Analysing correlation between sequential event in student's learning path

Sulfikar Sallu[1*], Tekad Matulatan[2], Muhammad Resha[3]

{sulfikar.sallu@usn.ac.id[1], tekad.matulatan@umrah.ac.id[2,] mresha@unitama.ac.id}
Universitas Sembilanbelas November, Kolaka[1], Universitas Maritim Raja Ali Haji, Tanjungpinang[2], Universitas Teknologi Akba Makassar, Makassar[3]

Abstract. The relationship between a course and other subjects described in the curriculum does not explain how much the relationship is supposed to influence the success of students' learning path. Also, some courses are not explicitly written as prerequisites to any other course in the curriculum, but still contribute significantly to the success of the subsequent course. One common way to find the correlation between these courses is by using the classical Pearson Test, but this might ignore the detail of the relationship. This can be improved by incorporating the PCA cross-matrix technique with an improved and more-detailed per-grade relation, not just a general view. This means that students who perform excellently in the previous course might also perform very well in a subsequent one. From the obtained results, it can be seen that the incorporation of the PCA cross matrix technique with a detailed description of the correlation's strength between the prerequisite course and the course it is prerequisite to, based on the results of the former, added more details to the classical Pearson test.

Keywords: Learning Analytics, Educational Data Mining, Correlation Analysis.

1 Introduction

The field of learning analytics is focused on processing the data from students' study results or behavior in order to enhance their study performance and to help improve skills or knowledge, which will be reflected in the students' future grade results. The research sample data were obtained from two informatics courses at Universitas Maritim Raja Ali Haji in the 2019–2021 academic year. This research does not address the complete analytical learning process, rather, it examines a small subset of problems that measure the strength of courses correlation. Following this, various research related to the field of learning analytics were reviewed, but none of them were focused on the time-series relationship that is based on students' grades [1],. Majority of the research analyzed the learning process by employing sample techniques, such as brief tests, exam result analysis, or questionnaires. Some research made use of dynamic time sampling and clustering to search for correlations between the courses [2]. Furthermore, in some learning analytics-related research, teaching techniques such as constructivist learning, were proposed to aid university students in the acquisition of generic skills [3], while others analyzed factors that help students navigate digital learning activities using LMS (*Learning Management System*) tools [4] or CMLS [5]. There are other computational approaches related to learning analytics as noted in [6][7], they may be curriculum-related [8][9], centered on

ICSEDTI 2022, October 11-13, Tanjungpinang, Indonesia

DOI 10.4108/eai.11-10-2022.2326327

competency analysis [10][11], involve the creation of an adaptive tutoring system [12], or may require sorting out the data of students having difficulties in their studies [13]. Another area specifically searches for causal relationships between temporal occurrences using time series as data streams [14][15][16]. The latter also uses time series but is less concerned with learning analytics.

2 Assumption and Boundaries

Situations and conditions where the data used in this research is believed to greatly affect the value in the data, so it is necessary to set limits and assumptions on several things in order to get a common persepective. This condition will obviously change when using data from other places.

These are the assumption used:

- All grades obtained in the course are fair and objective
- Grading scores range has little affect on quality
- Students who take courses do not have social, psychological or health problems that can affect their learning performance.
- Parallel classes with different lecturers have the same quality of assessment.

The boundaries for this research are:

- The higher education system where this research took place is the system that implemented in Republic of Indonesia
- Quality of grade are used despite of the range of grading value changes or differs.
- The curriculum that becomes the reference is the standard curriculum and does not take into account local subjects.
- This study uses the results of each course that has been carried out.
- The approach offers in this paper is not about acuracy but instead giving more details on facts that should be explored more.

Another things that related to data, is the data also been pre-processed by cleaning any unwanted data that can interfere with the result, for example withdraw courses and repeated courses for failed student.

3 The process

The experimental method used in developing the proposed solution as well as the overview or general concept of the processes is explained below. The detailed explanation of the implementation process is explained in the discussions section. The aim of this research is to develop a mechanism that is capable of analyzing the input data, which comprises of the grades obtained from previous courses, and making accurate predictions based on the input.

