# Effect of Students' Activities on Academic Performance Using Clustering Evolution Analysis

Djoni Haryadi Setiabudi[1*] and Michael Santoso[2]
[1−2]Informatics Department, Petra Christian University
Surabaya 60236, Indonesia
Email: [1]djonihs@petra.ac.id, [2]santosomichael94@gmail.com

***Abstract*—Educational data mining is a technique to evaluate educational process of university students, especially in their early stages. Most preliminary studies focus on observing courses undertaken by students from one semester to the next to predict their success rate. However, besides studying, many students are also involved in non-academic activities, which tends to affect their grades. Therefore, the research aims to determine the effect of student activities on grades while taking into account their academic activities. The method used for clustering is K-Means. Data are collected by observing students' activity patterns in lectures. The research is conducted in two study programs at Petra Christian University: Business Management and Architecture. The results show that the K-Means method gives good results. The clusters formed from the data show non-homogenous groups and produce insights from several groups. The results show a tendency for students' performance to increase along with the number of activities and points earned. Most students have increased activities during busy times in the third, fourth, fifth, and sixth semesters. The peak is between the fifth and sixth semesters. Then, it starts to decrease in the seventh and eighth semesters. Therefore, students' activities in the Business Management study program affect performance significantly. Meanwhile, in the Architecture study program, it has an insignificant effect on performance.**

***Index Terms*—Students' Activities, Academic Performance, Clustering Evolution Analysis**

## I. Introduction

UNIVERSITIES or tertiary institutions culturally differ from both primary and secondary academic levels. It is mainly distinguished by the rules and freedom associated with learning patterns [1]. At the primary and secondary levels, students are required to comply with various existing regulations and standards. These regulations and standards govern their dress, appearance, and lessons to be taken. When students enroll in the university, they are free to engage in diverse activities previously prohibited. This freedom has positive and negative impacts, although they are trained to manage their time effectively and independently. It also involves the selection of activities and associations that certainly affect their study outcomes. Meanwhile, the preference for numerous non-academic activities and poor time management can affect the students' scores in that semester [2].

Non-academic activities are also important in universities because they teach students to build relationships with other people and develop various soft and hard skills. Although the tertiary institution mainly focuses on their graduation and capability to apply the acquired abilities in the world of work competently. Several students are complacent with non-academic activities, affecting their scores. It negatively impacts their confidence level in the subsequent semester, especially those who are unaware and continue to participate in many activities.

However, some students have a better adaptation rate. Therefore, engaging in more non-academic activities makes them properly manage their time and set priorities. It does not affect their scores which will even improve next semester. Although the activities differ among universities, proper analysis is needed to understand the balance between non-academic and academic activities of students. It ensures an increased success rate in studies and helps universities to produce more quality students with non-academic skills.

The previous studies conducted [3–10] focus on determining the academic performance of undergraduates. The majority of these studies state that the observation process can be assisted using data mining [3–7, 9, 11–15], and clustering [3, 5, 11, 12, 14] techniques. Some studies state that the data used to evaluate performance are the scores realized from the

Received: Oct. 08, 2022; received in revised form: March 6, 2023; accepted: April 10, 2023; available online: Sept. 08, 2023.
*Corresponding Author

courses offered [3–9, 15, 16]. These analyses also focus on the initial study [3–6, 17] period. The progress is realized from one cluster to another during the study period [3, 13], and its effect on the probability of dropping out is observed further [4, 6, 10, 15, 17, 18].

Several studies have been carried out on data mining in education. However, it is better known as educational data mining. The discussions and analysis involved utilize large datasets in the field of education to predict the students' success or failure, especially in the early stages of their undergraduate studies. Generally, the focus of attention is the courses offered per semester. The clustering approach produces a pattern that can predict the students' success or failure. Stakeholders also use it to make several decisions about implementing policies that can improve the university's standards.

Previous research studies the changes in learning patterns from studying at home in senior high school and the university level [1]. It examines the psychological stresses students face and actions that can support their mental health and well-being during the transition to early-level undergraduate studies. However, it fails to investigate student activities' impact on yearly performances. Moreover, the reason why undergraduates usually drop out of university is investigated [17]. It considers variables related to academic and socioeconomic factors influencing dropouts. The results show that students with the highest risk of dropping out are those in vulnerable situations, such as low grades, and those enrolled in a leveling course for an engineering degree. The university authorities can use this model to identify possible dropout cases. Next, the effect of time management on the achievements of engineering students is examined [2]. It discusses self-reported time management behavior by these undergraduates using the Time Management Behavior Scale. However, it fails to link the students' success with non-academic activities and its effect on their Grade Point Average (GPA).

Next, previous research observes the changes associated with the courses offered from one semester to the next while studying at the university [3]. This process is realized by adopting a modified approach called profile-based cluster evolution analysis, although it fails to consider student activities' impact on learning patterns. Another previous research studies educational data mining to predict the number of students who drop out due to their engagement in several extracurricular activities, considering that certain events, such as sports, can affect academic performance [4]. It has proven that extracurricular activities are an accurate predictor of dropouts. However, it only discusses the effect of non-academic activities on dropout predictions in certain subjects. It does not examine students' scores from one semester to the next.

Prior research also examines the application of data mining in the field of education alongside cluster analysis and three decision methods for e-learning [11]. The results show that the respective lecturers can identify e-course contents that require more attention and encourage the students to use them while preparing for exams. On the other hand, another prior research publishes a review on educational data mining [5]. Data mining techniques are used to predict students' outcomes according to their academic performances, especially in the early stages of university education. Meanwhile, another research adopts data mining and decision tree methods to analyze a dropout case in the Electrical study program by observing students during the first semester [6]. Interestingly, useful and relevant knowledge aids in predicting the students' inability or failure to continue their studies after the initial semester exams.

