Jurnal Teknik Informatika C.I.T Medicom, xx (x) (20xx) xx-xx

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Long Term High School to Higher Education Trend Analysis (2018-2022) using K-Means Methodology

Mohamad Farizal Arifin[1]
Teknik Informatika, Pelita Bangsa University, West Java, Indonesia

Article Info

Article history

Received : diisi oleh editor
Revised : diisi oleh editor
Accepted : diisi oleh editor

Kata Kunci:

Higher Education Analysis;
K-Means Methodology;
Long-Term Trends;
PySpark Analysis;

Abstract

This study, titled "Long-Term High School to Higher Education Trend Analysis (2018-2022) using K-Means Methodology," meticulously explores the intricate landscape of academic trajectories. Employing the robust K-Means methodology, our research aims to unveil patterns characterizing the educational journey from high school to advanced levels. Adhering to Jurnal Teknik Informatika C.I.T's formatting guidelines, the manuscript is meticulously crafted in English with Constantia font size 9, single spacing, and one-sided A4 paper. The study, spanning 3 to 10 pages, integrates graphics and tables for comprehensive presentation, methodically organized into sections like Introduction, Methods, Results and Discussion, Conclusion, and References. Linguistically coherent, the manuscript adheres to Standard English grammar, brief and informative titles, and italicization for foreign terms. Rigorous adherence to formatting guidelines is evident, ensuring readability through the absence of bullet points and numbering, activation of widow or orphan control, and column balance. In conclusion, this study not only advances educational analytics but also provides valuable insights for educational policymakers. The nuanced understanding of long-term trends fosters informed decision-making in the dynamic landscape of education. The abstract serves as a concise summary, falling within the prescribed 180-300 words, offering a glimpse into the aim, research method, results, and conclusions of the paper without references or displayed equations.

Corresponding Author:

Mohamad Farizal Arifin,
Teknik Informatika
Pelita Bangsa University
Jl. Inspeksi Kalimalang Tegal Danas, Bekasi, West Java, Indonesia
farizalarifin@mhs.pelitabangsa.ac.id

This is an open access article under the CC BY-NC license.

1. Introduction

In the realm of educational analytics, our study titled "Long-Term High School to Higher Education Trend Analysis (2018-2022) using K-Means Methodology" delves into the intricate landscape of academic trajectories. Employing the robust K-Means methodology, this research aims to unearth patterns and insights characterizing the educational journey from high school to advanced levels over the period of 2018 to 2022[1], [2]. By adhering to the formatting guidelines set by Jurnal Teknik Informatika C.I.T, this manuscript promises a meticulous and insightful analysis, aligning with

stipulated page limits and language requirements. The use of abbreviation is permitted, but the abbreviation must be written in full and complete when it is mentioned for the first time and it should be written between parentheses. Terms/foreign words or regional words should be written in italics. Notations should be brief and clear and written according to the standardized writing style. Symbols/signs should be clear and distinguishable, such as the use of number 1 and letter l (also number 0 and leter O). In this manuscript doesn't allow to use bullet and numbering. At the end of this paper both of the colomns should be in balance. You also have to activated widow or orphan control in order to ensure that there are no single line of sentence at the end of column section[3], [4], [5].

Executing the study with precision, the manuscript follows the specified formatting guidelines. Written in English with Constantia font size 10, single-spaced, and one-sided on A4 paper, it adheres to strict margin dimensions. Comprising 3 to 10 pages, the manuscript incorporates graphics and tables for a comprehensive presentation. Headings, numbered in Arabic numerals, facilitate structured reading. The manuscript adopts a methodical organization, embracing sections such as Introduction, Methods, Results and Discussion, Conclusion, and References, adhering to the journal's prescribed sequence[6].

To maintain linguistic coherence, the manuscript employs Standard English grammar. The title, succinct within 15 words, offers a brief yet informative encapsulation. Keywords follow the abstract, enhancing discoverability. The first letter of headings is capitalized, and abbreviations are introduced in full upon first mention. Terms in foreign or regional languages appear in italics, ensuring clarity. Notations adhere to standardized writing styles, while symbols are distinct and easily discernible, preventing confusion[7], [8].

