SYSTEMIC: Information System and Informatics Journal

ISSN: 2460-8092, 2548-6551 (e)

Vol 7 No 2 - Desember 2021

Educational Data Clustering Menggunakan K-Means Pada Seleksi Nasional Peserta Didik Baru Madrasah Aliyah Negeri Unggulan

Noor Wahyudi[1], Yunita Ardilla[2], Nanik Puji Hastuti[3]

[1,2)] Universitas Islam Negeri Sunan Ampel Surabaya
[3)] Direktorat Kurikulum, Sarana, Kelembagaan, dan Kesiswaan Madrasah Kementerian Agama

n.wahyudi@uinsby.ac.id[1], yunita.ardilla@uinsby.ac.id[2], nanikpujihastuti@gmail.com[3]

Kata Kunci

EDC, K-Means, Penerimaan Siswa Baru, Madrasah.

Abstrak

Seleksi Nasional Peserta Didik Baru (SNPDB) Madrasah Aliyah Negeri Unggulan dikelola oleh Direktorat Kurikulum, Sarana, Kelembagaan dan Kesiswaan Madrasah. Menjadi penting bagi Direktorat dan Madrasah untuk menggali pola dan pengetahuan dari data seleksi dalam penyusunan kebijakan dan program pada MAN Unggulan.. MAN Insan Cendekia (MAN-IC) merupakan madrasah aliyah negeri unggulan yang paling diminati. Educational Data Clustering (EDC) merupakan metode data mining yang dimplementasikan dibidang pendidikan. K-means diterapkan untuk mengelompokkan Siswa berdasarkan hasil tes potensi belajar dan potensi akademik yang akan digunakan untuk penyusunan program dan kebijakan seleksi siswa pada MAN-IC. Hasil terbaik dari ekperimen yang diuji dengan Silhouette membagi data menjadi 2 klaster sangat baik dan baik. Nilai Silhouete menunjukkan struktur klaster pada predikat medium. Hasil pengelompokan menyajikan sebaran klaster di 23 MAN-IC, sebaran profil kepribadian dari calon peserta didik, serta rekomendasi untuk pelaksanaan tes di Madrasah.

Keywords

EDC, K-Means, Student Admission, Madrasah.

Abstract

The National Students admissions (SNPDB) for Madrasah Aliyah is managed by the Directorate of Madrasah Curriculum, Facilities, Institutions and Student Affairs. It is essential for the Directorate and Madrasah to explore patterns and knowledge from admission data in formulating policies and programs from to MAN. Educational Data Clustering (EDC) is a data mining method that is implemented in the education area. K-means is applied to group students based on the results of learning potential and academic potential tests that will be used for development program and student admission policies at MAN-IC. The best results from the experiments tested with Silhouette dividing the data into 2 clusters are excellent and good. The Silhouete value indicates the cluster structure in the medium predicate.. The results present the distribution of clusters in 23 MAN-IC, distribution of personality profiles of prospective students, as well as recommendations for conducting tests in Madrasah.

1. Pendahuluan

Saat ini, animo masyarakat terhadap Madrasah Aliyah Negeri (MAN) Unggulan semakin tinggi. Hal ini bisa dilihat dari perkembangan pendaftar yang selalu mengalami kenaikan setiap tahun. Tercatat sebanyak 17.422 pendaftar pada tahun 2021 meningkat signifikan dari tahun sebelumnya. Berangkat Jumlah dan lokasi yang tersebar, mulai tahun 2019 seluruh MAN Unggulan se-Indonesia merintis penyelenggaraan Seleksi Nasional Peserta Didik Baru (SNPDB) secara online yang di kelola oleh Direktorat Kurikulum, Sarana, Kelembagaan dan Kesiswaan Madrasah (KSKK).

