


RESEARCH ARTICLE

Open Access



Examining students' course trajectories using data mining and visualization approaches

Rabia Maqsood^{1*} , Paolo Ceravolo², Muhammad Ahmad¹ and Muhammad Shahzad Sarfraz¹

*Correspondence:
rabia.maqsood@nu.edu.pk

¹ National University
of Computer and Emerging
Sciences, Chiniot-Faisalabad
Campus, Chiniot 35400, Pakistan

² Università degli Studi di
Milano, Via Giovanni Celoria 18,
20133 Milan, Italy

Abstract

The heterogeneous data acquired by educational institutes about students' careers (e.g., performance scores, course preferences, attendance record, demographics, etc.) has been a source of investigation for Educational Data Mining (EDM) researchers for over two decades. EDM researchers have primarily focused on course-specific data analyses of students' performances, and rare attempts are made at the domain level that may benefit the educational institutes at large to gauge and improve their institutional effectiveness. Our work aims to fill this gap by examining students' transcripts data for identifying similar groups of students and patterns that might associate with these different cohorts of students based on: (a) difficulty level of a course category, (b) formation of course trajectories, and, (c) transitioning of students between different performance groups. We have exploited descriptive data mining and visualization methods to analyze transcript data of 1398 undergraduate Computer Science students of a private university in Pakistan. The dataset includes students' transcript data of 124 courses from nine distinct course categories. In the end, we have discussed our findings in detail, challenges, and, future work directions.

Keywords: Educational data mining, Course trajectories, Hierarchical clustering, Markov chain

Introduction

Educational institutes (e.g., schools, colleges, and universities) acquire an abundance of data related to students, which has been a concern of investigation for Educational Data Mining (EDM) researchers for over two decades. Students-related data come from many different sources including their demographics, scores in assignments and quizzes, course grades, course preferences, students' logged interactions with a learning management system, etc. (Romero & Ventura 2013). These different kinds of data sources have led researchers to analyze students' characteristics from various perspectives. For example, grade prediction, clustering similar learning styles, identifying at-risk students, creating (personalized) student models, course recommendation, and many more (Baker & Yacef, 2009; Dutt et al., 2017).

As highlighted in the extensive literature review performed by Dutt et al. (2017), EDM researchers have primarily focused on course-specific data analyses of students' performances, and rare attempts are made at the domain level that may benefit the educational

institutes at large to gauge and improve their institutional effectiveness. This research work is an attempt to fill that gap by extracting information from students' transcripts data that will be useful to determine the characteristics of undergraduate Computer Science students in a private university in Pakistan and make future decisions.

This work focuses on examining undergraduate students' transcripts data which contain information about each student's performance scores/grades in all the taken courses. An undergraduate program offered by a university either follow a *flexible curriculum* that allows students to take any combination of courses, or a *structured curriculum* wherein courses are labeled as core (or mandatory) and electives, and pre-requisite relations are also defined between different courses. Despite the curriculum design, students are generally required to take courses from multiple disciplines (e.g., Mathematics, Social Sciences, Computer Science, Engineering, etc.) in any degree program. We refer to these course disciplines as *course categories* in the remaining text.

Usually, a student's performance is summarized quantitatively for all the taken courses by a semester grade point average (GPA) and an overall cumulative grade point average (CGPA). These quantitative performance measures are used to determine a student's progress however, they do not provide a complete picture. Some course categories might be favorably easy or more difficult for a cohort of students which can be useful to identify the strengths and weaknesses of those students. On the other hand, a student's performance level may change over time as he/she progresses to advanced courses. Also, different courses belonging to a particular category may vary in difficulty levels. Thus, it is important to examine students' course trajectories to comprehend their varying levels of performance in distinct course categories that can reveal useful insights for teachers, student advisors, and university administration. To the best of our knowledge, this topic is not addressed in a detailed manner in the existing literature.

Data mining methods are broadly divided into two groups: *descriptive* and *predictive* (Tan et al., 2016). In this work, we focused on the descriptive methods of data mining that support us to derive patterns from the input data and answer the question "*what has happened in the past?*". Data visualization and clustering (or cluster analysis) are very prominent methods in data mining for this purpose. Findings of the existing works have shown that clustering can be useful for identifying characteristics that are pertinent to different groups of students (Dutt et al., 2017).

The objectives of this research study are to analyze students' transcripts data for identifying groups of students with similar performances in distinct course categories and differentiating between their characteristics. The dataset used in this study was obtained from the Department of Computer Science of the National University of Computer and Emerging Sciences (NUCES), Pakistan. We formulated four well-defined research questions as given below, which specifies in detail the scope of this research study.

- RQ-1: How can we identify different performance groups of students based on distinct course categories?
- RQ-2: What are the course categories that can be related as *hard* and *easy* for students in the pre-identified performance groups?
- RQ-3: Is there a difference in course trajectories of different students' performance groups?

- RQ-4: How likely students can migrate from one performance group to another?

The rest of the paper is organized in the following manner. A review of the existing related works is presented in [Related works](#) section. In [Data description](#) section, we provided data description and steps performed for cleaning purposes. Details of the methodology to answer the pre-mentioned research questions are given in [Methodology](#) section. Finally, we present our conclusions in [Conclusions and recommendations](#) section.

Related works

Clustering or unsupervised learning has been employed in the EDM domain for various purposes, for example, grouping students using their demographics and other characteristics, academic performance data, and logged interactions (Romero & Ventura, 2013). Some researchers have also focused on pattern mining techniques to find frequently occurring behaviors in students' data. We reviewed the existing literature with an emphasis on studies that have employed data mining and visualization techniques to examine students' course or grade data for analyzing their performance behaviors.

The most extensive related work is performed in Darcan and Badur (2012) for student profiling using their course grades data. The dataset comprises 467 students, 31 courses, and obtained grades of each student in all the taken courses. Factor analysis was used for dimensionality reduction to minimize the heterogeneity in the data that may occur due to 31 different courses. As a result, four factors were obtained that grouped courses into programming, mathematics, managerial, and theoretical disciplines. Then, K-means clustering was applied to the pre-processed data with $k = 6$ as the most suitable value. The obtained clusters were found to have associations with respect to students' different performance groups and the type of high school.