Next, a graph-based process discovery algorithm is developed [16]. It can reveal the sequence of learning activities engaged by several students. However, the result has not been associated with their success rate in pursuing higher education. Additionally, previous research has been conducted on data mining techniques in e-learning platforms, such as Moodle, WebCT, Claroline, and virtual learning systems [12]. The results of the student groupings are presented in tabular and visual representation forms, making those easier to read because Moodle does not provide special visualization tools. However, it does not provide good results regarding the student's performances and the possibility of dropping out.

The characteristic of the evolution cluster that is only applied for a certain period by ignoring the previous one is analyzed [13]. The results prove that the developed model can identify recurring clusters at various points and detect patterns to predict possible loss of value. It also indicate stability and preservation of value over time. Next, logs from students' learning processes in the form of modules and working quizzes are traced to produce menus and analyze the academic procedure [19]. The log analysis is carried out automatically. However, manual evaluation needs to be performed with Moodle Learning Management System (LMS) to ensure the learning outcome is unique. Moreover, students' progress and performance in Malaysia are analyzed and monitored, showing that existing methods were unsuitable [7]. The evaluated methods are Decision Tree, Neural Network, Naïve Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The Neural Network and Decision Tree techniques produce the highest accuracies.

Previous research performs a systematic literature review on educational data mining methods [14]. The procedure for preprocessing is clustering, a suitable technique for answering the research question to obtain valid results. Moreover, another research analyzes dropouts using an ensemble approach because it has not received much attention [15]. It proposes the use of an ensemble using Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict the number of undergraduates who drop out of the university.

Non-academic factors are also studied to predict the qualified prospective software engineering program trainees in Ukraine [20]. The factors affecting academic performance include degrees, age, work experience, and course additions. Moreover, an ensemble approach based on a heuristic system is utilized to develop a prediction paradigm to predict student performance [8]. In addition to employing boosting, it uses basic classifications, including Decision Tree, J48, KNN, Naïve Bayes, and Filtering Oversampling (SMOTE). As a result, the ensemble and screening approach has shown a substantial improvement in predicting student performance compared to the application of conventional classifiers. Similarly, previous research carries out a predictive analysis to determine the extent of the fifth and final-year achievements at Nigerian Universities [9]. The result shows that the study program, year of entry, and Grade Point Average (GPA) for the first three years are used to input the mining data based on the Konstanz Information Miner (KNIME) model. Meanwhile, six data mining algorithms are evaluated, and the maximum accuracy of 89.15% is obtained. The results are further verified using linear and quadratic regression models, thereby creating an opportunity to identify students who are about to graduate with poor GPAs or may not pass at all. There is a need for early intervention before it is too late.

Moreover, a predictive information system is used to prevent students from dropping out of university [10]. The developed model calculates the dropout risk per student by providing a warning procedure for the coordination of early preventive measures adjusted to the risk level. Similarly, previous research designs uplift modeling in the form of predictive analysis to encourage students individually [18]. However, the purpose is to reduce dropout rates among undergraduates. It is better than the conventional model. Another prior research also identifies the profile of psychological disorders on students' dropout rates at a personal level [21]. The factors observed are gender, educational path, one normative asymptomatic profile, and three symptomatic profiles, such as externalization, internalization, and comorbidities. The research shows that one in five students experiences symptoms of a psychological illness. The student is exposed to a greater risk of dropping out compared to their peers.

Based on the ideas from other preliminary studies that only focus on values such as [3], the research analyzes it from a different perspective. The student's performance is also influenced by how much they actively participate in non-academic activities [4]. The majority are not only interested in the academic aspect but also participate in extracurricular activities on campus to socialize with friends. Others include assignments from study programs that can be directly related to lectures, such as math or physics competitions, and building friendships with sports, hiking, and others. These undergraduates spend significant time managing these activities whenever they are on duty as the organizing committee. Hence, it affects their studies and prevents them from working on assignments given during lectures. Furthermore, some students are slow in learning, and these activities consume a lot of their study time. On the other hand, some are quick learners, and these activities motivate and encourage them to perform exceptionally in their studies.

The research aims to analyze the data acquired from students' grades and non-academic activities. Data mining and clustering are carried out using previously acquired information to determine the existing pattern in an educational institution. Hopefully, this method enables universities to strike a balance between activities and study and helps them to understand their students more. The research makes certain contributions, such as observing the pattern of students' activities from one semester to the next and its effect on their performances. It also includes identifying the migration patterns from one cluster to the next over time. This topic is rare because most research focuses on the courses offered and graduation. Meanwhile, the research observes the impact of non-academic activities on the student's studies.

## II. Research Method

The research is carried out at Petra Christian University. Utilized data are from the two study programs: Architecture and Business Management of the 2012 classes. These two study programs are selected based on their student population and high participation in activities at Petra Christian University. The tested samples include student performance and activity data from the Academic Administration Bureau and Student and Alumni Administration Bureau.

