|  |  |  |
| --- | --- | --- |
| Long-Term High School to Higher Education Trend Analysis (2018-2022) using K-Means Methodology | | |
|  | | |
| Mohamad Farizal Arifin1, Author2, | | |
| 1)Teknik Informatika, Pelita Bangsa University, West Java, Indonesia | | |
|  |  |  |
| Article Info |  | Abstract |
| *Article history*  Received : diisi oleh editor  Revised : diisi oleh editor  Accepted : diisi oleh editor |  | *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.* |
|  |  |
| *Kata Kunci:*  *Higher Education Analysis;*  *K-Means Methodology;*  *Long-Term Trends;*  *PySpark Analysis;* |  |
| ***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*](https://creativecommons.org/licenses/by-nc/4.0/) *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. 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.

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.

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.

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.

1. **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.

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.

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.

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.

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.

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.

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.

1. **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.

1. 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.

A screenshot of a computer code

Description automatically generated

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

1. 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.

A graph with numbers and dots

Description automatically generated

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

1. 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.

A graph of a bar chart

Description automatically generated with medium confidence

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

1. 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.

A graph with a blue and orange line

Description automatically generated

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

1. 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.

A graph with blue and orange lines

Description automatically generated

Figure 5. Display visualisasi bar plot

1. 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.

A screenshot of a computer code

Description automatically generated

Figure 6. Display silhouette score

1. 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.

A screenshot of a computer

Description automatically generated

Figure 7. Display significant differences

1. 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.

A white background with black text

Description automatically generated

Figure 8. Display filter data for top rank

1. **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**

(one blank single space line, 10 pt)

References must be written in the order they appear in the text, using Arabic numerals in square brackets, as shown in the example. References of at least 25 journal articles in the last 5 years, 80% of which must be sourced from international reputable journals and try the reference has a URL/Link. References must use a Reference Management application such as Mendeley or similar application and use IEEE style and 9pt font size.

[1] R. E. Caraka *et al.*, “Impact of COVID-19 large scale restriction on environment and economy in Indonesia,” *Glob. J. Environ. Sci. Manag.*, vol. 6, no. Special Issue (Covid-19), pp. 65–84, 2020.

[2] M. Machida *et al.*, “Incorrect use of face masks during the current COVID-19 pandemic among the general public in Japan,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, p. 6484, 2020.

[3] W. Bourequat and H. Mourad, “Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine,” *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 36–44, 2021, doi: 10.25008/ijadis.v2i1.1216.

[4] P. Chen, R. Tong, B. Yu, and Y. Wang, “Reliable shortest path finding in stochastic time-dependent road network with spatial-temporal link correlations: A case study from Beijing,” *Expert Syst. Appl.*, vol. 147, 2020, doi: 10.1016/j.eswa.2020.113192.

[5] D.-H. Oh, J. Ahn, S. Lee, and H. Choi, “Measuring technical inefficiency and CO2 shadow price of Korean fossil-fuel generation companies using deterministic and stochastic approaches,” *Energy Environ.*, p. 0958305X20932547, 2020.

[6] Y. J. Han *et al.*, “Treatment of severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and coronavirus disease 2019 (COVID-19): a systematic review of in vitro, in vivo, and clinical trials,” *Theranostics*, vol. 11, no. 3, p. 1207, 2021.

[7] A. Hocine, M. S. Guellil, E. Dogan, S. Ghouali, and N. Kouaissah, “A fuzzy goal programming with interval target model and its application to the decision problem of renewable energy planning,” *Environ. Ecol. Stat.*, vol. 27, no. 3, pp. 527–547, 2020.

[8] Z. Zhai, J. Fernán Martínez, N. Lucas Martínez, and V. Hernández Díaz, “Applying case-based reasoning and a learning-based adaptation strategy to irrigation scheduling in grape farming,” *Comput. Electron. Agric.*, vol. 178, p. 105741, Nov. 2020, doi: 10.1016/j.compag.2020.105741.

[9] A. M. Catai, C. M. Pastre, M. F. de Godoy, E. da Silva, A. C. de Medeiros Takahashi, and L. C. M. Vanderlei, “Heart rate variability: are you using it properly? Standardisation checklist of procedures,” *Brazilian J. Phys. Ther.*, vol. 24, no. 2, pp. 91–102, 2020.

