Jurnal Teknologi Informasi dan Pendidikan
Volume 17, No. 1, March 2024
https://doi.org/10.24036/jtip.v17i1.571

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

The Correlation of High School's Report Values with Grade Point Semester in College's First Year Using Principal Component Analysis

Amalia Yuli Astuti[1*]✉, Nungky Dwi Putri[1]
[1]Industrial Engineering Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Corresponding Author: amalia.yuliastuti@ie.uad.ac.id

Article Information

Article history:

No. 571
Rec. March 28, 2022
Rev. January 04, 2024
Acc. January 05, 2024
Pub. January 11, 2024
Page. 62 – 74

Keywords:

- *First Year of College*
- *Grade Point Semester*
- *High School Report Value*
- *Correlation*
- *Principal Component Analysis*

ABSTRACT

Department X of a university in Yogyakarta required to evaluate the learning achievement result in the first year. There were students that failed the first year and become bottlenecks for graduating on time. The study period of graduates and the percentage of students that graduate on time were important in the accreditation of study program. The first academic year in college was a crucial year that becomes a fundamental learning momentum for every student. The research purpose was to examine the correlation of students' high school report values to their grade point semester in the first year using Principal Component Analysis. The students' data involved were the class from 2015 to 2018. Based on the results of data processing, it was found that there was weak correlation between the student report cards in senior secondary education with the grade point in semesters 1 and 2. However, there was strong correlation between the grades for Chemistry and Biology. Department X need to take action to determine others subjects for students' admission.

How to Cite:

Astuti, A. Y., Putri, N. D. (2024). The Correlation of High School's Report Values with Grade Point Semester in College's First Year Using Principal Component Analysis. Jurnal Teknologi Informasi Dan Pendidikan, 17(1), 62-74. https://doi.org/10.24036/jtip.v17i1.571

1. INTRODUCTION

First year in higher education is one of the factors that can determine students' academic success in the undergraduate program. It was found that the first semester determines whether students will graduate on time, drop out or not continue to the second year of study [1]. Each student in the undergraduate program is given an academic time

from four until seven years. The standard for graduating on time is four years. If there are students that graduating more than four years then it will have impact on academic loads on departments.

In the department X on a university, it was found that the lecturing implementation in the courses was not efficient. The data about students that graduated on time was not on the department's target. Department X has implemented an academic evaluation program for monitoring the students' study progress. But there was data about students that still continue their studies until seven years.

The educational background of students before college became an assumption for the students could finish their studies on time. Their scores on senior high schools became one of the factors that was considerate for Grade Points Average (GPA) in the first year [2]–[4]. A university in the Philippines had found that the academic performance at high schools influenced GPA in college for three groups of students [5].

One factor that influenced increasing the quality of undergraduate students was the quality of the students' candidates as the input for college. The parameters for determining the input quality were the score of students' candidates when they were in high schools. The processes for getting good qualified new students in college were done by admission screening or selection with using academic achievements tracks or university scholastic tests [6].

The university admission programs also have influenced on GPA. The admission systems in the university that became case study has two paths. The first is with the high schools' reports invitation system and the second is scholastic tests. It was found that the most students in the department X came from the high schools' reports invitation system. The students in the public university that came from high schools' reports invitation system, that have been called Joint Entrance Selection of State Universities or SBMPTN, have better academic achievements than students that came from the university's scholastic tests [7]. In the previous research, it was found that students' GPA that came from university scholastic tests have differences with students' that came from SBMPTN [8]–[11]. But the department X is in a private university. The scholastic tests on private universities didn't have the same systems with the public universities so the admission systems were also different. The other factor is also the ranking of the private universities. If private universities have high ranking on some systems, then the students' candidate of admissions system were more than the lower one. The department X didn't have a privilege to be the first choice of prospective students. Therefore, the admissions to private universities have opened early and used the high schools' reports for the students' candidate.

The purpose of this research investigated the correlation of students' high school reports with the student's grade points in the first and second semesters using Principal Component Analysis (PCA). In the other side, the relationship about students' scores in high school's reports and GPA needed to be examined because the success of student lectures were based on the first semester [12]. Department X has already implemented academic

strategies and used higher points for students' admission. But the results from the strategies didn't meet the expectation. The investigation on the relationship of data that existed in the higher education could explain the phenomena happened. The existing data that can analyzed are students' grade points in first year and high schools' reports that used in the students' admission.