Rigorous adherence to formatting guidelines is evident throughout the manuscript. Abbreviations, written in full upon first usage and enclosed in parentheses, enhance reader comprehension. The absence of bullet points and numbering maintains a clean presentation, fostering readability. Activating widow or orphan control and ensuring column balance at the document's conclusion exemplify the meticulous attention to detail. In concert, these elements contribute to a manuscript that not only advances educational analytics but also meets the exacting standards of Jurnal Teknik Informatika C.I.T[9], [10].

2. Research Method

Our study leverages comprehensive data collected from educational institutions spanning the years 2018 to 2022. High-quality datasets encompass student demographics, academic performance, and progression from high school to higher education levels. The diverse and expansive nature of the dataset allows for a nuanced exploration of trends and patterns[11].

Prior to analysis, a rigorous preprocessing phase is undertaken. This involves cleansing the dataset of any inconsistencies, missing values, or outliers that may distort the results. Standardization and normalization techniques are applied to ensure uniformity and comparability across variables, enhancing the reliability of subsequent analyses[12], [13].

To streamline the analysis, relevant variables are carefully selected based on their significance to the educational trajectory. Key indicators include academic performance metrics, socio-economic factors, and geographical considerations. The chosen variables aim to capture the multifaceted aspects influencing the transition from high school to higher education[14], [15].

The core analytical approach involves the application of the K-Means clustering algorithm. This unsupervised machine learning technique classifies the dataset into distinct clusters based on similarities among data points. The algorithm iteratively refines cluster assignments, optimizing the grouping of educational trajectories. The optimal number of clusters is determined through methodologies like the elbow method or silhouette analysis[16], [17], [18], [19].

The validity of the clustering results is assessed through internal and, if applicable, external validation metrics. Subsequently, the identified clusters are interpreted to extract meaningful insights into long-term educational trends. Patterns related to academic pathways, regional disparities, and

socio-economic influences are scrutinized, providing a rich understanding of the dynamics within the dataset[20], [21].

Statistical techniques complement the K-Means analysis to quantify the significance of observed trends. Descriptive statistics, inferential tests, and graphical representations are employed to enhance the interpretation of findings. Visualization tools such as charts and graphs aid in presenting the results in an accessible and insightful manner[22], [23], [24].

Throughout the research process, ethical considerations are paramount. Data privacy and confidentiality are rigorously maintained, adhering to institutional and legal standards. Transparent reporting ensures the responsible use of data and the dissemination of knowledge without compromising the integrity of individuals or institutions involved[25], [26].

3. Result and Discussion

The initial phase of our analysis involves meticulous data processing and feature engineering to extract meaningful insights from the "SMA25tahun.csv" dataset. In this section, we explore the steps taken to preprocess and refine the raw data, emphasizing the pivotal role of the chosen feature, "persentasi," representing the percentage of high school graduates pursuing advanced education.

a. Data Processing and Feature Engineering
The Spark session was initialized, and the dataset "SMA25tahun.csv" was loaded into a PySpark DataFrame. The dataset primarily focuses on the "persentasi" feature, indicating the percentage of high school graduates pursuing higher education. The feature was then selected for further analysis, and a vector assembler was employed to combine the selected feature into a vector.

prediction	nama_tahun	stddev(persentasi)
0	2018-01-01	6.047917435198482
0	2019-01-01	5.889582729271832
0	2020-01-01	5.3996001065820876
0	2021-01-01	5.513971745517396
0	2022-01-01	5.098022115196008
1	2018-01-01	6.5147850198640915
1	2019-01-01	7.237504177208004
1	2020-01-01	6.653735691218713
1	2021-01-01	7.096818426982313
1	2022-01-01	7.353080827262011
2	2018-01-01	4.886179086014558
2	2019-01-01	4.70149132318723
2	2020-01-01	4.787037350662963
2	2021-01-01	4.962729197408835
2	2022-01-01	4.793755908263133

Figure 1. Display table that showcases a snippet of the preprocessed data.

b. K-Means Clustering
The K-Means clustering algorithm was applied to identify inherent patterns within the dataset. The number of clusters was set to three for this analysis. The resulting model was trained, and predictions were made for each data point. The clustering analysis was visualized using a scatter plot, providing insights into the distribution of education percentages across different clusters based on gender.