In Almatrafi et al. (2016), authors have compared the course-taking patterns of 630 low and high-performance achieving students belonging to three different majors of Engineering discipline. The Apriori algorithm is applied with a support of 0.25 to find the frequent courses taken in a semester. Results are visualized by a graphical structure with nodes representing the courses and edges representing the relation between the courses. Analyses of the obtained charts have shown clear differences in the combination of courses taken by low and high-performing students. Another similar study is done in Yoo et al. (2017) in which the PrefixSpan algorithm is used to mine frequent sequential patterns of courses taken by diverse groups of students in different semesters. In particular, their findings have shown distinct combinations of courses that were found to be relatively easy or harder for students who graduated with CS degrees, male and female students, and, students who dropped out or changed their major.

Ferral (2005) explored relationships between students' subject choices within three different years and their demographics which include gender, school, and ethnic groups. The hierarchical agglomerative clustering method was used to find similar groups of students based on courses taken using the Jaccard similarity measure. Their results showed that the formed clusters have a strong relationship with all the demographic variables.

Students' academic results were analyzed by Priyambada et al. (2017) to perform curriculum mining and find discrepancies between the originally designed curriculum

versus the one actually followed by students. This behavior profile was merged with the performance profile for each student which contains his/her semester grade point average. Three clusters were obtained using K-means clustering, relating to different performance groups of students (i.e., high, average, and low). Their profile-based cluster evolution analysis model showed that students depicted different migration patterns in each semester. Another related application of K-means clustering is found in Tuyishimire et al. (2022) to cluster students' learning behaviors using their performance scores in subsequent quizzes. Their proposed model allows for dynamic monitoring of students' progress in subsequent quizzes and identifies poor and high-performing students. Students' cluster migration analysis was performed to identify a positive or negative change in a student's individual performance in quizzes over time.

Some other essential variables describing students' characteristics have also been used in the literature to find similar groups of students for different purposes. For example, Bresfelean et al. (2008) applied K-means clustering to determine students' academic failure/success profile using their responses to a questionnaire that includes their background information and perspective towards the quality of education. The obtained two clusters comprise students who passed all the exams in the last semester and the ones who failed at least one course. Sya'iyah et al. (2019) on the other hand used the following four variables to identify cohorts of students, namely: grade point average, length of study, English proficiency score, and length of thesis writing. Three clusters were formulated using the K-means clustering method to differentiate between high, average, and, low performing students.

To summarize, study of the existing literature has shown that different data visualization and unsupervised learning or clustering methods are used to examine students' course trajectories. And, main focus has been given to students' enrollment data, that is, the combination of courses taken in a semester irrespective of students' attained performances in those courses, see for example, (Almatrafi et al., 2016; Ferral, 2005; Priyambada et al., 2017; Yoo et al., 2017). In fact, only two related works have considered students' performance scores in the courses (Darcan & Badur, 2012) and quizzes (Tuyishimire et al., 2022). Our research questions as stated in section: sec: Intro, suggest that this work aims to fill some research gaps in the existing literature by examining students' transcripts data for identifying groups of students with similar performances in distinct course disciplines and analyzing their characteristics.

Data description

We retrieved transcript data of undergraduate Computer Science students of NUCES, Pakistan. The data was obtained before the Covid-19 pandemic period, i.e., from Fall 2012 to Fall 2019 (22 semesters in total including Fall, Spring, and Summer terms). The original data contains 58,113 transcript records of 2073 students; each record representing the details of a course taken by a student, e.g., enrolled semester, course code, course name, credit hours, earned letter grade, earned grade points (a numeric value), semester grade point average (SGPA), cumulative grade point average (CGPA), etc.

The university offers a very structured BS Computer Science (BSCS) program for undergraduate students. That is, students are required to pass a variety of courses from multiple disciplines which include some *core* (i.e., mandatory) and *elective* (i.e., optional)

Table 1 Summary of the course categories

Course code	Course category	New label	Unique no. of courses taught
CS*	Theory/Non-Programming	NP	34
	Programming	P	22
CL	Programming Lab	PL	10
MT	Mathematics	MT	9
EE	Electrical Engineering	EE	13
EL	Electrical Engineering Lab	EL	4
MG	Management	MG	12
SS	Social Science	SS	17
SL	Social Science Lab	SL	3
	Total course categories = 9		Total no. of courses = 124

*The CS course category is further split into two types by two domain experts to separate the programming and theoretical/non-programming courses

Table 2 Summary of the cleaned data

Total no. of semesters	15
Total no. of students	1398
Total no. of semester-wise students' records	10,211
Total no. of enrolled courses	50,392

courses. Course contents may be updated at the beginning of a term, however, to maintain quality, students (studying in different sections) are taught the same contents of a course offered in a specific semester. Each course is assigned a unique code that comprises two alphabet letters followed by three digits (for example CS101, MT104, etc.). The first two letters define the discipline of a course; for example, CS stands for Computer Science and MT stands for Mathematics. A summary of all the course categories is shown in Table 1. In total, we have 9 course categories, 6 of them are 3-credit hours courses¹ and 3 of them are 1-credit hour lab courses (ending with a label 'L' i.e., PL, EL, and SL). Each lab course is offered with a core course from a corresponding category.

Data cleaning: After performing some preliminary data analysis, we cleaned the data by removing the following records: (a) summer semester records were deleted due to fewer course enrollments, (b) records of students with insufficient information is deleted (i.e., transcript data is missing or students who have studied only one semester), and, (c) final year project grades were deleted due to its different nature than the regular courses (a team of 2-3 students works on a project in the last two semesters). Table 2 contains the summary of the cleaned data which includes the transcript records of current, graduated, and terminated students.

¹ Except for the Social Science (SS) course category whose courses are of 2-credit hours.