2. RESEARCH METHOD

The object of this research was the data of students' admissions from 2015 until 2018. The other data needed were the first and second semesters' grade point. The data were collected from academic database in the university's information systems. The method for data processing was Principal Component Analysis (PCA) to found out the correlation and dimensions of students' report cards and their grade points.

PCA is a data processing technique based on multivariate that can analyze observed data by describing the correlation of the variables quantitatively [13]. PCA worked by reducing the data dimensions to become principal components that could be identifying the direction of data relationship and examined the data variations maximally [14]. PCA has a purpose to extract the new information from the data [13] and could develop the new variables from data combinations that correlated [14]. PCA became a method that was suitable for this research to examine the factors correlated in the scores of students' high school reports and the GPA in the first and second semesters.

The data processing in this research were using RStudio version 3.5. R Studio is one of data mining processing software. Now, RStudio was used for data mining broadly cause the packages for data processing could produce outputs' visualization. The research flowchart can be found in Figure 1.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

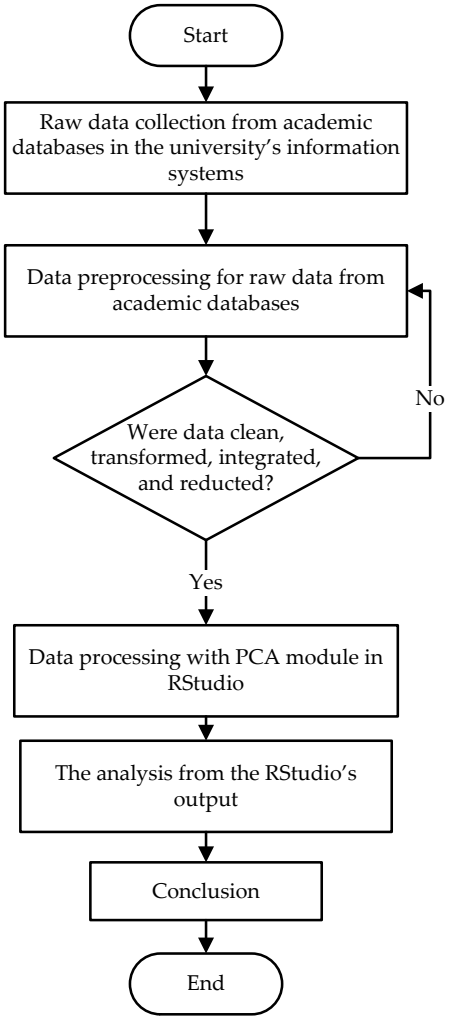


Figure 1. The research flowchart

The explanation of this research step as follows:

1. Raw data collection from academic databases. Mostly raw data came from university's information systems that managed data about students' admission and academic reports.
2. The raw data must be in the refined form so this research used data preprocessing. The raw data were examined in the phase of:
   a. Data cleaning
   b. Data transformation
   c. Data integration
   d. Data reduction

   If all of the data have already refined, the next step could be data processing with the refined data.

Volume 17, No. 1, March 2024
https://doi.org/10.24036/jtip.v17i1.571

3. The data that came from data preprocessing was ready to calculate. The data were processed using PCA modules in RStudio. The PCA modules are FactoMineR dan factoextra packages. The steps to use PCA modules in R:
   a. Installing FactoMineR and factoextra packages
   b. Inputting the data
   c. Data setting for calculating with active data for PCA
   d. Use the PCA's residue into the new data variable name
   e. Calculate the eigen value and variance
   f. Installing the visualization packages with ggplot2 and corrplot to get the output
   g. Using the visualization packages to get results
   4. Analysis the output from visualization processing from PCA residue's plot and correlation plot.
5. Conclusion of the analysis and discussed the results.

3. RESULTS AND DISCUSSION

The raw data from the database could not be examined on RStudio. The raw data should be checked using data preprocessing for better results and quality. Data cleaning and data transformation gave many benefits for getting refined data. The use of data cleaning because there were many missing data. Missing data was caused by the students that were not active in the system or the data were not completed when they input it. The data transformation was used to standardize the scale in the high school scores reports. The raw data proportion that came after data preprocessing can be found in Figure 2. The total refined data after data preprocessing was 503 students. The refined data after data preprocessing were processed with PCA and used RStudio version 3.5.