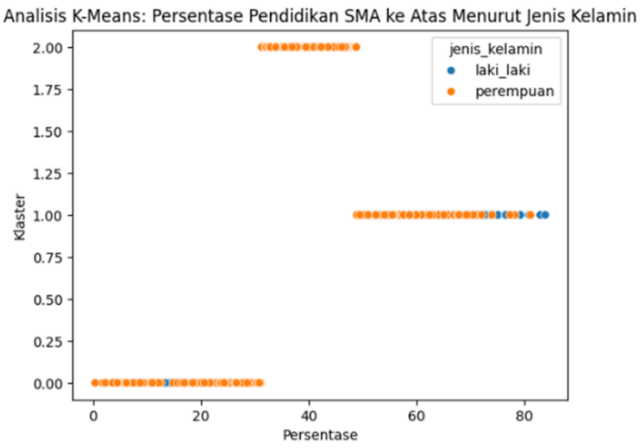

Figure 2. Display describes cluster assignment for each data point along with the value.

c. Cluster Analysis

Further insights into the clusters were gained through box plots, displaying the spread of education percentages within each cluster. The descriptive statistics for each cluster, including mean and standard deviation, were computed and presented. The distribution of education percentages in each cluster, segmented by gender, was visualized using histograms.

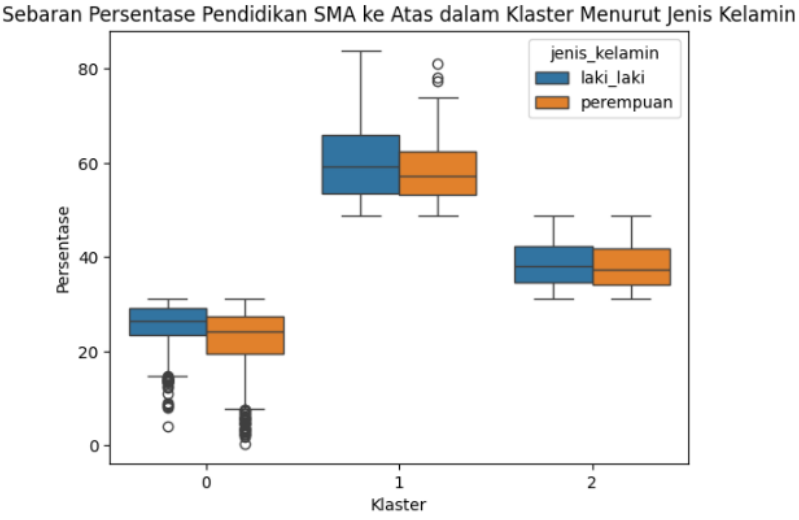

Figure 3. Display visualisasi box plot to see the distribution of percentages in each cluster.

d. Temporal Analysis

The temporal dimension was introduced by converting the "nama_tahun" column to a timestamp format. Descriptive statistics for each cluster across different years (2018-2022) were computed, providing a nuanced understanding of temporal trends. A trendline visualization illustrated changes in education percentages within each cluster over the specified time frame.

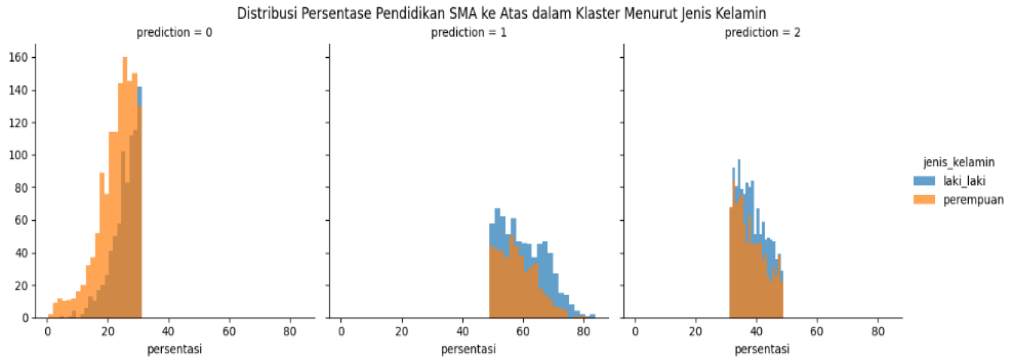

Figure 4. Display visualisasi cluster graph from 2018 to 2022.

e. Correlation and Comparative Analysis

The correlation between education percentages and cluster assignments was examined using a heatmap. Additionally, a comparison of average education percentages between clusters and genders was conducted. Bar plots effectively visualized these comparisons, offering a clear understanding of the variations.