SNPDB secara reguler diagendakan setiap tahun pada awal tahun secara online dengan mekanisme *Computer Based Test* (CBT). SNPDB dilaksankan secara serentak di seluruh MAN Unggulan di Indonesia. Dalam Seleksi ini diujikan tiga jenis tes yaitu tes Akademik, tes potensi belajar dan tes kepribadian. Tes Potensi belajar terdiri dari *Verbal ability* (kemampuan verbal), *Numerical reasoning* (penalaran numerik/angka), *Analytical reasoning* (penalaran analisis). Tes potensi belajar (*learning potential test*) digunakan untuk mengukur tingkat kesiapan belajar dan

DOI : 10.29080/systemic.v7i2.1768

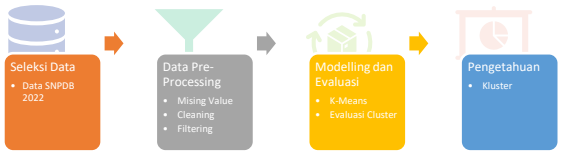
performansi akademik siswa[1]. Tes akademik terdiri dari mata uji Matematika, Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS), Bahasa Inggris, Bahasa Arab dan Pendidikan Agama Islam. Tes Akademik bertujuan untuk mengukur pengetahuan dan kemampuan siswa pada subyek mata pelajaran. Tes Kepribadiaan atau personality bertujuan untuk mengetahui profil calon peserta didik bereaksi dan berinteraksi dengan lingkungan ataupun individu lainnya. Ada 12 profil kepribadian yang dihasilkan dari tes kepribadian yaitu: *anxiety, balanced, causal, control, courage, efficacy, egoism, emotional control, empathy, optimism, regulation, sensitiv*. Hasil dari ketiga test tersebut akan menentukan kelulusan dari calon peserta didik yang mendaftar pada madrasah yang dituju. Data hasil SNPDB tahun 2021 berpotensi memiliki *insight* yang jelas mengenai tingkah laku dari siswa, potensi prestasi, serta deteksi dini faktor yang mempengaruhi pembelajaran. Menjadi penting bagi Direktorat KSKK dan Madrasah untuk menggali pola dan pengetahuan dari data selesi untuk digunakan dalam penyusunan kebijakan, strategi atau program dari MAN Unggulan untuk peningkatan mutu pembelajaran. MAN Insan Cendekia (MAN-IC) merupakan madrasah aliyah negeri unggulan yang paling diminati. Tercatat ada 23 MAN-IC di Indonesia yang tersebar di seluruh Indonesia dari mulai Aceh hingga Sorong. Beberapa penelitian dari dataset SNPDB atau MAN-IC telah dilakukan dengan menggunakan pendekatan statistik seperti[2] yang meneliti tentang pengaruh personality dengan capaian akademik, [3] pengaruh tes potensi belajar terhadap capaian akademik dan [4] yang meneliti tentang potensi kesalahan dalam pengukuran untuk seleksi penerimaan peserta didik baru. Dalam situasi tertentu pendekatan statistik tidak dapat digunakan berkenaan dengan permasalahan jumlah data yang semakin besar, *missing value* dan *imbalanced data*. Salah satu area riset tentang penambangan data (data mining) di bidang Pendidikan biasa disebut *Educational Data Mining* (EDM)[5].

EDM merupakan area riset interdisipliner yang memadukan bidang pendidikan dan ilmu komputer. EDM memanfaatkan metode dan teknik data mining pada dataset di bidang pendidikan yang akan diekstrak menjadi insight atau pengetahuan yang berguna dan bermakna. EDM dikategorikan dalam 6 bagian besar yaitu: 1) *Distillation of data for human judgment* 2) *Prediction methods* 3) *Relationship mining methods* 4) *Structure discovery methods* 5) Discovery with models 6) Miscellaneous other methods[6]. *Structure discovery methods* termasuk dalam model pembelajaran unsupervised untuk mencari struktur dalam sebuah dataset. Salah satu metode dari *Structure discovery methods* adalah *clustering* atau bisa disebut *Educational Data Clustering* (EDC). EDC umumnya digunakan untuk mengelompokkan peserta didik dengan kemampuan akademik atau karakteristik yang serupa. Teknik *Clustering* yang banyak digunakan dan umum digunakan adalah K-Means karena ringkas dan mempunyai performa yang baik[7]. Algoritma K-means cukup populer diimplementasikan dalam dalam EDC seperti penelitian yang memprediksi performa akademik[8][9], dan pengelompokkan sekolah di Kalimantan Timur berdasarkan mutu sekolah[10].

Dalam Penelitian ini K-means digunakan untuk mengelompokkan siswa berdasarkan hasil tes potensi belajar dan potensi akademik yang akan digunakan untuk penyusunan program dan kebijakan seleksi siswa pada tahun berikutnya bagi MAN unggulan, serta kebijakan lain yang sesuai arah peningkatan kualitas MAN unggulan khususnya MAN Insan Cendekia.