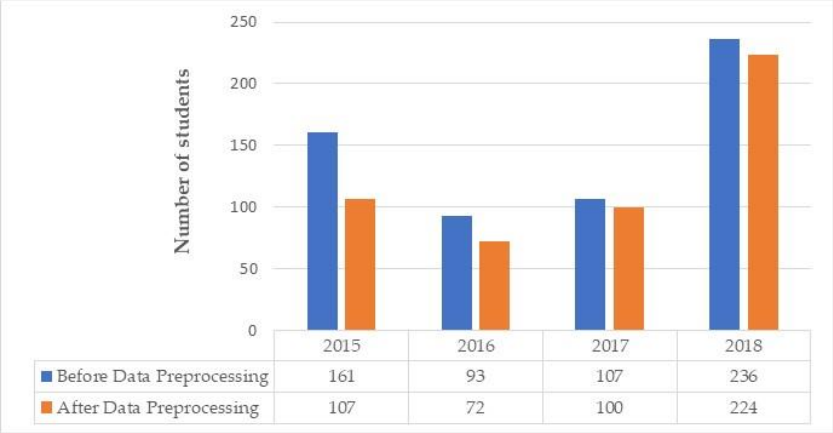


Figure 2. The data proportion before and after *data preprocessing*

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

3.1. The output from data processing in PCA module of RStudio

The results from data mining with using PCA in the RStudio were found that the orthogonal dimensions of each variable have not yet seen clearly. In Figure 3, it can be seen that there was a buildup of the orthogonal directions of each variable between one another. It was found that the results of the variable's contributions percentage in the first dimension was 43.07% and in the second dimension was 25.26%. The total two dimensions' contributions percentage was 68.33%.

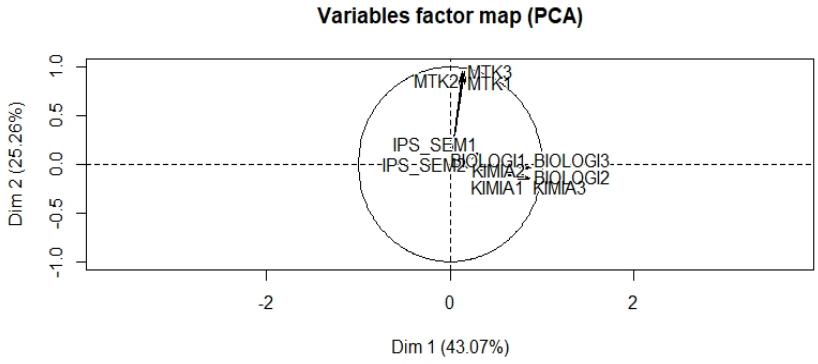


Figure 3. Orthogonal dimensions from PCA's visualization result

The next step for PCA was to observe the formation of dimensions using the correlation plot visualization in the RStudio. The result of the correlation plot's visualization can be seen in Figure 4. It can be found that the correlation of the scores from chemistry and biology in the students' high school reports' cards. The scores of students' chemistry and biology formed in the first dimension. The scores of mathematics formed in the second dimension. The third dimension was formed by the first and second semesters' grade points in college.

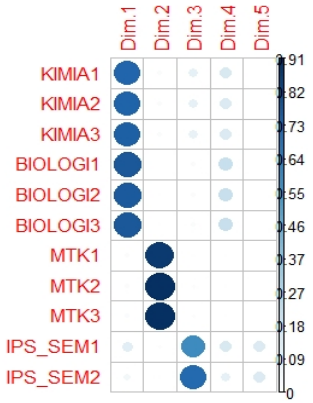


Figure 4. The correlation plot's visualization using PCA in the RStudio

Jurnal Teknologi Informasi dan Pendidikan
Volume 17, No. 1, March 2024
https://doi.org/10.24036/jtip.v17i1.571