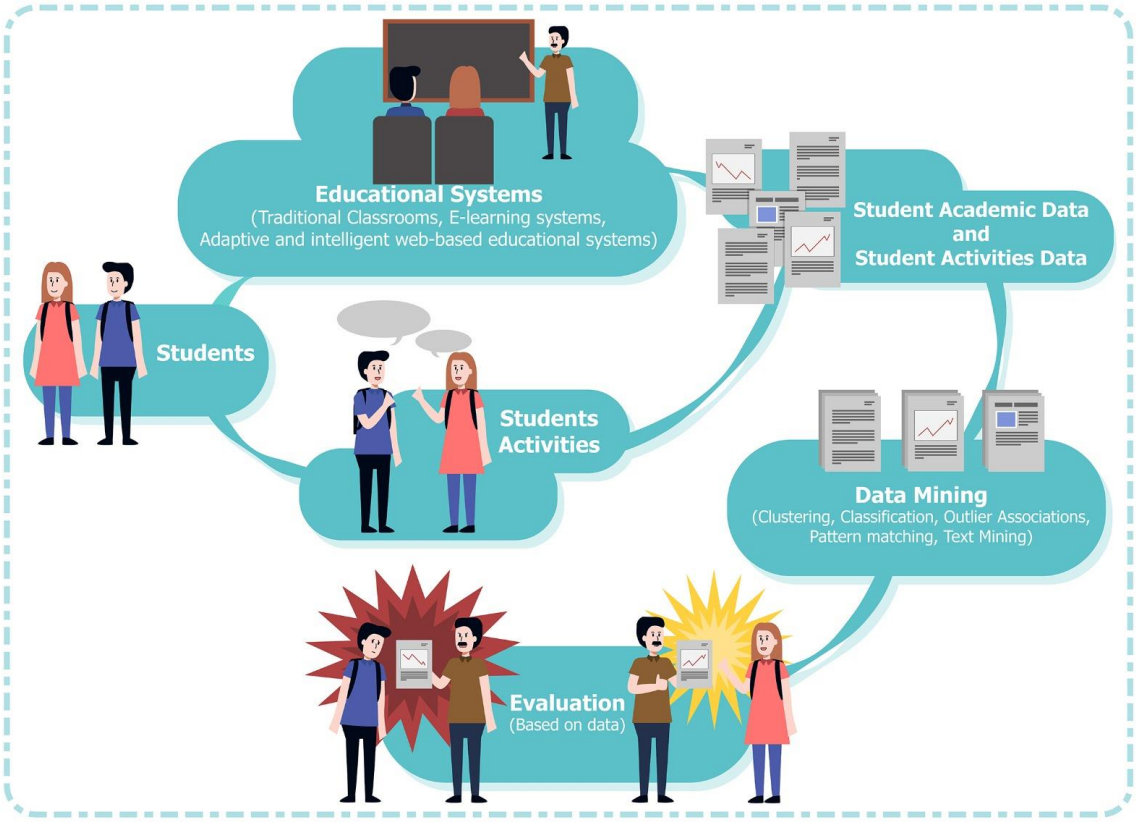


Fig. 1. The evaluation concept for the effect of student activities on academic performance with data mining.

The planned concept is shown in Fig. 1. Based on the diagram, it is evident that students participate in academic and non-academic activities during the study period. The information acquired from the scores and diverse activities is processed using a data mining technique to evaluate its effect on students' performance.

The data mining methods widely used tend to vary, such as Naïve Bayes, SVM, Decision Tree, K-Means, and KNN [3, 5, 6, 8, 9, 11, 15]. Some research also specializes in comparing these methods [7, 9]. Varying data sets are also used, such as those related to scores awarded in the early years and complex ones acquired from various fields of study [16]. The dataset used is the scores awarded to the academic and extracurricular activities of students from the Business Management and Architecture study programs in 2012. Batch 2012 is selected to examine the full study length before the COVID-19 pandemic. The criterion used is that students must have graduated by the end of 2019 (just a few months before the pandemic hit the country). The length of study at Petra Christian University is 3.5 years or 7 semesters, and the maximum is 7 years or 14 semesters. Generally, batches studied during the COVID-19 pandemic have very different activity behaviors than before because of the massive online adoption for non-academic and academic activities.

The clustering method selected is K-Means, considering that previous research has proven its success [3]. The K-Means method is used because the cluster is unknown with no labeling process. Hence, it cannot be predicted at the beginning. Objects are divided into groups that have something in common which is determined by using Euclidean Distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

In equation, $d(x,y)$ represents the Euclidean distance between data points $x$ and $y$. The sum of the squared difference between the corresponding components of $x$ and $y$ illustrates the geometric distance between these two points in a multi-dimensional space. It measures the length of the straight line connecting the two points, considering each dimension as an axis. Smaller Euclidean distances indicate a higher degree of similarity between data points, whereas larger distances signify greater dissimilarity.

Furthermore, data points are allocated to clusters

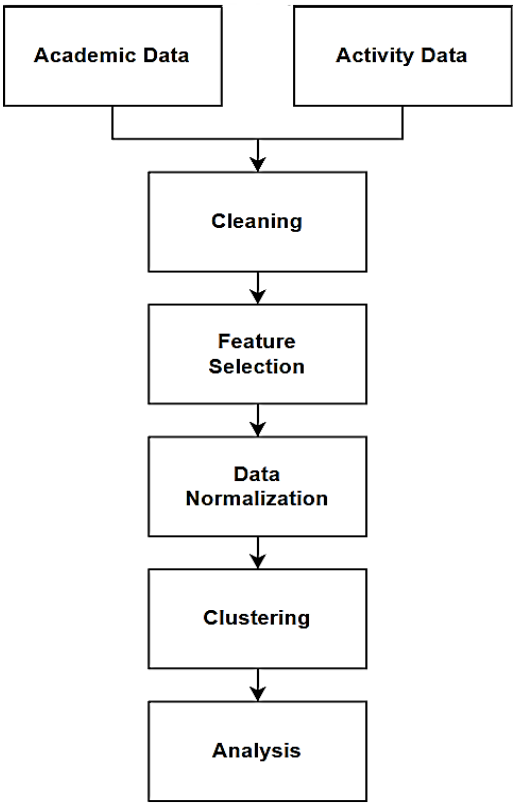


Fig. 2. K-Means algorithm for clustering evaluation analysis.

where the sum of the squared distance between the data points and the cluster centroids is minimize. The less variation within clusters is, the higher the homogeneity among the data points within the same cluster. The observations $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ consist of $n$ dimension. Therefore, only $k$ needs to be determined.

The cluster group is selected based on the Smallest Sum of Squared Error (SSE) compared to the average. The K-Means algorithm is used to cluster data per semester, while the elbow method determines its value. The points, SPI, and period columns are used for the clustering process. Some patterns are expected to facilitate the analysis for further evaluation.