2. Metode Penelitian

Tahapan yang ditempuh pada penelitian ini ditunjukkan pada gambar 1 dengan urutan sebagai berikut: Seleksi Data, Data Pre-Processing, modelling dan evaluasi, dan pengetahuan dan interpretasi.



Gambar 1. Tahapan Penelitian

2.1 Seleksi Data

Penelitian ini menggunakan data hasil seleksi Peserta Didik Baru tahun 2021 yang terkumpul 17422 baris data dengan atribut sebanyak 90 yang selanjutnya disebut sebagai dataset. Kemudian dari 90 atribut pada dataset dipilih 2 atribut yaitu nilai tes potensi belajar dan nilai tes akademik.

2.2 Pemrosesan Awal

Beberapa teknik pemrosesan awal data diterapkan dalam beberapa langkah agar data bisa diproses dengan metode clusteing menggunakan K-means. Langkah pertama dilakukan dengan menghapus atribut yang tidak digunakan. Langkah berikutnya menyaring baris dengan kriteria data siswa yang tidak diterima akan dibersihkan dari dataset yang akan digunakan. Tahap ketiga pemeriksaan *missing value* dari atribut nilai tes potensi akademik dan tes potensi belajar. Baris yang kosong atau bernilai nol akan dihapus dari atribut tes potensi belajar dan tes akademik. Pemrosesan awal ini menghasilkan 2974 baris data yang kemudian menjadi masukan pada algoritma K-means.

2.3 Modelling dan Evaluasi

Modelling merupakan proses mengekstrak atau menemukan pola atau informasi yang menarik dari data dengan menggunakan teknik, metode atau algoritma tertentu. Dalam penelitian ini algoritma K-Means diaplikasikan untuk menemukan pola atau pengetahuan dari data seleksi yang telah menjalani pemrosesan awal. Algoritma K-means dijalankan sesuai alur berikut:

1. Dimulai dengan pemilihan *k* titik secara acak pusat klaster.
2. Setiap titik dalam dataset ditandai dan masuk ke klaster terdekat berdasarkan jarak Euclidean antara setiap titik dengan pusat klaster.
3. Pusat klaster dihitung ulang sesuai nilai rata-rata dari titik-titik dalam klaster tersebut.
4. Ulangi Langkah 2 dan 3 sampai klaster-klaster terjadi konvergensi. Konvergensi berarti jika output dari pengulangan Langkah 2 dan 3 tidak membuat perbedaan signifikan pada klaster atau tidak ada perubahan dalam klaster.

Keluaran dari algoritma K-means berupa klaster akan dievaluasi menggunakan *Silhouette*. *Silhouette* merupakan salah satu metode evaluasi untuk teknik *clustering* yang mengukur kualitas dari klaster yang dihitung dari jarak dari setiap data dengan pusat klaster [11]. Setiap data dihitung jaraknya dengan pusat klaster. Evaluasi kluster menggunakan *Silhouette* dimasudkan untuk melihat seberapa dekat data dari pusat klaster dan jarak dengan pusat klaster lain. Hasil dari Silhouette mempunya interval -1 sampau dengan 1. Nilai -1 menunjukkan hasil clustering yang jelek dan nilai 1 menunjuukan hasil clustering yang baik[12]. Dengan menggunakan tabel Kaufman dan Rousseeuw akan diketahui predikat Struktur dari klaster yang dihasilkan[13]. Nilai Silhouette dapat diperoleh dengan menggunakan persamaan 1 dan 2.

$$sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

$$Sil = \frac{1}{n}\sum_{i=1}^{n} sil(i) \quad (2)$$

2.4 Pengetahuan

Modeling dan evaluasi akan menghasilkan pola atau klaster yang terbaik berdasarkan evaluasi yang menggunakan *Silhouette*. Dari pola yang didapatkan kemudian akan diidentifkasi dan diinterpretasikan sehingga bisa menjadi pengetahuan atau *insight* yang dapat dipahami dan bermanfaat bagi pihak yang berkepentingan.