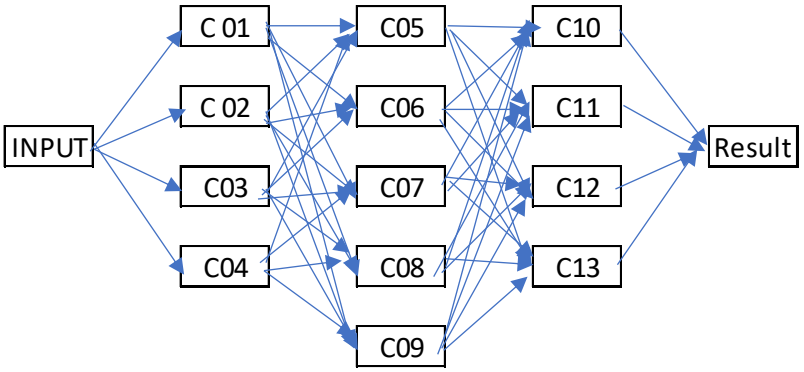


Fig 1. Finding course that rely on other course

Figure 1 shows that the implementation of the proposed method in an artificial neural network model. The students' grade result from the already-offered prerequisite course was used as the input data, after which the network will generates the best scenario based on that input. The process of identifying thllese courses is explained in the next section.

We also implement Pearson correlation test to quickly sort out which courses that are closely depend on other course. The Pearson formula is describe as follow:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

Where r is the coefficient

x_i is the ith of previous course variables

$\bar{x}$ is the grading value central tendency of previous course

y_i is the ith of the following course variables

$\bar{y}$ is the grading value central tendency of the following course

For Principal Component Analysis (PCA), we will not describe the detail process with equation but only algorithm, since it will be consume more pages.

Step 1: Mean Centering/ Normalize data. ...

Step 2: Compute the covariance matrix. ...

Step 3: Compute eigen vectors of the covariance matrix. ...

Step 4: Compute the explained variance and select N components. ...

Step 5: Transform Data using eigen vectors

Step 6: Reconstruct to the original data

4 Correlation with Pearson

Before we discuss further, lets have a look on data. The data show in table 1 is the recapitulation on course A01 dan course A07 which are pair courses where A01 precede

A07. And then we look at another pair which is not consifer independent (not correlated)

Table 1. Summarization on quality of grade on course A01 (left table) and course A07 (right table).

Grade	Freq
A	10
A-	15
B+	19
B	22
B-	25
C+	24
C	21
D	4
	140

Grade	Freq
A	8
A-	12
B+	15
B	22
B-	24
C+	22
C	22
D	1
	126

The number of students who took the course is not the same where the A01 course has bigger number than the A07 course. This indicate that students who failed in previous course could not take the following course. We could also find where the number of the following course is bigger the previous course, which cause by repeated student who failed in the following course. As mentioned before any repeated courses would be removed from the data.

Based on the data show in the table (Table 1), we could put on line graph, to display the trend happened on both courses. This pattern could describe what is happened in relation on both course that could signal if there is same affect on both side (Fig. 2)

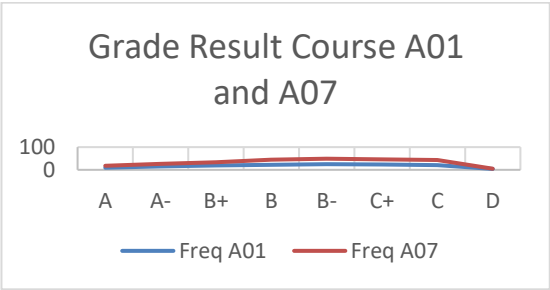


Fig 2. Distribution of grading on both courses

From the Fig.2 we can not assume that whoever performs well in A01 also performs well in A07, because it might show different student in that case.