The planned concept of analysis consists of the stages of acquiring students' academic and activity data, which are further cleaned. During the cleaning process, unnecessary data are deleted first. Then, it is followed by feature selection which is further processed with the clustering procedure. Then, it is analyzed to draw conclusions. The details of all processes are explained in Fig. 2.

### A. Dataset

Two types of data are obtained from the university. First is the score data for each course offered by the individual students from the Academic Administration Bureau. Second, is the activity data containing the points earned and the event period obtained from the Student and Alumni Administration Bureau. The research uses data on students' performance and activities from the Architecture and Business Management study program classes of 2012.

### B. Cleaning

There are several errors in the data obtained from the two bureaus, so cleaning needs to be carried out to prevent errors from the clustering process. The final score data has several anomalies and empty datasets, such as Not a Number (NaN), 'A1', '2', ' ', and 'D'. Letter adjustments are made for data 'A1' and 'D'. Meanwhile, invalid data (NaN, '2', ' ') are given a different application based on the analysis. The blank data (NaN) depicts students who drop out and are given an F score. The score '2' is only given to one student, leading to an average assessment where B+ is realized. The data ' ' is found for courses with zero credit units, meaning that it does not contribute to the assessment and is given an F score.

Moreover, there is null information in the position column in the activity data, such as the 'Theater Student Activity Unit Participants'. An analysis is carried out based on the points obtained while similar activities in the position column are labeled. For the period column, there are 200 blank data. Therefore, it is checked, and an appropriate value is assigned. Some data also have a blank period column. In this situation, an analysis is carried out using other datasets to obtain the appropriate value. After filling in the blank period column, 563 rows with inappropriate period formats such as 1.0, 2.0, 130.0, 11.0, 31.0, 909.0, 592.0, 15.0, 14.0, 155.0, 631.0, 88.0, 32.0, 16.0, 160.0, 219.0, 999.0, 19.0, 22021.0, and 6.0, are discovered. So, an analysis is carried out based on existing data to provide the appropriate value for the period.

### C. Feature Selection

The selection of data for the clustering process is performed. The information used results from combining the score and activity columns, as shown in Table I. The four columns used for analysis are Nomor Registrasi Pokok (NRP), SPI, activity points, and period. NRP is a sign of student identity, while SPI is the undergraduates' semester scores ranging from 0.0 to 4.0. It shows 0.0 as the lowest score, meaning they

TABLE I
DATA ON SEMESTER PERFORMANCE INDEX (SPI), ACTIVITY POINTS, AND PERIOD.

| No | NRP | SPI | Activity Points | Period |
|---|---|---|---|---|
| 1 | 22412009 | 3.570 | 0.00 | 151 |
| 2 | 31412016 | 3.520 | 6.80 | 131 |
| 3 | 22412145 | 2.500 | 3.00 | 181 |
| 4 | 22412066 | 3.375 | 4.95 | 161 |
| 5 | 31412038 | 3.110 | 30.90 | 131 |
| 6 | 31412116 | 3.380 | 39.90 | 121 |

are given an F in each course. Meanwhile, 4.0 is the highest score, with an A in each course.

Then, the activity points are the scores of activities the undergraduates engaged in, ranging from 0.0 to 150. Higher activity points indicate that the students actively participate in many activities within that period. It marks the year and semester when the student is given the existing scores and points. An example is 141. The first two numbers represent the year, while the third depicts the odd (1) or even (2) period. The odd period ranges from September to December, while the even is from February to July. The SPI, activity points, and period columns are used for the clustering process to find the pattern.

### D. Data Normalization

Furthermore, the existing data are normalized in the SPI and activity points columns because these two have quite different ranges. SPI has values from 0.0 to 4.0, while points are usually 0 to 150. The two columns are within a similar range from -1 to 1.

### E. Clustering

From the normalized data, the scores and activities for each semester are included in the plot to determine the optimal total number of clusters. The aggregation of grades and activities in every semester is carried out to assess the existing data migration. From the total amount of data and the existing distribution in both, a minimum total cluster of 4 is selected. Furthermore, the Business Management study program in Table II shows that the quite optimal value is found in the total cluster of 6. Meanwhile, Table III shows that from the Architecture study program, a quite optimal silhouette score is in the total cluster of 5.

### F. Results Analysis

After the clustering process has been completed, the next step is to analyze the existing results, search for the overall pattern of each class and study programs used, and compare the distribution procedure during each semester. Once the clustering process has been

TABLE II
SILHOUETTE SCORE FOR BUSINESS MANAGEMENT STUDY PROGRAM.

| Total Cluster | Silhouette Score |
|---|---|
| 2 | 0.4300798 |
| 3 | 0.4721699 |
| 4 | 0.3882946 |
| 5 | 0.4022031 |
| 6 | 0.4060820 |
| 7 | 0.3718710 |
| 8 | 0.3917870 |

TABLE III
SILHOUETTE SCORE FOR ARCHITECTURE STUDY PROGRAM.

| Total Cluster | Silhouette Score |
|---|---|
| 2 | 0.4467967 |
| 3 | 0.4695039 |
| 4 | 0.3975454 |
| 5 | 0.4030637 |
| 6 | 0.3843393 |
| 7 | 0.3813336 |
| 8 | 0.3740467 |

completed, the next step is to analyze the existing results. It searches for the overall pattern of each class and study programs used and compares the distribution procedure during each semester. This analysis can provide valuable insights into the effectiveness of the clustering algorithm and help identify any trends or patterns in the data. Additionally, the results of this analysis can be used to make data-driven decisions and improvements in various domains, such as academic program planning, student's performance evaluation, and resource allocation.

## III. RESULTS AND DISCUSSION

### A. Explanation of the Data Used

Students' data from the two study programs, Architecture and Business Management of the 2012 classes, are utilized. Regarding courses offered, students with a GPA of less than three can only get a maximum of 20 credits. Meanwhile, those with a GPA greater than three have 20 to 24 credits.