Table 3 The original 12-level letter grades and equivalent numeric grade points

Letter Grade	A+	A	A–	B+	B	B–	C+	C	C–
Grade Point	4.00	4.00	3.67	3.33	3.00	2.67	2.33	2.00	1.67
Letter Grade	D+	D	F						
Grade Point	1.33	1.00	0.00						

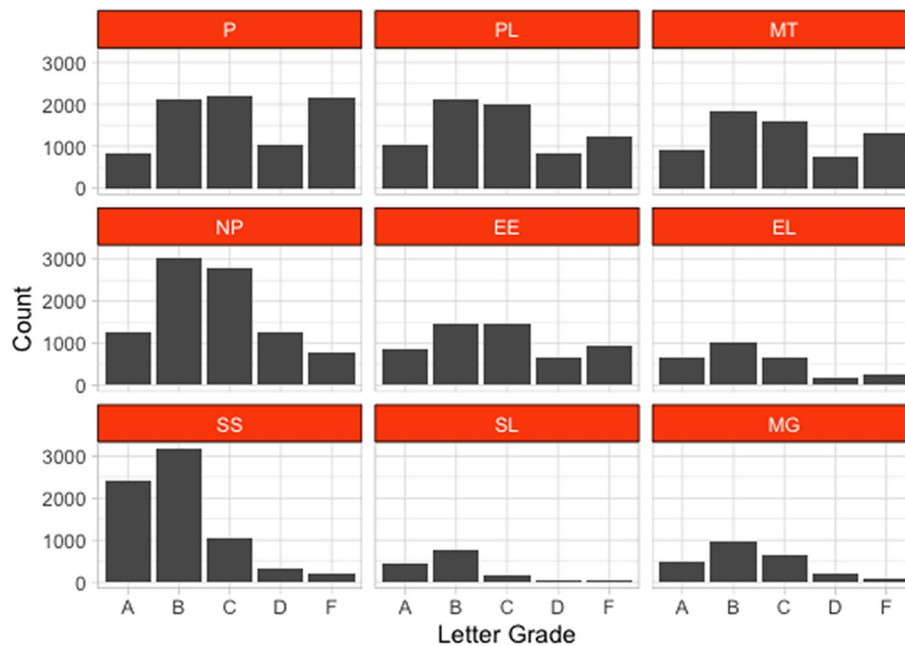


Fig. 1 Grades distribution per course category

The university follows a 12-level letter grades scheme as shown in Table 3; each letter grade is shown with its equivalent numeric grade point which is used to calculate a student’s SGPA and CGPA. A CGPA of 2.00 is the least requirement for graduation (which equals to C grade). As can be expected, bar chart visualization of students’ grades using the original 12-level scheme resulted in a highly varied multimodal distribution. To minimize the potential effect of different unknown variables (not available in the dataset, e.g., faculty details, students’ gender, previous academic background, etc.) causing a high variance in the students’ obtained grades, we reduced the letter grades to 5-level by merging all the A’s together, B’s together, and so on. A bar chart of the students’ grades (using the 5-level scheme) per course category is shown in Fig. 1.

The following five course categories have the most ‘F’ grades: programming (P), programming lab (PL), mathematics (MT), non-programming (NP), and, electrical engineering (EE) courses. To understand if a course type (which can be either *core* or *elective*) have any impact on grades distribution, we plot another bar chart, see Fig. 2. Not to much surprise, we can see that the core courses have maximum failed attempts (or F grades) in all the course categories. The bar charts in Fig. 2 also reveal two important facts about the course categories; first, elective courses do not have any lab (see that PL, EL, and SL are all empty); secondly, there is no MG core course. In general, course

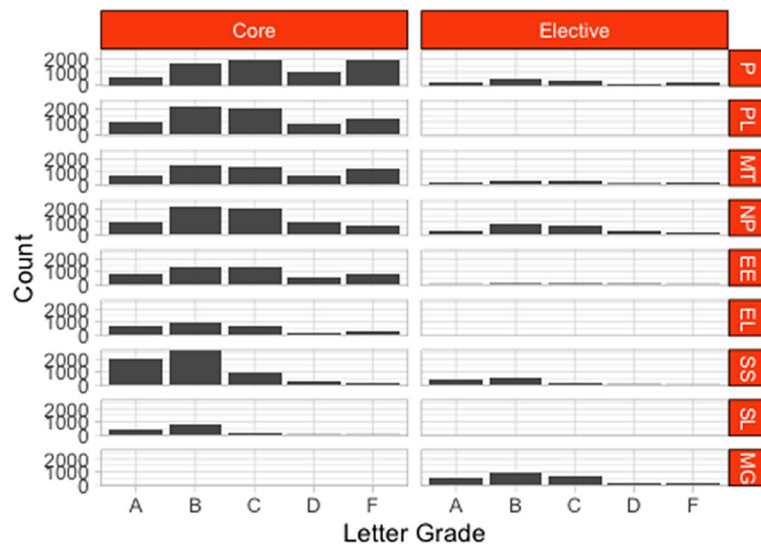


Fig. 2 Grades distribution for core and elective courses in each course category

Table 4 Preprocessed data sample for performing clustering—each column represents the frequency counts for a specific course category and a letter grade; each row represents a unique student’s performance profile

	P/A	P/B	P/C	P/D	P/F	PL/A	PL/B	...	MG/D	MG/F
St1	4	2	0	0	0	5	1	...	0	0
St2	1	3	3	0	0	2	3	...	0	0
St3	0	2	4	0	0	1	4	...	0	0
St4	0	0	2	3	2	0	0	...	1	0
St5	0	0	0	4	3	0	0	...	2	1
...

categories having both core and elective courses almost follow the same distribution—for example, in programming (or P), grades ‘B’, ‘C’ and ‘F’ are in high frequency for core and elective courses. Thus, we do not differentiate between the courses based on their types in further analyses.

Methodology

To answer these research questions, we employed different data mining and visualization techniques, as described in the following sub-sections.