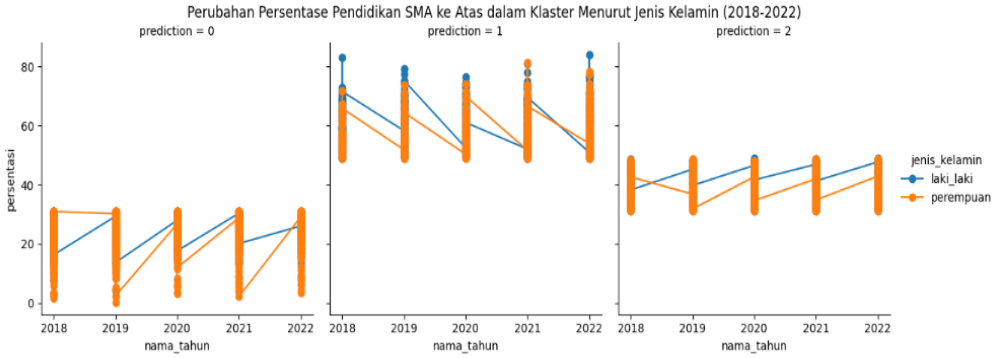

Figure 5. Display visualisasi bar plot

f. Quality Evaluation

The internal quality of clusters was assessed using the Silhouette Score. This metric provided a quantitative measure of how well-defined the clusters are within the dataset. The result, along with statistical analyses such as quartile calculations, was presented to evaluate the distribution of education percentages within each cluster and gender.

```
# Evaluasi model
evaluator = ClusteringEvaluator()
silhouette_score = evaluator.evaluate(df)
print(f"Silhouette Score: {silhouette_score}")
```

```
Silhouette Score: 0.7032226180520462
```

Figure 6. Display silhouette score

g. Top-Ranked Insights

To identify significant differences, statistical tests, such as t-tests, were applied. Top-ranked entries within each cluster, based on education percentages and gender, were isolated for in-depth examination. This allowed for a more targeted analysis of key observations and potential influential factors.

prediction	jenis_kelamin	Q1	Median	Q3
2	perempuan	34.04	37.45	41.78
0	perempuan	19.56	24.13	27.53
1	perempuan	53.23	57.28	62.37
0	laki_laki	23.34	26.54	29.08
1	laki_laki	53.6	59.13	65.83
2	laki_laki	34.59	38.07	42.49

Figure 7. Display significant differences

h. Result

In result, the comprehensive analysis utilizing K-Means clustering, visualizations, and statistical evaluations provides a nuanced understanding of the long-term trends in high school to higher education transitions. The presented results contribute valuable insights to educational policymakers and stakeholders, fostering informed decision-making in the dynamic landscape of education.

nama_wilayah	nama_tahun	jenis_kelamin	persentasi	features	prediction	rank
Cirebon	2020-01-01	laki_laki	31.2	[31.2]	0	1
Kota Banda Aceh	2022-01-01	laki_laki	83.85	[83.85]	1	1
Malinau	2022-01-01	laki_laki	48.88	[48.88]	2	1

Figure 8. Display filter data for top rank

4. Conslusion

In conclusion, our study "Long-Term High School to Higher Education Trend Analysis (2018-2022) using K-Means Methodology" navigates the intricate landscape of academic trajectories, employing the robust K-Means methodology. Adhering to the stringent formatting guidelines set by Jurnal Teknik Informatika C.I.T, the manuscript presents a meticulous and insightful analysis of educational trends over the specified period. The study's precision is reflected in its adherence to standardized writing styles, linguistic coherence, and methodical organization.

The manuscript, crafted in English with Constantia font size 10, single-spaced, and one-sided on A4 paper, ensures clarity and coherence in presentation. Incorporating graphics and tables within the 3 to 10-page range enriches the comprehensiveness of the analysis. The structured organization, including sections such as Introduction, Methods, Results and Discussion, Conclusion, and References, follows the prescribed sequence, facilitating a systematic and reader-friendly experience.

