IMPLEMENTASI DATA MINING MENGGUNAKAN METODE K-MEANS CLUSTERING UNTUK MENENTUKAN STATUS KEMATIAN BAYI DI JAWA BARAT

Anisa Pujianti, Mulyawan
Program Studi Manajemen Informatika D3, Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon, Indonesai
Pujiantianisa212@gmail.com

ABSTRAK

Angka Kematian Bayi (AKB) digunakan sebagai indikator untuk mengukur tingkat kesehatan di suatu daerah karena bayi adalah anak-anak berusia antara 0 dan 12 bulan, yang sangat rentan terhadap kesehatan dan kesejahteraan yang buruk. Provinsi Jawa Barat perlu mendapat perhatian khusus untuk menekan tingginya AKB karena memiliki jumlah kasus kematian bayi yang cukup besar—menempati urutan ketiga di Indonesia dalam hal ini. Teknik K-Means Clustering membagi data menjadi dua kelompok yang hampir identik satu sama lain dan berbeda satu sama lain. Tujuan dari penelitian ini adalah untuk mengevaluasi efektivitas metode K-Means Clustering dalam klasterisasi kematian bayi berbasis desa-desa di Jawa Barat. Berdasarkan informasi dari desa dan volume kejadian, pengukuran ini dilihat. Data penelitian ini berasal dari data.jabarprov.go.id dan berdasarkan statistik kematian bayi di Jawa Barat. 27 dari 601 data yang dikumpulkan dari 2019 hingga 2021, berjumlah 601, berasal dari Jawa Barat.

Kata kunci: *Kematian, Bayi, K-Means, KDD, Data Mining, Cluster.*

1. PENDAHULUAN

Angka Kematian Bayi adalah metrik penting untuk menilai tingkat kesehatan populasi karena menangkap kondisi kesehatan populasi secara umum.

Jumlah ini sangat responsif terhadap pergeseran kondisi kesehatan dan kesejahteraan secara umum. Data Bank Dunia menunjukkan bahwa pada tahun 2021, Indonesia memiliki angka kematian bayi baru lahir sebesar 11,7 per 1000 kelahiran hidup. Provinsi Jawa Barat memiliki angka kematian bayi yang relatif tinggi dan tertinggi ketiga di Indonesia setelah Jawa Tengah dan Jawa Timur dengan 2.851 kasus. Di seluruh Indonesia, AKB pada 2019 sebesar 5,53 per 1000 kelahiran hidup, atau 26.395 kasus. Provinsi Jawa Barat merupakan salah satu provinsi dengan jumlah penduduk terbanyak, itulah sebabnya hal ini terjadi.

Angka kematian bayi Provinsi Jawa Barat cukup tinggi dan tertinggi ketiga di Indonesia setelah Jawa Tengah dan Jawa Timur dengan 2.851 kasus, meskipun faktanya masih di bawah rata-rata nasional. AKB Indonesia pada 2019 sebesar 5,53 per 1000 kelahiran hidup, atau 26.395 kasus. Provinsi Jawa Barat merupakan salah satu provinsi dengan jumlah penduduk terbanyak, itulah sebabnya hal ini terjadi.

Pada penelitian ini, bayi dikelompokkan berdasarkan jumlah kematian bayi baru lahir di Jawa Barat menggunakan pendekatan klasterisasi K-Means. Untuk menyederhanakan klasifikasi angka kematian berdasarkan isu-isu terkini, diperlukan pengolahan data menjadi sebuah clustering.

Berdasarkan dataset yang berkaitan dengan masalah kesehatan yang disediakan oleh Dinas Pemberdayaan Masyarakat dan Desa, yang diambil pada tahun 2019-2021 dengan total data 15936, data yang mendukung signifikansi penelitian ini dilakukan. Penulis mengambil 601 data untuk alasan apa pun, dan ada 16 atribut

Nurul Qisthi, Yudhie Andriyana, dan Yuyun Hidayat menggunakan teknik Quantile Regression Analysis untuk melakukan penelitian angka kematian bayi baru lahir di Jawa Barat. Regresi kuantil digunakan untuk mengatasi outlier yang mengungkap kelemahan model regresi rata-rata dalam penelitian ini. Dalam penelitian ini, tiga tingkat kuantil—25%, 50%, dan 75%—akan digunakan dalam analisis regresi kuantil dengan model. Empat klasifikasi level AKB akan dibuat dari model regresi kuantil, yaitu low, medium, high, dan extremely high. Enam kabupaten/kota di Jawa Barat telah tergolong memiliki tingkat AKB rendah dan sedang, delapan telah diklasifikasikan memiliki tingkat AKB tinggi, dan tujuh telah diklasifikasikan memiliki tingkat AKB sangat tinggi, menurut klasifikasi AKB di Jawa Barat pada tahun 2019.[1]