Using the formula we calculate the Pearson coefficient using the data describe in table 1, and the result is 0.984 which indicate strong positive correlation. Again, keep in mind that this result using the summarization data as the general fact and not discrete based on per-student.

We took 2 courses as sample that runs on subsequent semester, we name it course

A01 and course A07, where course A01 and course A05 are independent of each other, in other words none of these courses are prerequisite.The data at Table 2, are the total the number of students who receive specific grade on both courses.

Table 2 Grading result on both course A01 and course A05

Grade	Freq A01	Freq A05
A	10	26
A-	15	21
B+	19	28
B	22	17
B-	25	4
C+	24	0
C	21	4
D	4	0

If we draw a line graph (Fig. 3) just like we do on previous section, we could find that the line is not very similar to each other accept for the lowest grade.

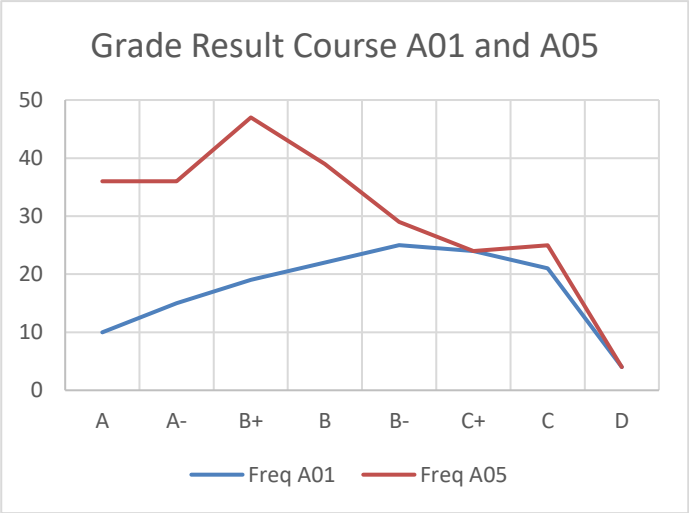


Fig 3. Distribution of grading between course A01 (previous course) and course A05 (following course) showing unsimilar pattern

Using the pearson correlation test we got the coefficient values is -0.14, which is indicates that both courses has weak negative correlation. Now we should try to look at one on one cases to find out if we could confirm the finding.

5 Correlation with Cross Matrix PCA

Cross matrix is the matrix that populate based on per-student grade result on both pair courses, as showed in Table 3 for first case A01 and A07.

Table 3. The Cross Matrix from previous course A01 to subsequent course A07

Grade A01	t A	t A-	t B+	t B	t B-	t C+	t C	t D
s A	4	2	2	2				
s A-	1	4	1	6	1	2		
s B+	2	5	4	1	2	2	3	
s B	1	1	5	9	3	3		
s B-			3	3	9	3	7	
s C+				1	6	7	10	
s C					3	5	2	1
s D								

On columns we see the target grade (in this case A07) range from A to D, and on rows we are using the source grade from A01 course. We removed the failed, E grade, because the grade means can not continue, so it did not make sense failed student could proceed to the next pair, unless it is not a pair (not correlated). The then populate number of student who receive grade in source course (rows) and grade in target course (columns). For example number of student who receive A grade in course A01 and also get A grade in course A07 is 4, get B grade is 2.

Using the data on the matrix above, we then calculate the PCA to find which component has more significant

Table 4. PCA on grading distribution from previous course A01 and subsequent course A07

	principal component 1	principal component 2	Grade A01
0	-2.377478	0.764995	s A
1	-1.434190	0.007498	s A-
2	-1.521272	-0.515815	s B+
3	-1.111236	-1.858600	s B
4	1.739296	-1.697674	s B-
5	3.031714	-0.551813	s C+
6	1.991303	2.091608	s C
7	-0.318135	1.759801	s D

On second case (course A01 and A05), we consolidate on cross matrix (Table 5), we could see the data is spread mostly on left side, with central tendency on B+,

Table 5 One on One Grading between unrelated course A01 and A05

Grade A01	t A	t A-	t B+	t B	t B-	t C+	t C	t D
s A	10							
s A-	5	6	4					
s B+	2	6	9	2				
s B	3	3	7	9				
s B-	2	2	4	2	2			
s C+	2	2	1	2	2		3	
s C	2	2	3	2			1	
s D								

Using the data describe on Table 5, we calculate the PCA values and the result as follow.