3.2. The correlation output of PCA's dimensions

The output from each dimension in PCA explained the contribution of variables. The output of dimension-1 can be found in Figure 5 (a). The variables in the dimension-1 were Biology-1, Biology-3, Biology-2, Chemistry-3, Chemistry-2, and Chemistry-1 that have most contributions. While the variables in the dimension-2 that have most contribution can be found in Figure 5 (b). Three variables in dimension-2 were Math-3, Math-2, Math-1 that have most contribution. Then there were two variables in dimension-3 that have most contribution. The variables in dimension 3 were Grade Point Semester-1 (GPS-1) and Grade Point Semester-2 (GPS-2). The contribution from each variable were scattered in each dimension except for Biology and Chemistry. There were only a little contribution of GPS-1 and GPS-2 in dimension-1 so the correlation of Biology and Chemistry to GPS-1 and GPS-2 were too low. Then GPS-1 and GPS-2 also had low contribution in dimension-2 so the correlation of Math to GPS-1 and GPS-2 were low too. GPS-1 and GPS-2 had higher contribution in dimension-3 but its correlation to Biology, Chemistry and Math were too low.

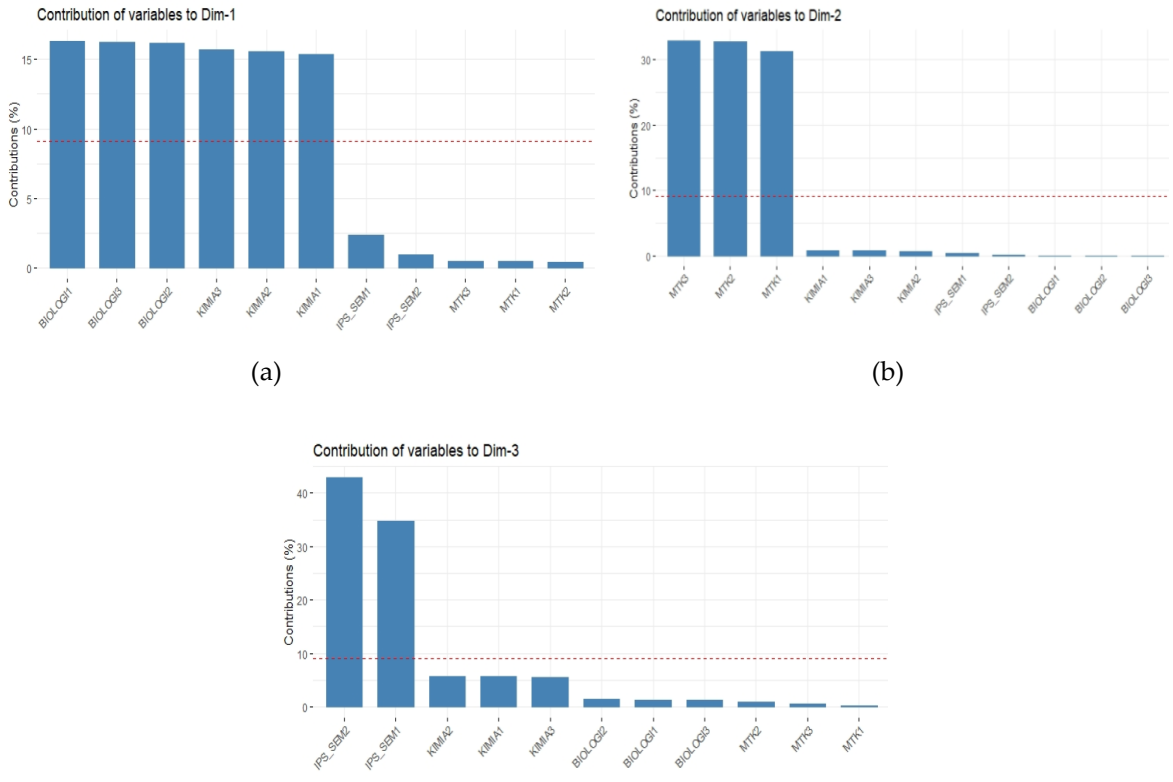


Figure 5. The contribution of variables for (a) PCA's dimension-1, (b) PCA's dimension-2, and (c) PCA's dimension-3

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

The correlation output from data processing with R from each dimension could be checked. These six variables have strong correlation in dimension-1 and the data can be seen in Table 1. The correlation table output in the Table 1 contrasted to the output from Figure 5 (a). The variables with strong correlation were Biology-1, Biology-3, Biology-2, Chemistry-3, Chemistry-2, Chemistry-1. GPS-1, GPS-2, and Mathematics had weak correlation in dimension-1.