Data preprocessing

To identify groups of students with similar performances in different course categories, we computed the count of each letter grade per course category for all the students. In other words, students’ transcript data was converted into a 45 columns table by combining 9 course categories and 5 letter grades, wherein each row represents a unique student’s performance profile. Table 4 shows hypothetical performance data of five students wherein each column represents the frequency counts for a specific course category and

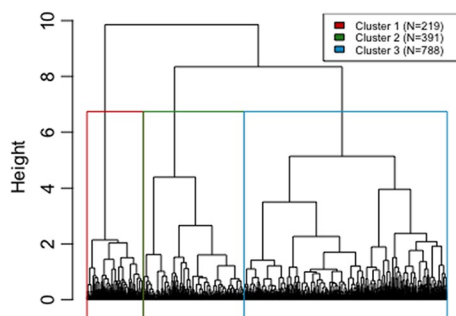


Fig. 3 The resultant dendrogram of the Hierarchical clustering algorithm performed on students' performance profiles with three major clusters

a letter grade. For example, P/A (the second column) means programming course category with an A letter grade. Following this terminology, row 1 shows that Student1 (or St1) has got 4 A and 2 B letter grades in programming (P) courses; 5 A and 1 B letter grades in programming lab (PL) courses. This shows that Student1 has done quite well in the programming courses. The data records of Student2 and Student3 suggest that they have achieved average performance in the programming-related courses. Subsequently, the performance data of Student4 and Student5 show that they struggled in the programming-related courses (i.e., P and PL) as well as the management (MG) courses.

Identification of students' performance groups

In our first research question (RQ-1), we intend to find groups of students having similar performances in different course categories. Given the sparse nature of our data (as illustrated in Table 4), we calculated Cosine Similarity measure between all the students. Then, we performed Hierarchical clustering algorithm using Ward's linkage method to minimize variance between the clusters (Tan et al., 2016). Once the clusters are formed, Hierarchical clustering produces a dendrogram as given in Fig. 3, which shows that how objects were grouped together to form clusters in a bottom-up manner. The next step is to identify the desired number of clusters, resultant dendrogram should be cut by a horizontal line where the line gets a maximum difference in height (on the y-axis). In our case, the maximum distance between clusters was found between height 5 and 8, hence, we retrieved the following three clusters: Cluster 1 with 219 students, Cluster 2 with 391 students, and Cluster 3 with 788 students.

To make sense of the resultant clusters, we plot the final CGPAs of the students belonging to each cluster. Figure 4 shows the boxplot for the three clusters, each boxplot also contains a mean value of the final CGPAs (marked as yellow diamonds). The mean final CGPA for Cluster 1 is 1.6; 3.15 for Cluster 2 and 2.42 for Cluster 3. Clearly, the distinct mean values of the three clusters show differences in students' academic performances. We used these mean final CGPA values to label each cluster with a suitable performance group name. That is, we refer to Cluster 1 as *Unsatisfactory* performance group, Cluster 2 as *Good* performance group, and Cluster 3 as *Satisfactory* performance group. We refer to these clusters as *students' performance groups* or sometimes simply as *performance groups* in the rest of the paper.

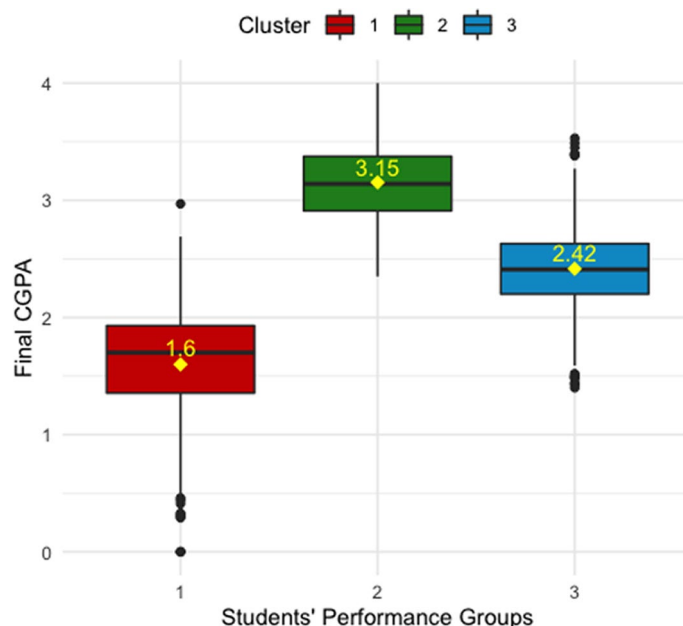


Fig. 4 Final CGPA boxplots for the three clusters

This distinction between clusters is quite logical since 2.00 is the least required CGPA by the university for a student to graduate. Thus, students in the unsatisfactory performance group (or Cluster 1) struggle to achieve the least required performance in order to graduate as well as to avoid the chance of termination of their admissions in the university.²

Classifying course categories as *hard* and *easy*

The second research question (RQ-2) focuses on analyzing the students’ performances in the nine course categories (given in Table 1) to determine if we can relate some course categories as *hard* and *easy* for the three performance groups (or clusters) of students identified in the previous step. In Fig. 5 we show boxplots of students’ performances³ in the nine course categories for each performance group.

The boxplot charts clearly show the difference in students’ performance with respect to distinct course categories. However, to quantitatively define a measure to categorize a course category as ‘hard’ or ‘easy’ for each performance group, we consider the least satisfactory grade point of 2.00 (which is equivalent to ‘C’ grade) as a threshold value. So, if the median grade point of students’ performance group in a specific course category is below 2.00 grade point, we refer to that as a ‘hard’ course category and an ‘easy’ course category otherwise. Table 5 shows the hard and easy courses for the three students’ performance groups.

² A student receives a *warning* from the university on obtaining CGPA below 2.00. The warning count increases by one on unsatisfactory performance in the subsequent semester. And, a student’s admission is canceled by the university on a warning count of three.

³ The equivalent numeric grade points of the original 12-level letter grades are used to plot the boxplots.