### B. Results Based on Cluster Grouping

By using K-Means, clusters are formed from the entered data. The data are further grouped based on the year of entry, the enrolled study program, and the number of clusters in each semester. In Table IV, the first column is the study program to be analyzed. The numbers in the next column heading, for instance, 1, 2, 3, . . . , 14, are the semesters where the maximum is 14. The number in each semester column is equivalent to the clusters, with the largest and smallest being seven and one, respectively.

TABLE IV
TOTAL CLUSTER CREATED FOR EVERY SEMESTER OF ARCHITECTURE AND BUSINESS MANAGEMENT STUDY PROGRAM DATA.

| Semester | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Architecture | 4 | 5 | 4 | 5 | 4 | 5 | 4 | 5 | 3 | 3 | 3 | 2 | 2 | 2 |
| Business Management | 5 | 6 | 6 | 6 | 5 | 6 | 4 | 6 | 3 | 3 | 3 | 3 | 2 | 1 |

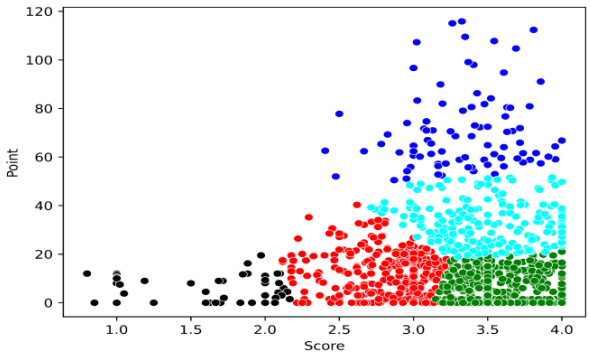


Fig. 3. Clusters of students' points and their scores per semester in the Architecture study program.

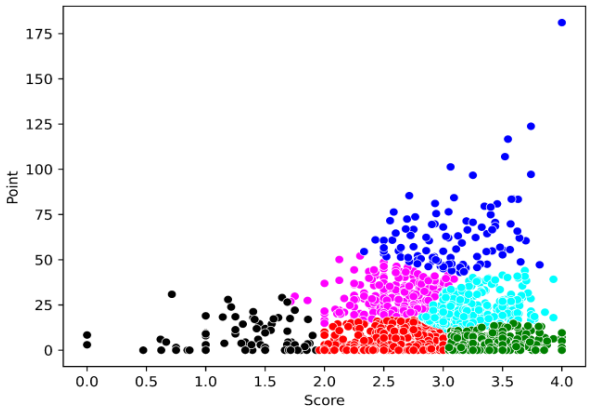


Fig. 4. Clusters of students' points and their scores per semester in the Business Management study program.

Figures 3 and 4 show the overall data from all semesters for the Architecture and Business Management Study programs. One point in the two figures represents students' data for a semester with final points and scores. The distribution of scores and activity points for the Architecture and Business Management study programs are shown in Tables V and VI.

Figure 3 and Table V show that the Architecture study program has a total of five clusters. The analysis shows that cluster zero or red represents the most active students with high points and good scores. Cluster one or blue depicts those with high scores but not very active. Cluster two or black illustrates low scores and activity. Cluster three or green represents students with

TABLE V
DISTRIBUTION OF SCORE AND POINT DATA FROM EACH CLUSTER IN THE RESPECTIVE SEMESTER OF THE ARCHITECTURE STUDY PROGRAM.

| Cluster | Color | Max. Score | Min. Score | Max. Point | Min. Point | Total Students |
|---|---|---|---|---|---|---|
| 0 | Blue | 4.00 | 2.40 | 115.9 | 50.5 | 86 |
| 1 | Green | 4.00 | 3.17 | 21.0 | 0.0 | 380 |
| 2 | Black | 2.17 | 0.80 | 19.5 | 0.0 | 65 |
| 3 | Red | 3.24 | 2.12 | 35.2 | 0.0 | 340 |
| 4 | Cyan | 4.00 | 2.62 | 51.6 | 19.0 | 200 |

TABLE VI
DISTRIBUTION OF SCORE AND POINT DATA FROM EACH CLUSTER IN THE RESPECTIVE SEMESTER OF THE BUSINESS MANAGEMENT STUDY PROGRAM.

| Cluster | Color | Max. Score | Min. Score | Max. Point | Min. Point | Total Students |
|---|---|---|---|---|---|---|
| 0 | Cyan | 3.93 | 2.81 | 43.9 | 11.1 | 348 |
| 1 | Magenta | 3.04 | 1.64 | 52.1 | 14.5 | 266 |
| 2 | Black | 1.93 | 0.00 | 30.9 | 0.0 | 114 |
| 3 | Blue | 4.00 | 2.33 | 181.1 | 43.0 | 107 |
| 4 | Green | 4.00 | 3.05 | 18.0 | 0.0 | 488 |
| 5 | Red | 3.05 | 1.96 | 16.6 | 0.0 | 735 |

median scores and activity. Cluster four or cyan depicts the group with absolutely high scores. It is more spread out to low scores but with fairly high activity.

Figure 4 and Table VI show that the Business Management study program has a total of six clusters. Based on the analysis, cluster zero or cyan represents active students with good scores. Cluster one or magenta depicts those with not too high scores but are more active than those in cluster zero. However, cluster two or black represents low scores and activeness, but it is not the least. Cluster three or blue illustrates undergraduates with quite low to highest scores and activity points. Cluster four or green depicts those with high scores but low activity. Cluster five or red has students with scores that are not as high as cluster one but have lower activity. Based on Fig. 4, cluster three (blue) is a case where students' activities have a positive effect because the highest points are balanced with high scores.

From the two clusters, the maximum score in both study programs is the same as the SPI of 4.0. Meanwhile, the lowest is for Business Management, with an SPI of 0. Then, the highest points achieved are 181.1 for Business Management and 115.9 for Architecture.