Table 1. The output of correlation in dimension-1

Variables	Correlation
Biology-1	0.878287
Biology-3	0.877427
Biology-2	0.875644
Chemistry-3	0.863046
Chemistry-2	0.858901
Chemistry-1	0.852041
GPS-1	0.333871
GPS-2	0.213979
Mathematics-3	0.153399
Mathematics-1	0.145956
Mathematics-2	0.136205

The correlation output from dimension-2 could be found in Table 2. The correlation output from Table 2 explained three variables with strong correlation and the others were weak in dimension-2. These variables were Mathematics-3, Mathematics-2 and Mathematics-1. GPS-1 had weak correlation in dimension-2 and GPS-2 didn't have correlation. Chemistry had weak negative correlation in dimension-2.

Table 2. The output of correlation in dimension-2

Variables	Correlation
Mathematics -3	0.955853
Mathematics -2	0.954664
Mathematics -1	0.931967
GPS-1	0.114992
Chemistry-2	-0.14223
Chemistry-3	-0.14634
Chemistry-1	-0.14766

The output about the correlation in dimension-3 in Table 3 explained only two variables that have strong correlation. These variables were GPS-2 and GPS-1. The school reports cards subject had weak correlation. Biology had very weak correlation. Chemistry and Mathematics had weak yet negative correlation.

Table 3. The output of correlation in dimension-3

Variables	Correlation
GPS-2	0.843546
GPS-1	0.758528
Biology-2	0.149393
Biology-1	0.147986
Biology-3	0.147756
Mathematics-3	-0.09978
Mathematics-2	-0.12035
Chemistry-3	-0.30004
Chemistry-1	-0.30671
Chemistry-2	-0.30695

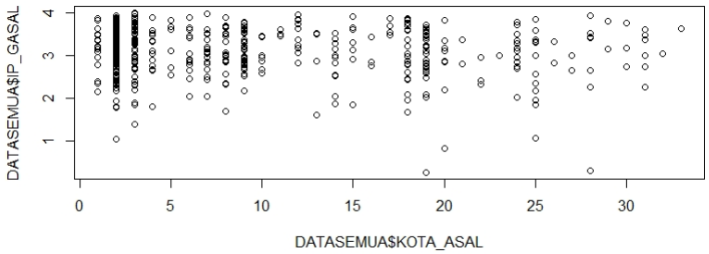
3.3. The analysis from the PCA's output in RStudio

It can be found that the students' high school reports had weak correlations with the first and second semesters' grade points. There were found correlations between the scores in chemistry and biology in PCA's dimension-1. The mathematics scores didn't have a correlation with the others' variables. The first semester's grade points only had strong correlation with the second semester's grade points in PCA's dimension-3. These results corresponded with the previous results that it was found that the means of high schools' reports didn't have correlation with the GPA [15]. The other researchers also found that the students' academic achievements before college have not been able to explain the students' academic success in the first year in the college but if the data of students' scholastic tests results were added together with the high schools' reports that will give more accurate results [16]. It was found that the students' biology scores in high schools reports and the biology scores in the college didn't have a correlation [17].

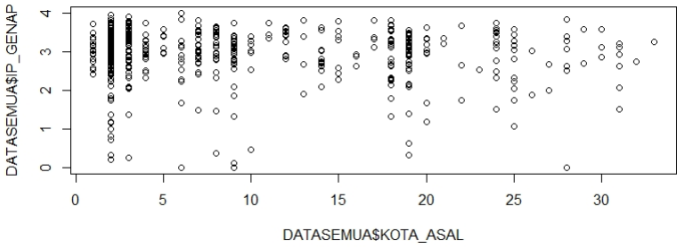
The students in this research had quite high variation, especially in their learning background before entering college. The students' hometown in this research were varied from 33 provinces in Indonesia. The students' hometown data can be obtained from Figure 6. Although the high school in Indonesia have already implemented the same curriculum but there was gap about the facilities and resources through all areas in Indonesia. There was consideration about curriculum in high school that differed across any schools and high school GPA still has shortcoming as predictors of undergraduate performance [18]. Indonesia still has problem with education's equity that there are implementation education gap in cities and villages [19].