Table 6. PCA analysis on grading distribution result A01 and A05

	principal component 1	principal component 2	Grade A01
0	-1.456117	2.475906	s A
1	0.921101	0.900143	s A-
2	2.293931	-0.216885	s B+
3	2.048628	-0.497065	s B
4	-0.382192	-0.902333	s B-
5	-2.050718	-2.234170	s C+
6	-0.223622	-0.095503	s C
7	-1.151011	0.569907	s D

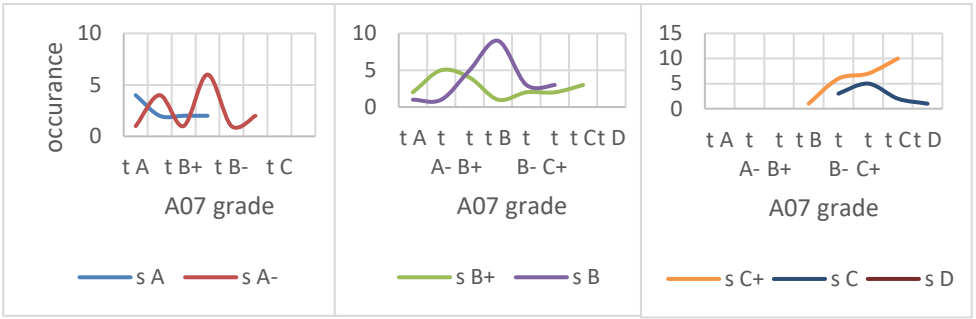
6 Discussion

For the first scenario (courses A01 and A07), the data trend used to calculate the Pearson coefficient, as shown in Figure 1, does not represent the actual situation on both courses, instead, it only represents the general trends. Furthermore, this general trend did not describe the track records of each individual with respect to their performance on the courses. Using Pearson coefficient for a fast glance to discover which courses are causally related may provide false indications.

The PCA result in Table 3 shows that A- and B+ are close to the origin and have a more significant positive impact compared to others grades. Meanwhile, lower grades

such as D and E resulted in fewer improvements.

The cross matrix above can also be plotted as a line chart to show the relationship between the courses. From Figure 3, it can be seen that students whose grades were A or A- in course A01 will most likely perform well in A07. Also, a similar trend was observed with students whose grades were B+ or B in the course. From the third graph (c), it can be seen that students whose grades are either C+ or C are most likely to perform on average. This shows that the grade achieved by students in course A01 is predicted to have an effect on their performance in course A07, i.e the correlation between these two courses is strong when a student's grade in A01 is above C+, otherwise, the correlation is weak. The second case (course A01 and A05) in the Fig. 2, we could see there is no similar pattern but with same trend on the end while the Pearson test result shows not correlated among two courses. We then will do the same technique as previous, which we draw based on grades (Fig. 4).



(a) (b)

Fig 4. The cross matrix grading distribution on course A07 where previous course A01 (a) student who receive grade A and A (b) grade B+ and B, (c) grade C+ and C

Plotting the PCA result in scatter graph (Fig. 5) showing the disperse situation on each grade that gives how much or how strong it will enforce the effect on the next sequence. Plot that stray away from origin could be assume that it might has a little force or no significant on the result.