2. TINJAUAN PUSTAKA

2.1. Data Mining

Data mining adalah teknik untuk menemukan informasi bernilai tambah secara manual. Data mining telah digunakan untuk mengungkap hubungan antara data yang akan dikelompokkan ke dalam satu atau lebih kelompok sehingga item dalam kelompok secara substansial mirip satu sama lain sejak tahun 1990-an sebagai pendekatan praktis dan terarah untuk mengambil pola dan data. [2]

Menemukan informasi tersembunyi dalam database adalah ide mendasar di balik penambangan data, yang merupakan komponen Knowledge Discovery in Databases (KDD) untuk

mengidentifikasi informasi dan pola yang relevan dalam data. Penambangan data secara berulang menggunakan prosedur otomatis atau manual untuk mencari informasi baru, penting, dan bermanfaat dalam kumpulan data yang dibantu komputer dan manusia. [2]

2.2. Clustering

Salah satu ilmu dari data mining adalah clustering, yang melibatkan pengelompokan data atau objek ke dalam cluster (group) sehingga setiap cluster memiliki data yang sama dengan cluster lain yang layak namun berbeda dari mereka. Para peneliti masih berusaha untuk memperbaiki model cluster dan menentukan berapa banyak cluster yang dibutuhkan untuk membuat cluster yang ideal.[3]

Data dikelompokkan ke dalam berbagai cluster atau kelompok menggunakan proses clustering untuk membuat data di dalam setiap cluster sebanding mungkin dan setidaknya agak mirip dengan data di seluruh cluster. [4]

Ada dua pendekatan pengelompokan yang digunakan dalam penambangan data: pengelompokan hierarkis dan pengelompokan non-hierarkis.

Untuk mengatur data, pengelompokan hierarkis terlebih dahulu mengelompokkan dua atau lebih objek yang paling sebanding. Saat memilih jumlah cluster yang Anda inginkan pertama-tama ditentukan oleh pengelompokan non-hierarkis (dua cluster, tiga cluster, dll.). Prosedur pengelompokan kemudian dilakukan tanpa mengikuti pendekatan hierarkis setelah jumlah cluster diketahui. Sering disebut sebagai pengelompokan K-means.

2.3. Algoritma K-Means

K-Means Menggunakan clustering, yang dalam hal ini mewakili rata-rata grup data yang diidentifikasi sebagai cluster, kita dapat mengelompokkan data dengan menyajikan jumlah cluster konstan yang diinginkan. Analisis data dan teknik penambangan seperti clustering melakukan pemodelan yang tidak terkontrol dan merupakan salah satu cara untuk mengatur data menggunakan skema partisi. K-Method mencoba untuk membagi data yang sudah ada menjadi sejumlah kelompok, yang masing-masing memiliki karakteristik yang mirip dan berbeda dari data dari yang lain. Algoritma k-means mengatur satu set item n ke dalam cluster k menggunakan parameter input k sedemikian rupa sehingga kesamaan antara anggota cluster tinggi dan kesamaan antara objek individu rendah.

Kedekatan suatu objek dengan cluster homogen dapat digunakan untuk mengevaluasi seberapa mirip anggota cluster, atau dapat ditegaskan bahwa objek tersebut adalah centroid atau pusat cluster.[3]

Langkah-langkah dalam algoritma K-Means Clustering adalah sebagai berikut:

1. Jumlah kluster yang akan ditentukan menentukan berapa banyak grup data yang akan dibentuk atau dibuat.
2. Sentroid pertama. Sentroid awal dipilih secara acak, dan jumlahnya sebanyak yang akan ada cluster. Sentroid cluster pertama adalah episentrumnya atau awal dari pusat cluster.
3. Jarak dari setiap centroid data. Rumus menentukan pemisahan antara setiap cluster.

$$d_{eucliden(x,y)} = \sqrt{\sum_{i=1}^{n}(x_1 - y_1)^2}$$

   Dimana :
   x dan y : mewakili data atribut dari dua rekaman.