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Volume 17, No. 1, March 2024
https://doi.org/10.24036/jtip.v17i1.571



(a)



(b)

Figure 6. The output plot that explained students' hometowns from observation data in the first year of college: (a) first semester, (b) second semester

The output plot in Figure 6 has codes from 1 until 33 that explain the origin of provinces of observed students. The codes can be interpreted as:

1. Special Region of Yogyakarta
2. Central Java
3. West Java
4. East Java
5. Banten
6. Bengkulu
7. Jambi
8. Lampung
9. South Sumatra
10. West Sumatra
11. North Sumatra
12. Central Kalimantan
13. West Kalimantan
14. East Kalimantan
15. South Kalimantan
16. North Kalimantan
17. East Nusa Tenggara
18. West Nusa Tenggara
19. Riau
20. Riau Islands
21. Maluku
22. North Maluku
23. South Maluku
24. East Maluku
25. Bangka Belitung Islands
26. Central Sulawesi
27. North Sulawesi
28. South Sulawesi
29. Aceh
30. Papua
31. West Papua
32. DKI Jakarta
33. Gorontalo

There was a presumption that the first and second semesters grade points have a correlation with the other variables. The courses in the first year may have contributed to

the GPA where the courses' propositions have the basic knowledge that was different and far out from the high schools' subjects. The other presumable was the courses in the first year may have correlated with others subjects in high schools' reports. The reason why courses in the first year had not been engaged in the research, caused by the data of courses' scores were the non-parametric type that didn't suitable for the PCA method.

There was strong correlation between chemistry and biology from the result that have obtained. This result suggested that the students' admission process could reduce the data of high schools' reports that must be inputted by applicants. Therefore, biology could be a representative of chemistry as high school values report in students' admission.

4. CONCLUSION

It was found that there was a correlation between chemistry and biology scores in the students' high school reports. Thus, it could choose one of the subject scores from chemistry or biology for the students' admissions process of department X. On another side, it found that students' high school reports like the scores in chemistry, biology, and mathematics didn't have strong correlation with first and second semesters grade points. There was an assumption that it was necessary to involve the first-year course variables to check their correlation with the student's report card scores data. There is also a need for research to analyze the relationship between students' report cards and the data from students' college course grades that require basic knowledge of high school subjects.

ACKNOWLEDGEMENTS

Thank you for study program secretary of department X, Mrs. Utaminingsih Linarti, S.T., M.T., in providing data collection and supporting the use of raw data from academic information systems.

REFERENCES

[1] D. Raju and R. Schumacker, "Exploring student characteristics of retention that lead to graduation in higher education using data mining models," *J. Coll. Student Retent. Res. Theory Pract.*, vol. 16, no. 4, pp. 563–591, 2015, doi: 10.2190/CS.16.4.e.

[2] Sulistiowati, "Pengaruh Jurusan Dan Nilai Sekolah Menengah Terhadap Prestasi Belajar Mahasiswa Program Studi S1 Sistem Informasi STIKOM Surabaya," in *Prosiding Seminar Nasional Sistem & Teknologi Informasi*, 2013, pp. LL-13-LL–20.

[3] K. Al Hazaa *et al.*, "The effects of attendance and high school GPA on student performance in first-year undergraduate courses," *Cogent Educ.*, vol. 8, no. 1, pp. 1–19, 2021, doi: 10.1080/2331186X.2021.1956857.

[4] T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder, "Data Mining and Machine Learning Retention Models in Higher Education," *J. Coll. Student Retent. Res. Theory Pract.*, vol. 0, no. 0,