[10] B. M. F. Viana, L. T. Pereira, C. F. M. Toledo, S. R. dos Santos, and S. M. D. M. Maia, “Feasible–Infeasible Two-Population Genetic Algorithm to evolve dungeon levels with dependencies in barrier mechanics,” *Appl. Soft Comput.*, vol. 119, p. 108586, 2022.

[11] S. R. Jayashree, G. Dias, J. J. Andrew, S. Saha, F. Maurel, and S. Ferrari, “Multimodal Web Page Segmentation Using Self-organized Multi-objective Clustering,” *ACM Trans. Inf. Syst.*, vol. 40, no. 3, pp. 1–49, 2022.

[12] A. Mughaid *et al.*, “An intelligent cybersecurity system for detecting fake news in social media websites,” *Soft Comput.*, pp. 1–15, 2022.

[13] O. Nahli, A. Sanna, M. Bandini, and F. Boschetti, “Commerce Numérique: traffic signals for the crossroads between cultures,” *Int. J. Inf. Sci. Technol.*, vol. 5, no. 3, pp. 36–45, 2022.

[14] M. Kumar, J. Aggarwal, A. Rani, T. Stephan, A. Shankar, and S. Mirjalili, “Secure video communication using firefly optimization and visual cryptography,” *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 2997–3017, 2022.

[15] M. K. Gupta *et al.*, “Detection and localization for watermarking technique using LSB encryption for DICOM Image,” *J. Discret. Math. Sci. Cryptogr.*, pp. 1–12, 2022.

[16] M. Salehi, G. De Borger, D. Hughes, and B. Crispo, “NemesisGuard: Mitigating interrupt latency side channel attacks with static binary rewriting,” *Comput. Networks*, p. 108744, 2022.

[17] G. Palumbo *et al.*, “A Structured Approach for the Design and Manufacturing of Titanium Cranial Prostheses via Sheet Metal Forming,” *Metals (Basel).*, vol. 12, no. 2, p. 293, 2022.

[18] J.-S. Park, J.-H. Kim, Y.-C. Jeong, H.-T. Kim, S.-K. Kim, and J.-M. Lee, “Effect of Corrugated Sheet Diameter on Structural Behavior under Cryogenic Temperature and Hydrodynamic Load,” *Metals (Basel).*, vol. 12, no. 3, p. 521, 2022.

[19] P. Stachelek, L. MacKenzie, D. Parker, and R. Pal, “Circularly polarised luminescence laser scanning confocal microscopy to study live cell chiral molecular interactions,” *Nat. Commun.*, vol. 13, no. 1, pp. 1–8, 2022.

[20] D. Mezad-Koursh *et al.*, “Response to: Comment on:‘Peripapillary hyperreflective ovoid mass-like structures—a novel entity as frequent cause of pseudopapilloedema in children,’” *Eye*, p. 1, 2022.

[21] P. Chen, L. Xu, and Z. Li, “Modeling methods of self-centering energy dissipation braces using OpenSees,” *J. Constr. Steel Res.*, vol. 193, p. 107267, 2022.

[22] D. Stanford, Z. Yang, and S. Yao, “Subleading weingartens,” *J. High Energy Phys.*, vol. 2022, no. 2, pp. 1–50, 2022.

[23] B. Joksimoski *et al.*, “Technological solutions for sign language recognition: a scoping review of research trends, challenges, and opportunities,” *IEEE Access*, 2022.

[24] J. K. Lê and T. Schmid, “The practice of innovating research methods,” *Organ. Res. Methods*, vol. 25, no. 2, pp. 308–336, 2022.

[25] J. Miller and R. Ulrich, “Optimizing Research Output: How Can Psychological Research Methods Be Improved?,” *Annu. Rev. Psychol.*, vol. 73, pp. 691–718, 2022.

[26] S. Hosseinpour, A. Gaudin, and O. A. Peters, “A critical analysis of research methods and experimental models to study biocompatibility of endodontic materials,” *Int. Endod. J.*, 2022.