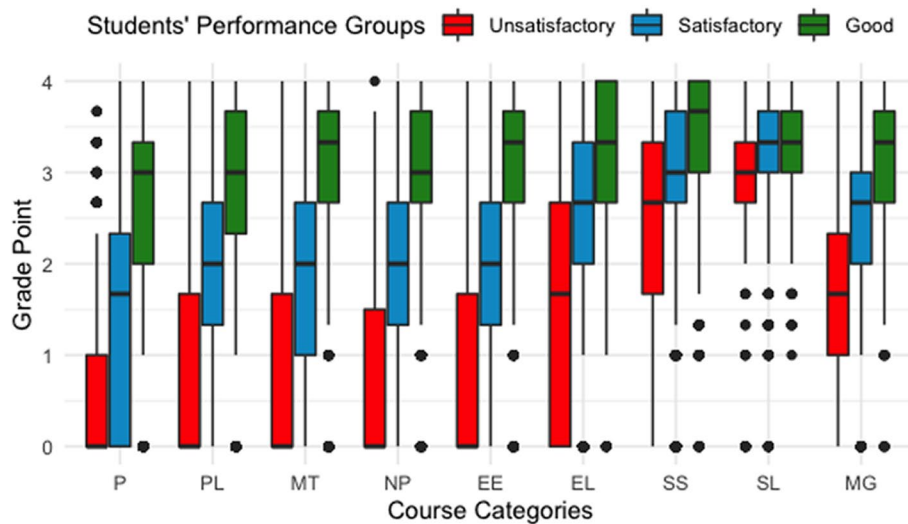


Fig. 5 Comparisons of the three students' performance groups in the nine course categories

Table 5 Categorization of the nine course categories as *hard* or *easy* for the three students' performance groups (or clusters)

Students' performance groups	Hard course categories	Easy course categories
Good	–	P, PL, MT, NP, EE, EL, SS, SL, MG
Satisfactory	P, PL, MT, NP, EE	EL, SS, SL, MG
Unsatisfactory	P, PL, MT, NP, EE, EL, MG	SS, SL

As expected, most of the course categories appear to be challenging for students in the *Unsatisfactory* performance group. That is, all the course categories except Social Science (SS and SL) are categorized as hard for this group (based on the criterion mentioned earlier). Whereas, students in the *Satisfactory* performance group found the following course categories as hard: all the CS-related courses (i.e., P, PL, and NP), Mathematics (MT) and Electrical Engineering (EE) courses⁴. Not to much surprise, no course category is found to be hard for *Good* performance students. Note that the SS and SL course categories are categorized as *easy* for all the performance groups. This shows that students in general perform well in Social Science courses and labs (SS and SL, respectively). This information can be useful for course advisors who can suggest a SS or SL elective course instead of an MG course to a struggling student to improve his/her CGPA.

Analyzing course trajectories of students

Next, we were interested in analyzing differences (if any) in the combination of courses taken in a semester by students from different performance groups (see RQ-3). As mentioned earlier, the university offers a very structured BSCS program which defines the

⁴ The median grade points for these course categories are on the boundary value of 2.00, thus, we consider them as hard for the *Satisfactory* performance group.

Table 6 Summary of students’ transcript data for the three performance groups

Students’ performance groups	No. of students	No. of course trajectories CT	Average no. of courses per semester
Good	391	2625	6.10
Satisfactory	788	4779	6.14
Unsatisfactory	219	831	6.05

combination of core and elective courses to be taken in a semester. In the text below, we refer to a sequence of courses taken in a semester as a “course trajectory” (CT). Course trajectories of *regular* students (i.e., who have never failed a course) would match the university’s proposed curriculum. However, course trajectories of *lagger* students who failed at least a course(s) will show deviation(s) from the proposed curriculum. And, analyzing course trajectories of students belonging to different performance groups may unveil some useful insights for stakeholders which include students, teachers, and student administration.

Table 6 shows a summary of the transcripts data for the three students’ performance groups. The average number, of course, trajectories per student for each group are as follows: *Unsatisfactory* = 3.8; *Satisfactory* = 6.1; *Good* = 6.7. Since the *Unsatisfactory* performance group includes data of students with CGPA below 2.00, some students were dropped-out early from the university; hence, the average number of CTs for this group is very less in comparison to the other two groups. Also, note that the average number of courses per semester were 6 for all the performance groups (see the last column in Table 6). It would have been a challenge to manage this (regular) workload of 6 courses for students belonging to the *Unsatisfactory* group. By consulting the university’s academic rules, we find that a student with CGPA below 2.00 should enroll in only five courses at a maximum in a semester. A detailed look at such cases revealed that there were violations of the academic guidelines in many cases, and, in some of the cases, actually enrolled courses were five but a mandatory lab course increases the count to six.

Figure 6 shows the top five frequent course trajectories of the three students’ performance groups. The relative frequency of each unique course combination is shown on the right. The accumulative sum of the shown course trajectories for the three groups is: (a) *Good* = 34.85%, (b) *Satisfactory* = 25.47%, (c) *Unsatisfactory* = 33.69%. In the following, we discuss some important differences in the course trajectories of the three performance groups.

The first difference that one can observe is that accumulative sum of the top five course trajectories of the *Satisfactory* performance group is very less as compared to the other two groups. The second frequent course trajectory in *Good* students’ performance group shows two Mathematics (MT) courses. By looking at the university’s course catalogs, we came to know that initially two MT courses (namely: Linear Algebra and Calculus-I) were offered to freshmen students until the year 2015. However, the university revised the curriculum later and started offering a single MT course per semester to balance the students’ workload which naturally increases with two MT courses. This particular course trajectory is also found in the *Satisfactory* students’ performance group in the second place. However, double MT courses appear in two course trajectories of