3. Hasil Dan Pembahasan

3.1 Seleksi dan pemrosessan awal Data

Dari data mentah hasil seleksi penerimaan peserta didik baru yang diperoleh dilakukan pemilihan atribut yang akan dianalisis untuk menemukan pola dan pengetahuan mengenai siswa yang diterima di Madrasah. Dari 90 atribut, pada penelitian ini dipilih dua atribut yakni nilai tes potensi belajar dan nilai tes potensi akademik. Dua atribut ini kemudian masuk dalan tahap pemrosesan awal pembersihan data, penanganan missing value dan filtering supaya data siap untuk masuk proses clustering. Hasil dari pemrosesan awal sejumlah 3349 baris data peserta didik yang lulus seleksi dan masuk di 23 MAN-IC.

Tabel 1 Atibut Tes Potensi Belajar dan Tes Akademik

Variabel	Nilai minimal	Nilai maksimal	Mean
TPB	328,49	745,02	552,43
TA	351,64	748,40	547,74

3.2 Modelling dan evaluasi

Dari tahap seleksi dan pemrosesan data kemudian masuk pada modelling dan evaluasi. Tahap ini mengimplementasikan algoritma K-means pada data yang akan menghasilkan klaster. Pada penelitian ini dilakukan percobaan menggunakan k=2 hingga k=5. Hasil proses klaster kemudian diukur menggunakan Silhouette. Hasil pengukuran dari Silhouette dengan k=2 sampai k=5 ditunjukkan pada tabel 2.

Tabel 2 Nilai evaluasi klaster dengan Silhouette

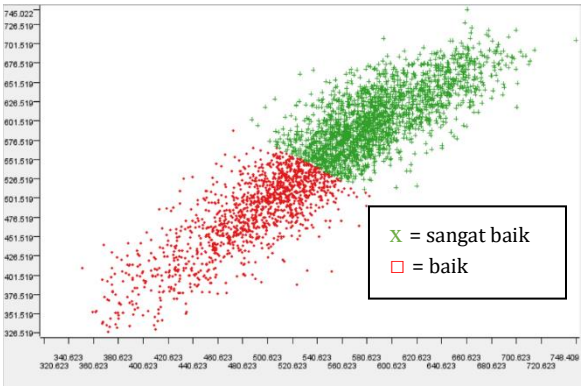
Klaster	Silhouette Coeficient
2	0,522
3	0,460
4	0,426
5	0,405

Tabel 2 menunjukkan eksperimen menggunakan *k*=2 menghasilkan nilai yang paling tinggi yaitu 0,522 dibandingkan dengan eksperimen dengan menggunakan nilai *k* lainnya. Menurut tabel Kaufmann k=2 mempunyai struktur yang baik (*medium Structure*). Hasil dari clustering menggunakan k=2 ditunjukkan pada tabel 3 dengan klaster 0 atau bisa disebut sangat baik (SB) dengan angka 604,67 untuk tes potensi belajar dan angka 590,22 untuk tes akademik, sedangkan klaster 1 atau bisa disebut baik dengan angka 479,95 untuk tes potensi belajar dan 487,79 untuk tes akademik. Pada gambar 2 memperlihatkan klaster sangat baik berada dibagian atas dengan notas x berwarna hijau, sedangkan klaster baik dengan notasi □ berwarna merah.

