2008)
2. Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F., Rego, J.: Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73, 247–256 (2017)
3. Kumar, V., Chadha, A.: An empirical study of the applications of data mining techniques in higher education. International Journal of Advanced Computer Science and Applications 2(3) (2011)
4. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6), 601–618 (2010)
5. Rustia, R.A., Cruz, M.M.A., Burac, M.A.P., Palaoag, T.D.: Predicting student's board examination performance using classification algorithms. In: Proceedings of the 2018 7th International Conference on Software and Computer Applications. pp. 233–237. ACM (2018)
6. Yang, F., Li, F.W.: Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers & Education 123, 97–108 (2018)