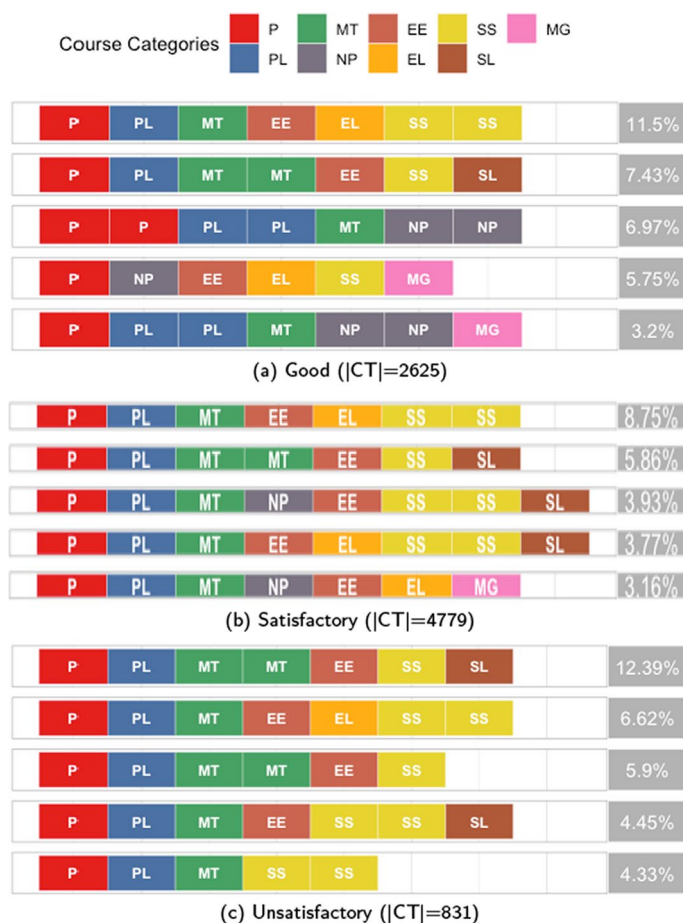


Fig. 6 Frequent course trajectories of students belonging to different performance groups, namely: **a** *Good*, **b** *Satisfactory*, and **c** *Unsatisfactory*. |CT| is the total number of course trajectories. The relative frequency of each unique course combination is shown on the right. The accumulative sum of all the frequencies equals 50%

the *Unsatisfactory* students’ performance group. We can expect that the two Mathematics courses in a semester were really challenging for poor-performance students. And, so they might have to repeat the same courses upon failure. But, we also wonder why those students were not advised to take only one MT course in a semester after repetitive failures.

We can also observe that the combination of two Social Science (SS) courses is quite persistent in the course trajectories of *Satisfactory* and *Unsatisfactory* students’ performance groups. As identified earlier, students with unsatisfactory performance are overburdened with the number of enrolled courses in a semester against the university’s academic guidelines. The count of courses may be reduced by eliminating an additional SS course from their course enrollment plans for a semester. Students belonging to *Good* performance group clearly have distinct course trajectories as compared to the other two groups. For example, non-programming (NP) courses are found in the last three trajectories, these CS courses usually start from the fourth semester (including Operating Systems and Databases). NP courses appear only twice in the *Satisfactory* group and never in the *Unsatisfactory* group. This shows that students from the last two groups

Table 7 Sample students’ semester performance data ordered temporally

Student	Semester Performance Data <S1, S2, S3, S4, ...>
St1	<good, good, good, average, good, good, good, good>
St2	<average, average, good, good, good>
St3	<good, average, average, average, good, average>
St4	<poor, good, poor, poor, average, good>
St5	<poor, poor, poor>
St6	<average, poor, poor, average, poor, poor, poor>

were majorly studying the first three semester courses. In view of these data-driven patterns, student administration can offer some courses in summer semester(s) particularly for *Unsatisfactory* performance group that may led them to complete fundamental courses early on and advance to other supporting and elective courses. Thus, visualization of complete course trajectories of students particularly those belonging to different performance groups can unveil some key differences and issues that may go unnoticed otherwise. And, by offering these data visualizations in academic dashboards, different stakeholders (i.e., department head and student advisor) can benefit to devise actionable strategies in a timely manner.

Transitioning of students between different performance groups

Another crucial aspect while investigating students’ transcripts data is to understand how likely a student can transition from one performance group to another, which was identified as the last research question of this research study (see RQ-4).

To answer this question, we labeled the students’ SGPA for each semester as ‘poor’ if the SGPA < 2.00, ‘average’ if the SGPA ≥ 2.00 and < 3.00, and, ‘good’ if the SGPA ≥ 3.00. Then, we temporally order each student’s registered semesters in an ascending order wherein each semester is represented by a label (poor, average, or good). A sample data is shown in Table 7.

Finally, we plot a *Markov* chain for the three students’ performance groups as shown in Fig. 7, to determine the student’s probable migration between different semesters performance labels subsequently. The three pre-mentioned SGPA performance labels (i.e., poor, average, and good) are represented as states of the *Markov* chains. Edges between different states are labeled with a transition probability that shows respectively the probability of students shifting or retaining a specific performance label in the next semester. The transition probabilities less than 10% are not shown for legibility. And, the thickness of the edges shows a high transition probability between the states. In the following, we interpret the three *Markov* chains to understand the performance transition behaviors of the students belonging to different performance groups. We primarily focused on transition probabilities higher than 50% only, so that the most probable behaviors can be understood clearly. The remaining edges and corresponding probabilities are however shown for elaboration of researchers who might perform a similar study in the future. Note that the cutoff value to interpret the resultant *Markov* chains is subjective to the size of the data and domain experts choice.