Tabel 3 Hasil klaster k=2

	Sangat baik	Baik
TPB	604,67	479,95

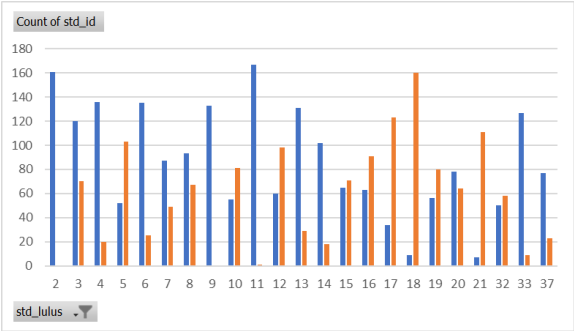
TA	590,22	487,79



Gambar 2. Klaster Peserta didik MAN-IC

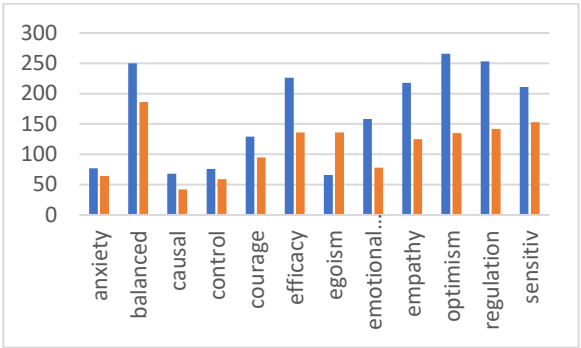
3.3 Pembahasan

Hasil eskperimen menunjukkan ada 2 klaster yaitu sangat baik dan baik. Gambar 3 menunjukkan distribusi klaster pada 23 MAN-IC. Klaster sangat baik ditunjukkan dengan warna biru dan klaster baik dengan warna orange. Tampak 13 MAN-IC mempunyai klaster sangat baik yang lebih tinggi. Lebih jauh pada MAN-IC dengan id 2 (MAN-IC Serpong), 9 (MAN-IC Pekalongan) dan 11 (MAN-IC Padang Pariaman) didominasi penuh dengan klaster sangat baik. MAN-IC yang masuk dalam kelompok ini tersebar di Jawa, Sumatera, Kalimantan dan Sulawesi. Pada kelompok lainya dengan klaster baik yang lebih tinggi dari sangat baik sejumlah 10 MAN-IC tersebar di Sumatera, Sulawesi dan wilayah Indonesia Timur. Data yang telah terklaster. Sebaran ini merupakan informasi berharga dalam penyusunan dan pengembangan MAN-IC ke depan khususnya pada MAN-IC yang peserta didiknya didominasi pada klaster baik.



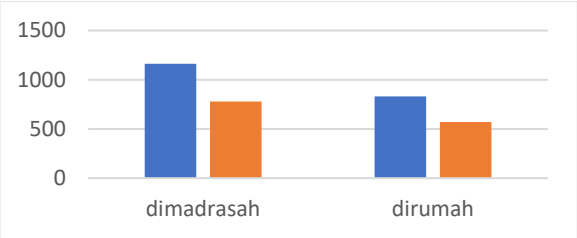
Gambar 3. Distribusi klaster di 23 MAN-IC

Dalam seleksi peserta didik, selain mencari calon peserta didik dengan nilai tes akademik dan tes potensi belajar yang tinggi, diperhatikan pula faktor kepribadian (*personality*). Sebagai salah satu dari tes yang diujikan, tes kepribadian akan menampilkan profil calon peserta didik sesuai hasil jawaban. Gambar 4 menampilkan sebaran kepribadian dari peserta didik. Ada 12 profil kepribadian calon peserta didik yang merupakan pegembangan dari profil kepribadian dari tes kepribadian tahun sebelumnya[2]. Profil kepribadian yang bersifat positif seperti *balanced, Optimis, Empathy, Efficacy, regulation* dan *courage* lebih diutamakan dari kepribadian yang bersifat negatif seperti *anxiety, egoism dan sensitiv*. Akan tetapi dalam beberapa kasus karena tidak tersedianya pilihan, maka beberapa kepribadian yang negatif tetap bisa masuk dalam seleksi. Meskipun demikian, pihak madrasah sudah mempunyai data profil kepribadian yang menjadi *baseline* dalam melakukan strategi dalam proses pembelajaran yang sesuai dengan kepribadian peserta didik.



Gambar 4. Distribusi klaster dengan kepribadian

Selain sebaran klaster pada madrasah dan kepribadian calon peserta didik, perlu juga untuk dibahas mengenai tempat pelaksanaan tes. Pada awal tahun 2021 pemerintah masih memberlakukan pembatasan sosial berskala besar (PSBB) karena pandemi Covid-19. Status ini menyebabkan kondisi dari setiap daerah yang menyelenggarakan tes seleksi mengalami keterbatasan karena larangan pengumpulan massa dan pelaksanaan protokol kesehatan. Oleh karena itu panita penyelenggara memberikan keleluasaan bagi calon peserta didik untuk menjalani tes dari rumah sebagai alternatif apabila tidak bisa hadir di madrasah karena kendala kondisi pandemi. Gambar 5 menampilkan sebaran jumlah peserta didik yang melaksanakan tes di madrasah dan di rumah. Jumlah peserta didik yang mengerjakan tes di madrasah lebih tinggi dari pada peserta didik yang mengerjakan dirumah. Klaster sangat baik dan baik menunjukkan bahwa jumlah siswa yang mengerjakan di madrasah lebih tinggi dari siswa yang mengerjakan dirumah. Beberapa temuan dilapangan yang dihimpun dati panitia lokal maupun pusat, kondisi geografis yang mempengaruhi koneksi internet, standar peralatan kompter, serta kesulitan penanganan kendala aplikasi menjadi kendala yang paling banyak diterima. Temuan ini mengarakan rekomendasi pelaksanaan SNPDB pada tahun berikutnya untuk calon peserta seleksi wajib mengerjakan tes di madrasah baik itu MAN-IC tujuan ataupun madrasah lain yang disiapkan oleh panitia sebagai tempat tes.