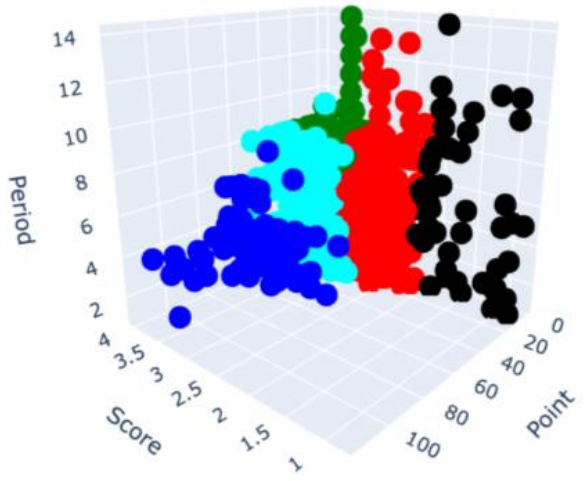


Fig. 5. 3D plot of every student point and their score per semester in the Architecture study program.

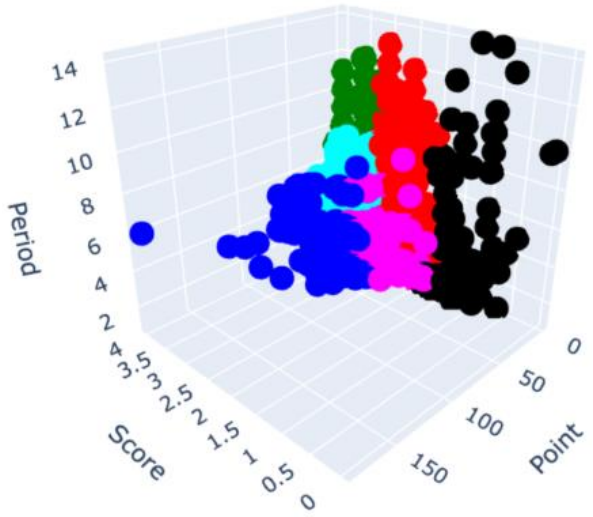


Fig. 6. 3D plot of every student point and their score per semester in the Business Management study program.

### C. Results Based on a 3D Graph Plot

Based on Fig. 5 in the Architecture study program, a fairly high point increase is evident in the fourth, fifth, and sixth semesters. The distribution of scores is also slightly dominated by an SPI that is greater than 2.5. The greater the points are achieved, the higher the scores will be. Unfortunately, students with low scores do not actively participate in these activities.

Based on Fig. 6 in the Business Management study program, an increase in points is also evident in the fourth, fifth, and sixth semesters. A student owns the most prominent points in the sixth semester. Moreover, the score is also quite dominated by an SPI greater than 2.5. An exceptional score also follows the distribution of high points.

### D. Results Based on Migration from One Semester to the Next

Figures 7 and 8 show the migration process from existing clusters in the Architecture and Business Management study programs, respectively. The vertical rectangular block represents one semester which is divided into several clusters. The block colors represent clusters that define the outcome of the k-Means analysis for the students. The connections among the blocks indicate the presence of a student who transitions from one cluster in the previous block to another cluster in the subsequent block. A total of 14 semesters represents the maximum cycle of lectures. It illustrates that during the initial semester, students are distributed across various clusters. This distribution gradually narrows in the subsequent semester due to a decline in the student population. The details have been shown in Tables V and VI previously.

Figure 8 shows that in the fourteenth semester, the Business Management study program only has one color, black. In Table VI, black shows that the score and points are less than 1.92 and 30.9, respectively. The line between the clusters shows the behavioral movement of existing users. In the twelfth and thirteenth blocks, an absence of lines is noticeable. This pattern signifies a break for students during the twelfth semester, with a subsequent continuation of their academic journey in the thirteenth semester.

In Tables VII and VIII, there are four rows of data. The first row indicates activity, followed by the increased students’ performances. The second line depicts increased performance but reduced activity. The third row represents increased activity which leads to reduced scores. The fourth line indicates reduced performances followed by decreased activity. The analysis is carried out in the first to the eighth semester because they have the most active students. Generally, these undergraduates actively participate in these activities and study during these semesters.

Based on Table VII for the Architecture study program, several cases trigger the students’ scores. It aligns with the number of activities engaged in from the first semester to the sixth semester. However, in the same semesters, several activities lead to a drastic reduction in scores. It simply implies that students’ activities do not significantly affect the given scores. The increase or decrease in performance depends on the student’s ability to manage their time effectively.

Based on Table VIII for the Business Management

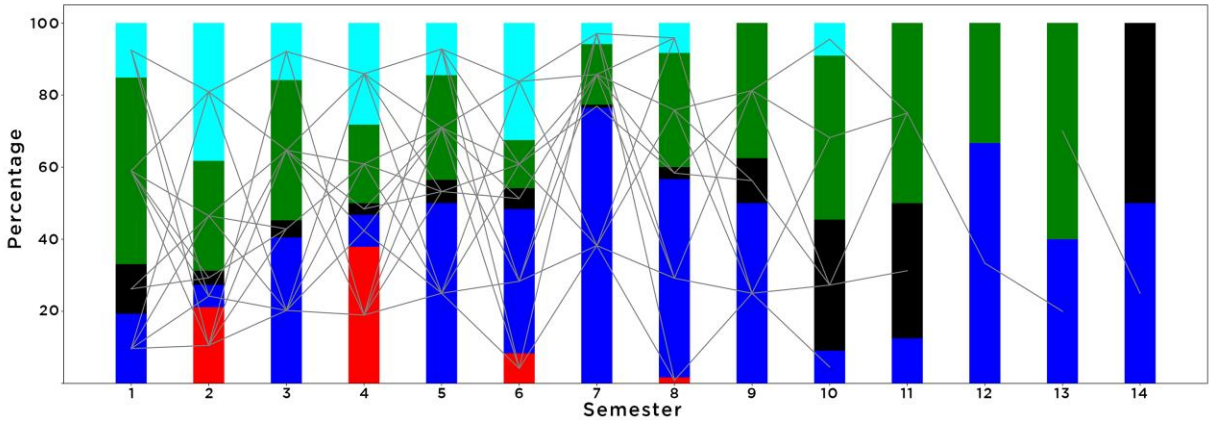


Fig. 7. Cluster created for the fourteenth semester (seven years) and students' migration every semester in the Architecture study program.