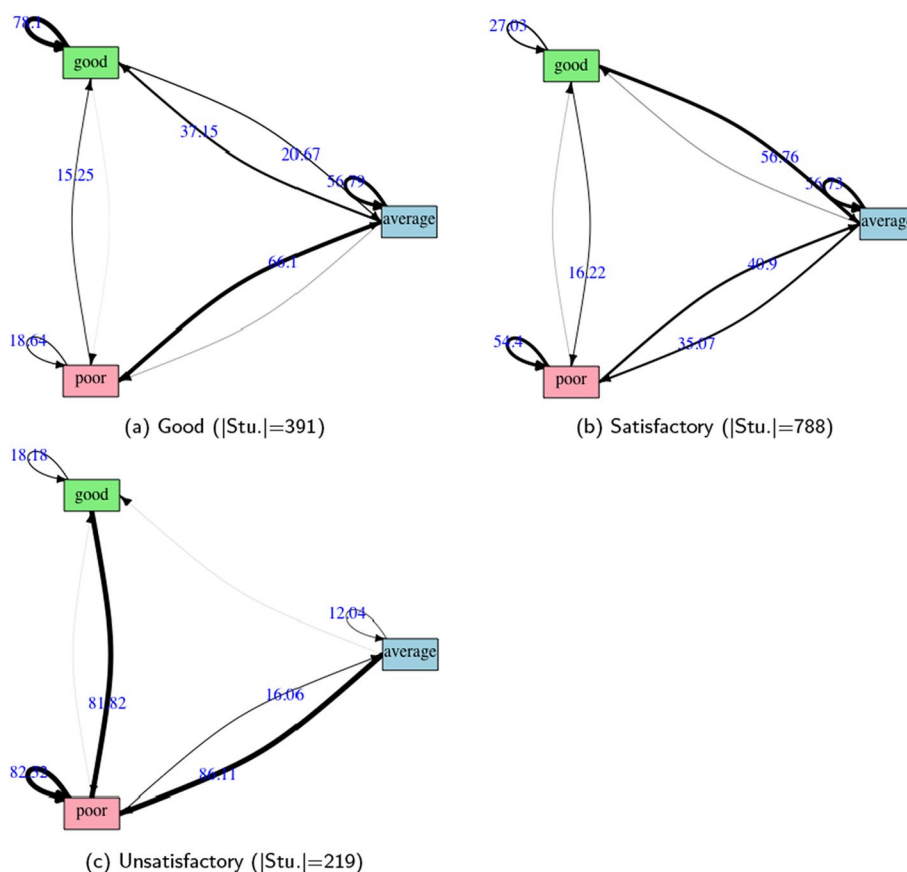


Fig. 7 Markov chains of the students belonging to different performance groups, namely: **a** *Good*, **b** *Satisfactory*, and, **c** *Unsatisfactory*. |Stu.| is the total number of students in a cluster. Transition probabilities below 10% are not shown for legibility

Figure 7a shows the *Markov* chain of *Good* performance group that comprises 391 students. The *good* state has a high probability (78.1%) self-loop which shows that students in this group are more likely to achieve high SGPA in the subsequent semesters. Students in this group who may have achieved poor SGPA at some point have shown a probability of 66.1% to achieve average SGPA later on. The average state has a 56.79% self-loop transition probability. Thus, students in this group have shown a potential to repeatedly achieve good and average SGPA, with a more likelihood of staying in the same state.

Figure 7b shows the *Markov* chain of *Satisfactory* performance group containing 788 students. Students in this group are more likely to retain their average or poor SGPA subsequently as shown by the self-loops on both states with above 50% transition probability. The only prominent migration behavior is depicted by the edge from good to average state with a transition probability of 56.76%. This shows that students who attained good SGPA initially, moved to average SGPA.

Figure 7c shows the *Markov* chain of *Unsatisfactory* performance group containing 219 students. Students in this group are more likely to retain their poor SGPA periodically with a very high probability of 82.32% (see the self-loop on the poor state). Also, students who may have achieved good or average SGPA at some point have more than

80% chance of moving to a poor SGPA state later. This kind of pattern is alarming for an educational institute where an absolute decline in the performance is observed, i.e., students who previously obtained good grades (between ~ 3.00 to 4.00) dropped their performance to a poor range (grade point below ~ 2.00). Many factors may be associated with such unfavorable observed patterns, for example, change in the course difficulty and/or course instructor. Change in a course difficulty is inevitable in a curricula, however, course advisors can pick a balanced combination of courses for such struggling students. Hence, further investigation is needed by student administration to determine the potential impact of change in faculty for a specific course category to improve students' learning outcomes.

In summary, these Markov chains have helped us to comprehend transitioning of the students between different performance labels based on their SGPA in the subsequent semesters. The semester grade point average or SGPA is already used in higher educational institutes (HEIs) to track a student's performance profile. However, HEIs are not certain which group of students are more likely to improve, retain, or decrease their performance over time. In this respect, the adoption of our method to utilize the Markov chains can provide useful insights for teachers, student advisors, and other stakeholders by offering visualization and a probability estimate based on the data. It is worth mentioning that the interpretation of the probabilities of transition cannot be immediately accepted as cause/effect relationships. For example, in Azzini et al. (2016) the authors observed that performance is more connected with students' quality than courses' trajectories. The identification of a trajectory connected to good performance may not necessarily imply forcing other students to follow the same trajectory to attain better performance. The attitude to follow a trajectory may be the effect of being good students and not the cause. Therefore, the transitions of the Markov chains should be interpreted as effective descriptors of the state of the art rather than predictors.

Conclusions and recommendations

Students have to study a combination of courses from different disciplines (e.g., Computer Science, Mathematics, Engineering, Social Science, etc.) during any degree program. The two quantitative measures to rank students' performances include a semester grade point average (SGPA), and, a cumulative grade point average (CGPA). However, neither of these measures can tell us which students have shown similar performances in certain course categories (RQ-1). Also, some course categories might be favorably easy or more difficult for a cohort of students which can be useful to identify their strengths and weaknesses (RQ-2). Furthermore, examining the combination of courses (which we referred to as course trajectories) of different groups of students can also be an interesting point of concern for the stakeholders (RQ-3). Lastly, understanding the migration behaviors of students between different performance groups based on their SGPA in the subsequent semesters can reveal useful insights for teachers and student advisors (RQ-4). These were some open research questions that are answered in this study using various descriptive data mining and visualization approaches. The dataset used in this case study comprises transcripts records of 1398 undergraduate Computer Science students of a private university in Pakistan. Although the university offers a very structured BSCS program, we have

identified some issues by examining students' transcript records grouped into nine distinct course categories. We expect that the findings of this research study not only help the NUCES, but the research questions and methodology devised in this work would also benefit the EDM community at large.