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

pp. 1–25, 2020, doi: 10.1177/1521025120964920.
[5] M. S. Canque, L. Matthew, C. Derasin, and L. L. Pinatil, "Senior High School Background and GPA of the Education Students in a State University in the Philippines," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 13, pp. 3560–3566, 2021.
[6] Nurhasanah, Purwati, and H. Ahmad, "Pengaruh Sistem Seleksi Masuk Perguruan Tinggi Terhadap Indeks Prestasi Mahasiswa Jurusan Pendidikan Matematika Universitas Papua (UNIPA)," in *Prosiding Seminar Nasional*, 2015, vol. 03, pp. 114–120, [Online]. Available: http://www.journal.uncp.ac.id/index.php/proceding/article/view/780/653.
[7] M. F. Qudratullah, "Pengaruh Jalur Penerimaan Mahasiswa Dan Asal Sekolah Terhadap Prestasi Mahasiswa Di Fakultas Sains Dan Teknologi UIN Sunan Kalijaga," *J. Fourier*, vol. 3, no. 1, pp. 9–15, 2014, doi: 10.14421/fourier.2014.31.9-15.
[8] A. K. Riezky, "Hubungan Hasil Seleksi Penerimaan Mahasiswa Baru dengan Indeks Prestasi Kumulatif pada Mahasiswa Program Studi Pendidikan Dokter Fakultas Kedokteran Universitas Abulyatama," *Seambi Akad.*, vol. IV, no. 2, pp. 91–95, 2016.
[9] M. R. Alkautsar, Susilawati, and M. B. Azhar, "Hubungan Akreditasi Sekolah , Asal Sekolah , Jalur Penerimaan Mahasiswa dan Tempat Tinggal dengan Indeks Prestasi Kumulatif Mahasiswa Proses pembelajaran merupakan kegiatan utama dalam dunia pendidikan , termasuk di Perguruan Tinggi (PT). Keberhasilan p," *Biomed. J. Indones.*, vol. 4, no. 3, pp. 140–148, 2018.
[10] A. M. Annizar and M. Arifin, "Perbedaan Prestasi Belajar Mahasiswa Ditinjau dari Jalur Seleksi Masuk Perguruan Tinggi," *SAP (Susunan Artik. Pendidikan)*, vol. 5, no. 3, pp. 197–204, 2021, doi: 10.30998/sap.v5i3.8411.
[11] M. Idris, "Analisis Prestasi Akademik Mahasiswa Teknik Informatika Institut Teknologi Sumatera (ITERA) Berdasarkan Jalur Seleksi Masuk Perguruan Tinggi Negeri," *J. Sci. Appl. Technol.*, vol. 5, no. 1, pp. 126–130, 2021, doi: 10.35472/jsat.v5i1.410.
[12] D. Raju, "Predicting Student Graduation in Higher Education Using Data Mining Models: A Comparison," The University of Alabama, 2012.
[13] H. Abdi and L. J. Williams, "Principal component analysis. wiley interdisciplinary reviews: computational statistics," in *Wiley Interdisplinary Reviews: Computational Statistics*, 2010, pp. 1–47.
[14] M. Ringnér, "What is principal components analysis?," *What is Princ. Compon. Anal.*, vol. 26, no. 3, pp. 303–304, 2008.
[15] L. Lukmanulhakim, "Nilai Rapor Sekolah Menengah Dan Hubungannya Dengan Indeks Prestasi Komulatif Mahasiswa," *J. Visi Ilmu Pendidik.*, vol. 10, no. 1, pp. 1–7, 2018, doi: 10.26418/jvip.v9i1.25958.
[16] J. Noble and R. Sawyer, "Predicting Different Levels of Academic Success in College Using High School GPA and ACT Composite Score: ACT research report series," 2002. [Online]. Available: http://www.act.org/research/researchers/reports/pdf/ACT_RR2002-4.pdf.
[17] M. T. Sinaga, B. Manurung, and T. Gultom, "Hubungan Nilai Rapor Biologi dengan Kompetensi Biologi Umum I Mahasiswa FMIPA Semester I T.P 2015/2016 Berdasarkan Jalur Masuk di Universitas Negeri Medan," *J. Pendidik. Biol.*, vol. 5, no. 2, pp. 94–98, 2016, doi: 10.24114/jpb.v5i2.4304.
[18] P. A. Westrick, H. Le, S. B. Robbins, J. M. R. Radunzel, and F. L. Schmidt, "College Performance and Retention: A Meta-Analysis of the Predictive Validities of ACT® Scores, High School Grades, and SES," *Educ. Assess.*, vol. 20, no. 1, pp. 23–45, 2015, doi: 10.1080/10627197.2015.997614.

[19] R. Ananda, S. Mulyani, H. Sagita, and H. M. Al Mubarok, "Government Problems and Solution to Improve the Quality and Equity of Education in Indonesia," *Edumaspul J. Pendidik.*, vol. 7, no. 1, pp. 902–909, 2023, doi: 10.33487/edumaspul.v7i1.5718

[20] H. Armi and I. Dewi, "Analysis of Feasibility Level of Interactive Learning Media on Workshop Work Subjects And Technical Drawing", JTIP, vol. 13, no. 2, pp. 81-88, Jan. 2021.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id