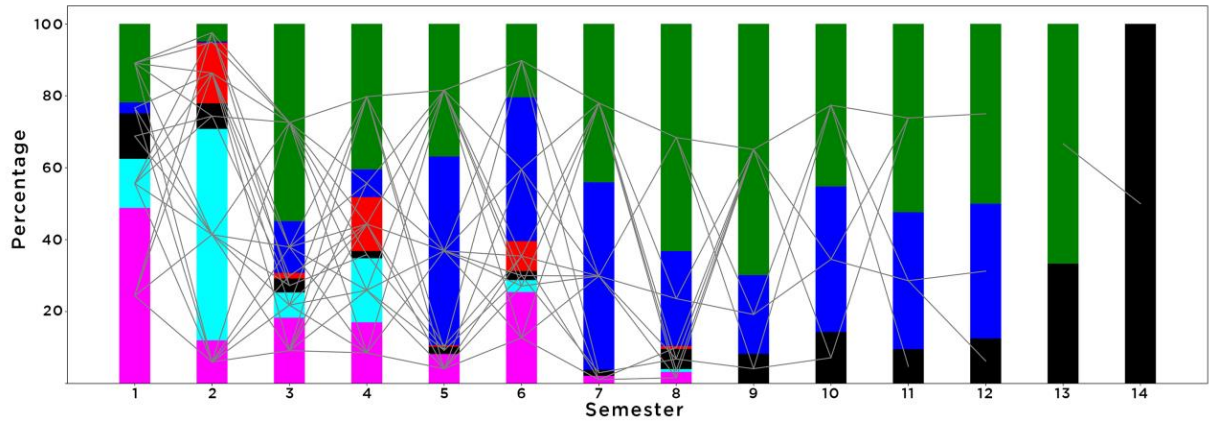


Fig. 8. Cluster created for the fourteenth semester (seven years) and students' migration every semester in the Business Management study program.

study program, a decrease in activities causes an increase in scores and vice versa. It is evident that there are increased scores from the second semester to the seventh semester while the points decrease. Meanwhile, from the first semester to the sixth semester, an increase in activities causes a decrease in the scores. There are mixed cases in other semesters, but the population is not large. Therefore, it is concluded that students' activities greatly affect their scores.

TABLE VII
TOTAL NUMBER OF STUDENTS PER SCORE AND POINTS FOR EIGHT SEMESTERS IN ARCHITECTURE STUDY PROGRAM.

| Changes | 1-2 | 2-3 | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 |
|---|---|---|---|---|---|---|---|
| Increased Scores, Increased Points | 56 | 2 | 61 | 4 | 56 | 6 | 19 |
| Increased Scores, Decreased Points | 16 | 64 | 1 | 77 | 27 | 35 | 19 |
| Decreased Scores, Increased Points | 44 | 8 | 57 | 4 | 18 | 10 | 29 |
| Decreased Scores, Decreased Points | 10 | 48 | 4 | 35 | 8 | 56 | 26 |

### *E. Discussion*

In previous studies, the research topics vary and are related to research from various perspectives. Most of the previous research focuses only on clustering and data mining methods [5, 7, 11, 13, 14, 16], including those that strengthen clustering with ensemble learning [8]. These studies show that K-Means is the most effective and widely used clustering method. Several other studies attempt to predict student dropout rates. However, they are limited to the engineering field of study [17], only discuss dropout predictions based on previous semester results [6], or are based on influencing factors and educational agents [15]. Some develop dropout prediction systems by calculating predictions per student [10] and subsequently develop an uplift model to prevent dropouts [18]. These studies observe that psychological factors can increase

TABLE VIII
TOTAL NUMBER OF STUDENTS PER SCORE AND POINT FOR EIGHT SEMESTERS OF BUSINESS MANAGEMENT STUDY PROGRAM.

| Changes | 1-2 | 2-3 | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 |
|---|---|---|---|---|---|---|---|
| Increased scores, increased points | 27 | 4 | 85 | 9 | 99 | 8 | 24 |
| Increased scores, decreased points | 1 | 186 | 24 | 181 | 21 | 86 | 23 |
| Decreased scores, increased points | 208 | 1 | 87 | 1 | 79 | 10 | 35 |
| Decreased scores, decreased points | 27 | 60 | 41 | 39 | 20 | 123 | 20 |

dropout cases [21]. However, in previous studies [3, 4] the effect of student activities on dropout occurrences is not evaluated. Although research that analyzes the effect of student activities on dropout occurrences has been conducted, it fails to address the effect of student activities on performance for all courses taken beyond some activities, such as sports. The effect on students' activities from one semester to the next is also not examined. Therefore, that is the strengths and differences in the research.

In addition to combining the analysis of all course data taken by students and their activities, it also evaluates the development of studies from one semester to the next. It is intended to determine in which semester the influence of student activities to have a decrease or increase in grades. It is clearly evident in Table VII for the Architecture study program, especially from odd to even semesters. For example, from semesters 1 to 2, semesters 3 to 4, and semesters 5 to 6, there is an improvement in students' performance both in grades and activities. Then, from even to odd semesters, i.e., from semesters 2 to 3, semesters 4 to 5, and 6 to 7, there is an increase in grades, but it is not as significant as the average improvement from odd to even semesters. Overall, in the Architecture study program, it can be said that students' activities do not have much influence on academic performance.