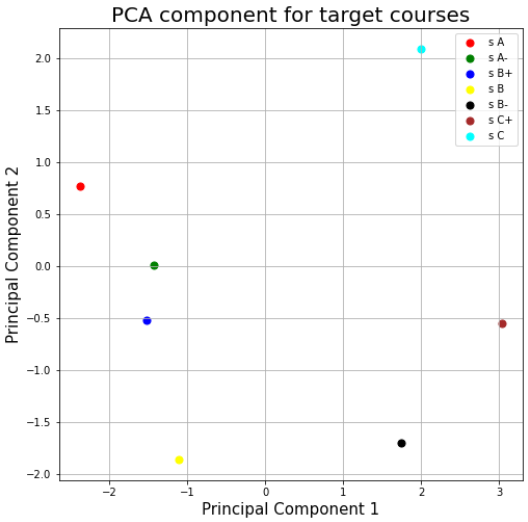


Fig 5. Scatter graph on PCA result

For case course A01 and A05, comparing each grade with each other as in Fig 6, display very little similar pattern on some grades.

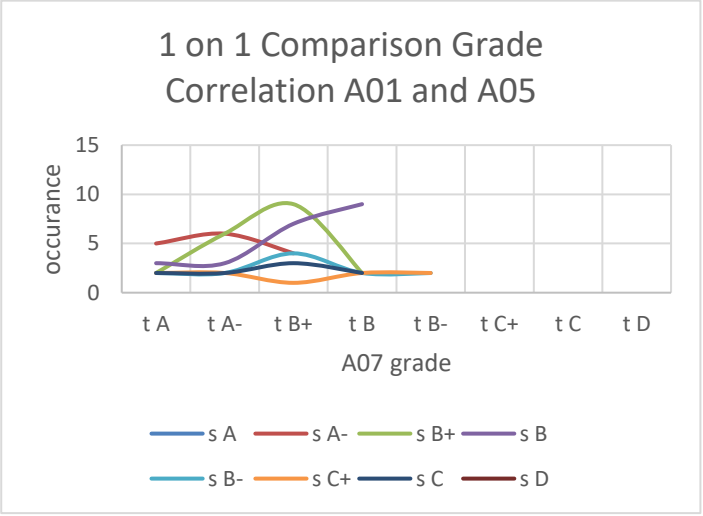


Fig 6. The grading distribution on A05 based on A01 result

Doing the similar step, by plotting the PCA result (Table 6) on relation between A01 and A07 (Fig. 7),

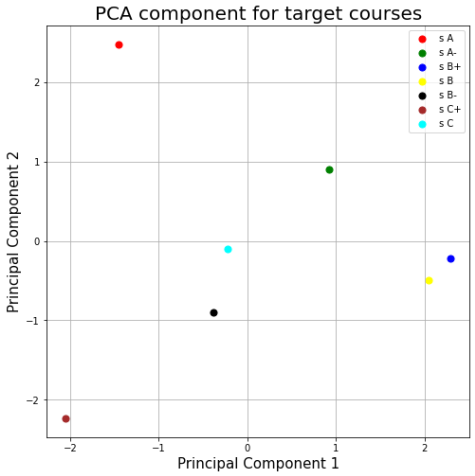


Fig. 7. PCA plotting based on grading distribution on A01 and A05

With two cases, now lets look the difference between using Pearson Correlation Test and with adding Cross Matrices PCA (Table 7) shows that using more information on cross matrices PCA, we could see which grade have more affects on students performance with the subsequent courses (prerequisite or not).

For pair courses A01 and A07, student with grade A- in A01 will have more chance to have A- or better in A07 compare with C+ on A01, that's also true for student who get D on A01 will likely to perform bad in A07. For general courses A01 and A05 (which has no related at all), student with C will likely also perform bad in A05 (which in this case, students could be not performing good in all other courses in that semester).

Table 7. Comparison using Pearson Test and with additional Cross Matrices PCA

s-t		Pearson Coeficient	distance Cross Matrices PCA
A01-A07	s A	0.984	2.49
	s A-		**1.43**
	s B+		1.60
	s B		2.16
	s B-		2.42
	s C+		3.08
	s C		2.89
	s D		1.78
A01-A05	s A	-0.14	2.86
	s A-		1.29
	s B+		2.30
	s B		2.10
	s B-		0.98
	s C+		3.03
	s C		**0.24**
	s D		1.28

7 Conclusion

In conclusion, using cross-matrix PCA can improve detection when examining individual cases rather than general cases. This gives a deeper insight into the interaction process among present courses based on the grades from previous ones. On the other hand, the process took longer time to complete due to the one-on-one population cross-process,

which significantly slows down the whole process when implemented on an artificial neural network. However, as more details are incorporated into the process, it is expected that the accuracy of the system will improve, regardless of the method that will be used.