Linguistically, the manuscript upholds Standard English grammar, with a succinct title within 15 words and discoverable keywords following the abstract. Capitalized headings, full introductions of abbreviations, and the italicization of foreign or regional terms contribute to linguistic clarity. Rigorous adherence to formatting guidelines is apparent throughout the manuscript, maintaining a clean and readable presentation. The meticulous attention to detail, including the absence of bullet

points and numbering, activation of widow or orphan control, and ensuring column balance, exemplifies the commitment to precision.

In harmonizing these elements, the manuscript not only advances the field of educational analytics but also upholds the exacting standards of Jurnal Teknik Informatika C.I.T. By providing a nuanced understanding of long-term educational trends, our study contributes valuable insights for educational policymakers and stakeholders, fostering informed decision-making in the dynamic landscape of education.

References

[1] A. Verdecchia, R. Capocaccia, and T. Hakulinen, "Methods of data analysis.," *IARC Sci Publ*, no. 132, pp. 32–37, Jan. 1995, doi: 10.1201/9781482280043-4/METHODS-DATA-ANALYSIS-XAVIER-PERRIER-ALBERT-FLORI-FRAN-OIS-BONNOT.

[2] "Badan Pusat Statistik Provinsi Jawa Barat." Accessed: Jan. 14, 2024. [Online]. Available: https://jabar.bps.go.id/indicator/26/121/1/indeks-pendidikan.html

[3] "Indeks Pendidikan Berdasarkan Kabupaten/Kota di Jawa Barat." Accessed: Jan. 14, 2024. [Online]. Available: https://opendata.jabarprov.go.id/id/dataset/indeks-pendidikan-berdasarkan-kabupatenkota-di-jawa-barat

[4] "Statistik Pendidikan 2022 - Badan Pusat Statistik Indonesia." Accessed: Jan. 14, 2024. [Online]. Available: https://www.bps.go.id/id/publication/2022/11/25/a80bdf8c85bc28a4e6566661/statistik-pendidikan-2022.html

[5] "Statistik Pendidikan 2021 - Badan Pusat Statistik Indonesia." Accessed: Jan. 14, 2024. [Online]. Available: https://www.bps.go.id/id/publication/2021/11/26/d077e67ada9a93c99131bcde/statistik-pendidikan-2021.html

[6] D. Kurniadi and A. Mulyani, "Implementasi Pengembangan Student Information Terminal (S-IT) Untuk Pelayanan Akademik Mahasiswa," *Jurnal Algoritma*, vol. 13, no. 2, pp. 437–442, Dec. 2016, doi: 10.33364/ALGORITMA/V.13-2.437.

[7] A. Faiz and P. Purwati, "Koherensi Program Pertukaran Pelajar Kurikulum Merdeka Belajar Kampus Merdeka dan General Education," *EDUKATIF : JURNAL ILMU PENDIDIKAN*, vol. 3, no. 3, pp. 649–655, Apr. 2021, doi: 10.31004/EDUKATIF.V3I3.378.

[8] "ANALISIS KOHESI DAN KOHERENSI WACANA BERITA RUBRIK NASIONAL DI MAJALAH ONLINE DETIK | Jurnal Sastra Indonesia." Accessed: Jan. 14, 2024. [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jsi/article/view/7359

[9] H. Mayatopani, "Multi-criteria decision making using weighted aggregated sum product assessment in corn seed selection system," *Jurnal Teknik Informatika C.I.T Medicom*, vol. 15, no. 1, pp. 21–31, Mar. 2023, doi: 10.35335/CIT.VOL15.2023.302.PP21-31.

[10] "Jurnal Teknik Informatika C.I.T Medicom." Accessed: Jan. 14, 2024. [Online]. Available: https://www.medikom.iocspublisher.org/index.php/JTI

[11] "Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan dan Provinsi - Tabel Statistik - Badan Pusat Statistik Indonesia." Accessed: Jan. 14, 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MTk4MCMy/tingkat-penyelesaian-pendidikan-menurut-jenjang-pendidikan-dan-provinsi.html

[12] H. Wickham, "Data Analysis," pp. 189–201, 2016, doi: 10.1007/978-3-319-24277-4_9.

[13] S. Brandt, "Data analysis: Statistical and computational methods for scientists and engineers, fourth edition," *Data Analysis: Statistical and Computational Methods for Scientists and Engineers, Fourth Edition*, pp. 1–523, Jan. 2014, doi: 10.1007/978-3-319-03762-2/COVER.