Gambar 5. Cluster dengan tempat tes

4. Kesimpulan

EDC telah diimplementasikan pada dataset SNPDB MAN Unggulan tahun 2021 menggunakan algoritma K-means. Ekperimen ini menghasilkan 2 buah klaster dengan predikat sangat baik dan baik. Hasil pengujian menunjukkan K-means dengan *k*=2 menghasilkan nilai Silhouette paling tinggi dengan nilai 0,55 yang memperoleh predikat struktur medium berdasarkan tabel kaufman.

Daftar Pustaka

[1] H. Klimusová and P. Květon, "Psychometric Properties of the Learning Potential Test," *Procedia - Soc. Behav. Sci.*, vol. 217, pp. 652–656, 2016, doi: 10.1016/j.sbspro.2016.02.089.

[2] A. Muhid, A. Ridho, A. Yusuf, N. Wahyudi, Z. Ulya, and A. H. Asyhar, "Big Five Personality Test for State Islamic Senior High School Students in Indonesia," *Int. J. Instr.*, vol. 14, no. 2, pp. 483–500, 2021.

[3] A. Muhid, A. Yusuf, Kusaeri, D. C. R. Novitasari, A. H. Asyhar, and A. Ridho, "Determining scholastic aptitude test as predictors of academic achievement on students of islamic school in indonesia," *New Educ. Rev.*, vol. 61, pp. 211–221, 2020, doi: 10.15804/tner.2020.61.3.17.

[4] A. Yusuf, K. Kusaeri, A. Hidayatullah, D. C. R. Novitasari, and A. H. Asyhar, "Detection of potential errors in measurement results of madrasa admission instruments in Indonesia," *Int. J. Eval. Res. Educ.*, vol. 10, no. 4, pp. 1334–1343, 2021, doi: 10.11591/IJERE.V10I4.21412.

[5] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, vol. 5, no. c, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[6] A. Aleem and M. M. Gore, "Educational data mining methods: A survey," in *Proceedings - 2020 IEEE 9th International Conference on Communication Systems and Network Technologies, CSNT 2020*, 2020, pp. 182–188, doi: 10.1109/CSNT48778.2020.9115734.

[7] E. Umargono, J. E. Suseno, and V. G. S. K., "K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median," no. Conrist 2019, pp. 234–240, 2020, doi: 10.5220/0009908402340240.

[8] Z. M. Ali, N. H. Hassoon, W. S. Ahmed, and H. N. Abed, "The Application of Data Mining for Predicting Academic Performance Using K-means Clustering and Naïve Bayes Classification," *Int. J. Psychosoc. Rehabil.*, vol. 24, no. 03, pp. 2143–2151, 2020, doi: 10.37200/ijpr/v24i3/pr200962.

[9] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[10] M. Yoalifa, M. Wati, N. Puspitasari, and U. Hairah, "Analisa Mutu Sekolah Pada Provinsi Kalimantan Timur," *SAINS, Apl. KOMPUTASI DAN Teknol. Inf.*, vol. 3, no. 2, pp. 53–60, 2021, doi: http://dx.doi.org/10.30872/jsakti.v3i2.4407.

[11] A. H. Asyhar *et al.*, "Graph Degree Linkage Clustering for Identify Student's Performance on Kompetisi Sains Madrasah in Indonesia," in *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020*, 2021, pp. 211–220.

[12] Kusaeri *et al.*, "Stepwise Iterative Maximum Likelihood Clustering Based on Kompetisi Sains Madrasah' Scores for Identifying Quality of Junior High School Grading Distribution," in *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020*, 2021, pp. 221–229.

[13] V. T. P. Swindiarto, R. Sarno, and D. C. R. Novitasari, "Integration of Fuzzy C-Means Clustering and TOPSIS (FCM-TOPSIS) with Silhouette Analysis for Multi Criteria Parameter Data," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 463–468, 2018, doi: 10.1109/ISEMANTIC.2018.8549844.