In a nutshell, our findings of RQ-1 show that similar groups of students can be found using a suitable clustering algorithm (considering data types and its nature, i.e., data sparsity, dimensionality, etc.). A post processing step should be performed to interpret the resultant clusters using some external criterion, for example, CGPA, SGPA, etc. These clusters or students' groups can be further utilized to identify patterns that might be associated with these groups. For example, we identified (in response to RQ-2) that students follow different course trajectories. Using the domain knowledge, we also devised a simple method for categorizing course categories as 'hard' or 'easy' to answer RQ-3. However, there are a number of caveats which concern the interpretation of the obtained results. As mentioned earlier, the identification of a trajectory connected to good performance students may not necessarily imply forcing other students to follow the same trajectory to attain better performance. Visual examination of course trajectories in fact can provide useful insights for student advisors and administration, who can mutually devise a suitable plan for different cohorts of students knowing which course categories are preferably suitable or more complex for a specific group of students. Finally, the Markov chain visualization and prominent migration patterns should be interpreted cautiously and not mistakenly taken as cause/effect relationships without further evidence. The focus should be on to devise strategies for future ramifications that might benefit students and academic institutes.

To conclude, only by investigating abundant data of students, educational institutes can identify evidence-based problems and devise a course of action to improve their academic processes and effectiveness. However, performing descriptive data analyses of educational data has its own challenges that include formulation of well-defined research questions which should be devised considering the needs of different stakeholders (e.g., student administration, teachers, and students). Another challenge is that educational institutes usually capture students' data in different formats and at different granularity levels (Romero & Ventura, 2013), thus data integration and data wrangling tasks probably need to be done repetitively for each research question by keeping in mind its objective. Additionally, we highlight that an accurate interpretation of the results in an educational setting is subject to some unknown variables including different faculty teaching the same course, changes in course syllabus, student's demographics, etc. Last but not the least, educational institutes need to realize the usefulness of data-centric methods which can offer valuable insights to them. We suggest that different data mining and visualizations methods should be integrated into dashboards used by academic institutes to extract useful information on a periodic basis, for example, after the completion of each semester or a major assessment activity (i.e., mid-term exams). We highlight that this work does not present answers to a comprehensive list of questions that might interests an academic institute. We however have shown in this case study that devising a list of meaningful open questions is the first step. To answer those questions, most appropriate data mining and/or visualization method along with domain knowledge can provide the right answers

to higher education institutes. We believe the findings of this work has laid a good basis for designing a learning analytic dashboard in the future.

Future work directions

A few research directions are discussed below which can help us to retrieve more meaningful data-driven insights to identify factors affecting students' performances. One of the most critical independent variable is faculty details, e.g., education level, years of experience, course-specific experience, etc. We suggest to investigate the potential impact of an individual faculty on students' learning outcomes in a future study. Another future research direction is to use a formal method for categorising course categories as 'easy' or 'hard', for example, item difficulty and item discrimination index (Karadag & Sahin, 2016). Lastly, we hope that devising and answering a new list of open questions about students learning behaviors and institutional effectiveness will follow this work.

Acknowledgements

The authors would like to sincerely NUCES, Pakistan, for providing the dataset. The authors also acknowledge the support of Rizwan Ul Haq and Munaza Akhter in course labeling. We also appreciate Muhammad Naeem and Meriam Chuhdary for their assistance in data understanding.

Author contributions

Conceptualization, RM; Data curation, RM and MSS; Formal analysis, RM, PC and MA; Funding acquisition, RM and MA; Investigation, RM, PC and MA; Methodology, RM and PC; Visualization, RM; Writing—original draft, RM and PC; Writing—review and editing, RM, PC, MA and MSS. All authors have read and agreed to the published version of the manuscript.

Funding

This research work was funded by the Office of Research, Innovation and Commercialization (ORIC) of the National University of Computer and Emerging Sciences, Pakistan, under the Faculty Research Support Grant No. 11-71-11/NU-R/21.

Availability of data and materials

The data used in this study cannot be made public due to restrictions of the authorized university.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 6 April 2023 Accepted: 26 September 2023

Published online: 16 October 2023

References

- Almatrafi, O., Johri, A., Rangwala, H., & Lester, J. (2016). Identifying course trajectories of high achieving engineering students through data analytics. In: 2016 ASEE Annual Conference & Exposition.
- Azzini, A., Ceravolo, P., Scarabottolo, N., & Damiani, E. (2016). On the predictive power of university curricula. In: 2016 IEEE Global Engineering Education Conference (EDUCON), pp. 929–932 . <https://doi.org/10.1109/EDUCON.2016.7474663>
- Baker, R. S., Yacef, K., et al. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.
- Bresfelean, V. P., Bresfelean, M., Ghisoiu, N., & Comes, C.-A. (2008). Determining students' academic failure profile founded on data mining methods. In: ITI 2008–30th International Conference on Information Technology Interfaces, pp. 317–322 . IEEE
- Darcan, O. N., & Badur, B. Y. (2012). Student profiling on academic performance using cluster analysis. *Journal of E-learning & Higher Education*, 2012, 1–8.
- Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991–16005.
- Ferral, H. (2005). *Clustering students by their subject choices in the learning curves project*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
- Karadag, N., Sahin, M. D., et al. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *Turkish Online Journal of Educational Technology-TOJET*, 15(4), 16–24.
- Priyambada, S. A., Mahendrawathi, E., & Yahya, B. N. (2017). Curriculum assessment of higher educational institution using aggregate profile clustering. *Procedia Computer Science*, 124, 264–273.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.

- Sya'iyah, K., Yuliansyah, H., & Arfiani, I. (2019). Clustering student data based on k-means algorithms. *Int J Sci Technol. Res*, 8(8), 1014–1018.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining.
- Tuyishimire, E., Mabuto, W., Gatabazi, P., & Bayisingize, S. (2022). Detecting learning patterns in tertiary education using k-means clustering. *Information*, 13(2), 94.
- Yoo, J.S., Woo, Y.-S., & Park, S.J. (2017). Mining course trajectories of successful and failure students: a case study. In: 2017 IEEE International Conference on Big Knowledge (ICBK), pp. 270–275 . IEEE

Publisher's Note

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

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