[14] S. YUNIATI, "METODE PENYEDERHANAAN DATA OBSERVASI MULTIVARIAT DENGAN ANALISA FAKTOR," 1998.

[15] ----------, "Analisa Efektifitas Perbandingan Metode Thevenin Dengan MetodeMatrik Rel Impedansi Dalam Kajian Perhitungan Arus Hubungan Singkat Simetris Sistim Tenaga Listrik 12 Bus Nernais Computer," 2013, Accessed: Jan. 14, 2024. [Online]. Available: https://repositori.uma.ac.id/handle/123456789/7295

[16] H. S. Kudale, M. V Phadnis, P. J. Chittar, K. P. Zarkar, and B. K. Bodhke, "A REVIEW OF DATA ANALYSIS AND VISUALIZATION OF OLYMPICS USING PYSPARK AND DASH-PLOTLY," *International Research Journal of Modernization in Engineering Technology and Science www.irjmets.com @International*

Research Journal of Modernization in Engineering, pp. 2582–5208, 2093, Accessed: Jan. 14, 2024. [Online]. Available: www.irjmets.com

[17] C. Ding and X. He, "K-means clustering via principal component analysis," *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 225–232, 2004, doi: 10.1145/1015330.1015408.

[18] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data," *Procedia Comput Sci*, vol. 78, pp. 507–512, Jan. 2016, doi: 10.1016/J.PROCS.2016.02.095.

[19] S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," *Proceedings of the 5th International Conference on Confluence 2014: The Next Generation Information Technology Summit*, pp. 59–65, Nov. 2014, doi: 10.1109/CONFLUENCE.2014.6949256.

[20] S. Sundari *et al.*, "Analisis K-Medoids Clustering Dalam Pengelompokkan Data Imunisasi Campak Balita di Indonesia," *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, vol. 1, no. 0, pp. 687–696, Sep. 2019, doi: 10.30645/SENARIS.V1I0.75.

[21] D. R. Ningrat, D. Asih, I. Maruddani, and T. Wuryandari, "ANALISIS CLUSTER DENGAN ALGORITMA K-MEANS DAN FUZZY C-MEANS CLUSTERING UNTUK PENGELOMPOKAN DATA OBLIGASI KORPORASI," *Jurnal Gaussian*, vol. 5, no. 4, pp. 641–650, 2016, doi: 10.14710/J.GAUSS.5.4.641-650.

[22] A. Testas, "k-Means Clustering with Pandas, Scikit-Learn, and PySpark," *Distributed Machine Learning with PySpark*, pp. 395–416, 2023, doi: 10.1007/978-1-4842-9751-3_15.

[23] T. Kanungo, D. M. Mounv, R. Silverman~, N. S. Netanyahu, A. Y. Wu, and C. Piatko~, "The Analysis of a Simple k-Means Clustering Algorithm," *Hong Kong China Copyright ACM*, 2000.

[24] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithms: Analysis and implementation," *IEEE Trans Pattern Anal Mach Intell*, vol. 24, no. 7, pp. 881–892, Jul. 2002, doi: 10.1109/TPAMI.2002.1017616.

[25] V. Veronica and R. Rodhiah, "PENGARUH PRIVACY, SOCIAL INFLUENCE TERHADAP ONLINE PURCHASE INTENTION: TRUST SEBAGAI VARIABEL MEDIASI," *Jurnal Muara Ilmu Ekonomi dan Bisnis*, vol. 5, no. 2, pp. 235–246, Jul. 2021, doi: 10.24912/JMIEB.V5I2.9657.

[26] "POPULASI-SAMPEL, TEKNIK SAMPLING & BIAS DALAM PENELITIAN - I Ketut Swarjana, S.K.M., M.P.H., Dr.PH - Google Buku." Accessed: Jan. 14, 2024. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=87J3EAAAQBAJ&oi=fnd&pg=PA1&dq=pertimbangan+dalam+sebuah+penelitian&ots=LODEsaoKFv&sig=qi53X8E5x-G2caKB2lafJtgZEaY&redir_esc=y#v=onepage&q=pertimbangan%20dalam%20sebuah%20penelitian&f=false