References

[1] Matulatan, T., Bettiza, M., Rathomi, M.R., Ritha, N., Hayaty, N.: Predictive Adaptive Test with Selective Weighted Bayesian Through Questions and Answers Patterns to Measure Student Competency Levels, Jurnal Teknologi dan Sistem Komputer, 7(2), Universitas Diponegoro, Indonesia (2019)

[2] Srishti M, Zohair S and Santanu P.:TIME SERIES EVENT CORRELATION, PeerJ Preprints, https://doi.org/10.7287/peerj.preprints.27959v1

[3] Virtanen A.,Tynjälä P.,: Factors explaining the learning of generic skills: a study of university students' experiences, Teach. High. Educ. 24(7) pp. 880–894. (2019)

[4] Krumm A., Everson H T.,Neisler J.: A Partnership-Based Approach to Operationalizing Learning Behaviours Using Event Data,. Journal of Learning Analytics, 9(2)pp. 24-37 (2022).

[5] McEneaney J ., Morsink,P. : Curriculum Modelling and Learner Simulation as a Tool in Curriculum (Re)Design. Journal of Learning Analytics, 9(2) pp 161-178 (2022).

[6] Buckingham, S. S., Deakin, R. C. : Learning Analytics for 21st Century Competencies, J. Learn. Analytic. 3(2) pp. 6–21 (2016)

[7] Mangaroska K., Giannakos, M. : Learning Analytics for Learning Design: A Systematic Literature Review of Analytics-Driven Design to Enhance Learning, IEEE Trans. Learn. Technol. 12(4) pp. 516–534 (2019)

[8] Brown, M., DeMonbrun, R M.,Teasley S. :Taken together: Conceptualizing students' concurrent course enrollment across the post-secondary curriculum using temporal analytics. Journal of Learning Analytics, 5(3), pp. 60–72 (2018)

[9] Hilliger I., Aguirre C., Miranda C., Celis S., Pérez-Sanagustín, M.: Design of a curriculum analytics tool to support continuous improvement processes in higher education. Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), 23–27 March 2020, Frankfurt, Germany , pp. 181–186. ACM Press (2020).

[10] Gottipati, S., Shankararaman, V.:. Competency analytics tool: Analyzing curriculum using course competencies. Education and Information Technologies, 23(1), pp. 41–60 (2017)

[11] Chou C-Y, Tseng S-F, Chih W-C, Chen Z-H, Chao P-Y, Lai K R, Chan C-L, Yu L-C. , Lin Y-L: Open student models of core competencies at the curriculum level: Using learning analytics for student reflection. IEEE Transactions on Emerging Topics in Computing, 5(1), 32–44, (2017)

[12] Gonzalez-Brenes, J. P., & Huang, Y.:. Using data from real and simulated learners to evaluate adaptive tutoring systems. In J. Boticario & K. Muldner (Eds.), Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education (AIED 2015), 22–26 June 2015, Madrid, Spain (**5**), pp. 31–34 (2015)

[13] Hershkovitz A., Ambrose A:. Insights of Instructors and Advisors into an Early Prediction Model for Non-Thriving Students. Journal of Learning Analytics, 9(2), 202-217 (2022)

[14] Box, G. E. , Jenkins, G. M., Reinsel, G. C. , Ljung, G. M.: Time series analysis: forecasting and control. John Wiley & Sons, (2015).

[15] Chatfield, C. : The analysis of time series: an introduction. CRC press, (2016).

[16] Guralnik, V., Srivastava,J. :Event detection from time series data," in Proceedings of the fifth ACM IGKDD international conference on Knowledge discovery and data mining., pp. 33–42. ACM, (1999)