The situation is different in the Business Management study program. In the second row of Table VIII, it can be observed that some students improve their academic performance but are forced to reduce their extracurricular activities. The same applies to the third row of Table VIII, where a considerable number of students actively participate in extracurricular activities, but their academic performance tends to decline, especially from odd to even semesters. Overall, in the Business Management study program, extracurricular activities significantly impact academic performance, and there is a general decline in performance as the level of extracurricular involvement increases.

## IV. CONCLUSION

The K-Means method gives good results. The clusters formed show non-homogenous groups and produce insights from several groups. It also facilitates research to be carried out from existing data. One cluster in the Architecture and Business Management study programs depicts several students' activities and increases academic performance. It proves that students' activities do not completely disrupt lectures. Besides, the majority of these undergraduates are motivated to learn. Most of them participated in these events during the middle semesters (four to six) because they are not used to engaging in these activities in the early semester, and they have been busy writing their final year thesis. Many students' activities are followed by high performance. However, there are differences between the two study programs based on the migration process. Students' activities in the Architecture study program do not significantly affect their scores. Meanwhile, in the Business Management study program, it affects their performance significantly.

The cluster evolution method is proven to be quite effective for observing changes in the data, especially its analysis with multiple periods. The insights obtained from migration characteristics in each semester can be used by universities to understand the characteristics of their students. It is a first step that can be applied later to all departments of universities for a better understanding of each department.

In conclusion, although the research has given satisfactory insight into the effect of students' activities, there are limitations, such as not knowing what activities they engage in. An activity may require more effort than another. The role of the student, as well as the length and scope of the activity, also determine how much effort that needs to be exerted. Therefore, future research needs to gain deep details of their activities and determine the weight based on other factors. It is also recommended to research other study programs and use data from different years to get comprehensive results for all students.

## REFERENCES

[1] J. D. Worsley, P. Harrison, and R. Corcoran, "Bridging the gap: Exploring the unique transition from home, school or college into university," *Frontiers in Public Health*, vol. 9, pp. 1–12, 2021.

[2] R. V. Adams and E. Blair, "Impact of time management behaviors on undergraduate engineering students' performance," *Sage Open*, vol. 9, no. 1, pp. 1–11, 2019.

[3] S. A. Priyambada, M. Er, B. N. Yahya, and T. Usagawa, "Profile-based cluster evolution analysis:

Identification of migration patterns for understanding student learning behavior," *IEEE Access*, vol. 9, pp. 101 718–101 728, 2021.

[4] T. Hasbun, A. Araya, and J. Villalon, "Extracurricular activities as dropout prediction factors in higher education using decision trees," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. Austin, TX, USA: IEEE, July 25–28, 2016, pp. 242–244.

[5] C. Anuradha, T. Velmurugan, and R. Anandavally, "Clustering algorithms in educational data mining: A review," *International Journal of Power Control and Computation(IJPCSC)*, vol. 7, no. 1, pp. 47–52, 2015.

[6] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009*, Cordoba, Spain, July 1–3, 2009, pp. 41–50.

[7] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.

[8] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471–1483, 2020.

[9] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, pp. 1–21, 2019.

[10] S. Guzmán-Castillo, F. Körner, J. I. Pantoja-García, L. Nieto-Ramos, Y. Gómez-Charris, A. Castro-Sarmiento, and A. R. Romero-Conrado, "Implementation of a predictive information system for university dropout prevention," *Procedia Computer Science*, vol. 198, pp. 566–571, 2022.

[11] S. Križanić, "Educational data mining using cluster analysis and decision tree technique: A case study," *International Journal of Engineering Business Management*, vol. 12, pp. 1–9, 2020.

[12] A. J. Viloria Silva, T. J. Crissien Borrero, J. Vargas Villa, M. Torres Samuel, J. E. Garcia Guiliany, C. Vargas Mercado, N. Orellano Llinas, and K. Batista Zea, "Differential evolution clustering and data mining for determining learning routes in Moodle," in *International Conference on Data Mining and Big Data*, vol. 1071. Chiang Mai, Thailand: Springer, July 26–30, 2019, pp. 170–178.

[13] R. Ramon-Gonen and R. Gelbard, "Cluster evolution analysis: Identification and detection of similar clusters and migration patterns," *Expert Systems with Applications*, vol. 83, pp. 363–378, 2017.

[14] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15 991–16 005, 2017.

[15] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, pp. 1–12, 2022.

[16] N. Patel, C. Sellman, and D. Lomas, "Mining frequent learning pathways from a large educational dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.11125

[17] I. Sandoval-Palis, D. Naranjo, J. Vidal, and R. Gilar-Corbi, "Early dropout prediction model: A case study of university leveling course students," *Sustainability*, vol. 12, no. 22, pp. 1–17, 2020.

[18] D. Olaya, J. Vásquez, S. Maldonado, J. Miranda, and W. Verbeke, "Uplift modeling for preventing student dropout in higher education," *Decision Support Systems*, vol. 134, pp. 1–11, 2020.

[19] W. Hachicha, L. Ghorbel, R. Champagnat, C. A. Zayani, and I. Amous, "Using process mining for learning resource recommendation: A Moodle case study," *Procedia Computer Science*, vol. 192, pp. 853–862, 2021.

[20] I. Dronyuk, V. Verhun, and E. Benova, "Non-academic factors impacting analysis of the student's the qualifying test results," *Procedia Computer Science*, vol. 155, pp. 593–598, 2019.

[21] M. Parviainen, K. Aunola, M. Torppa, A.-M. Poikkeus, and K. Vasalampi, "Symptoms of psychological ill-being and school dropout intentions among upper secondary education students: A person-centered approach," *Learning and Individual Differences*, vol. 80, pp. 1–11, 2020.