4. Pengelompokan data dalam cluster
5. Jika masih ada cluster pemindahan data, perubahan nilai sentroid, sesuatu yang berada di atas nilai ambang batas yang ditetapkan, atau perubahan nilai fungsi objektif yang digunakan olehnya yang lebih besar dari nilai ambang batas yang dinyatakan, kembali ke langkah 3 dari proses. Sentroid baru sedang dibangunkan oleh rumus.

$$C = \frac{\sum m}{n}$$

Dimana :
c = data untuk sentroid
m = anggota data yang terkait dengan sentroid tertentu
n = ukuran kumpulan data yang membentuk sentroid tertentu. [5]

2.4. Angka Kematian Bayi

Angka Kematian Bayi (AKB) adalah metrik penting untuk menilai tingkat kesehatan populasi karena menangkap keadaan penduduk secara keseluruhan. Perubahan dalam kesehatan dan kesejahteraan memiliki dampak signifikan pada angka ini. Kematian bayi adalah istilah yang digunakan untuk menggambarkan kematian yang terjadi setelah bayi lahir dan sampai anak berusia satu tahun. Erga Aprina Sari, Departemen Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, menyarankan metode berikut untuk menghitung kematian bayi:

IMR= $do\ B\ xk$ (1)

Dengan:
d0 = Jumlah kematian bayi di bawah satu tahun
B = Jumlah kelahiran hidup yang dilaporkan selama periode 12 bulan
K = Konstanta (1000)

2.5. Rapid Miner

Perangkat lunak yang disebut Rapid Miner berfungsi sebagai alat untuk mempelajari penambangan data. Sebuah perusahaan yang berspesialisasi dalam semua aspek bisnis big data, penelitian, pendidikan, dan pembelajaran membangun platform. Lebih dari 100 solusi pembelajaran untuk pengelompokan, klasifikasi, dan analisis regresi

tersedia di RapidMiner. Selain itu, RapidMiner mendukung sekitar 22 jenis file, termasuk.xls dan.csv.

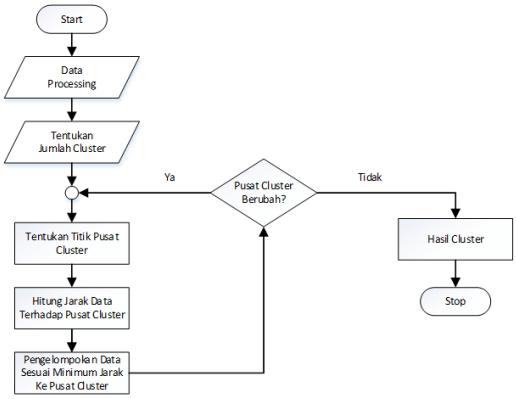
Pengguna dapat dengan mudah menggunakan Rapid Miner berkat antarmuka pengguna (UI) yang ramah pengguna. Viewpoint adalah nama tampilan yang ditawarkan oleh RapidMiner. Perspektif penyambutan, perspektif desain, dan perspektif hasil adalah tiga perspektif. [6]

2.6. Excel

Excel atau Office dari Microsoft Microsoft Corporation membuat dan merilis Excel, alat spreadsheet yang kompatibel dengan Windows dan Mac OS. Suite Microsoft Office mencakup program ini.

Berdasarkan sistem operasi Windows, Microsoft Excel adalah aplikasi dari keluarga Microsoft Office [7–11]. Microsoft Excel menggunakan tabel dengan baris dan kolom untuk memproses data yang disajikan sebagai angka atau bilangan bulat. Alat pengolah data dan angka yang paling populer saat ini adalah Microsoft Excel, yang dapat digunakan dalam berbagai pengaturan pada perangkat seperti desktop, tablet, dan smartphone. Selain sistem operasi Windows, Microsoft Excel juga tersedia untuk Macintosh, Android, dan Apple IOS. Excel menjalankan rumus pada lembar bentang. [7]

3. METODE PENELITIAN



Gambar 1. Tahapan Knowledge Discovery in Database

Penelitian ini menggunakan pendekatan knowledge discovery in databases, yang melalui langkah-langkah berikut:

1. *Data Cleaning* merupakan Proses menghilangkan kebisingan atau menolak data yang salah atau tidak relevan dikenal sebagai pembersihan data. Para peneliti memanfaatkan aplikasi Microsoft Excel untuk memfilter setiap kolom dan mencari data yang kosong atau berlebihan selama prosedur pembersihan data ini.
2. Proses penggabungan data dari banyak database ke database lain dikenal sebagai data Integration. Integrasi data tidak digunakan dalam penelitian ini karena tidak diperlukan.
3. *Data selection* adalah proses menghapus informasi yang tidak relevan dari database dan hanya menggunakan informasi yang diperlukan untuk analisis.
4. Proses transformasi data agar lebih sesuai dengan kebutuhan dikenal dengan istilah *Data Transformation*. Data berupa angka dan data yang dapat digunakan sama-sama dapat ditangani selama proses Algoritma K-Means Clustering. Transformasi data yang akan digunakan dalam pemodelan K-Means Clustering digambarkan pada Tabel 1.

Tabel 1. Hasil Transformasi

Kode Kelurahan	Nama kelurahan	jumlah kejadian	tahun
3201190006	WANAHERANG	1	2019
3201190010	BOJONG KULUR	0	2019
3201190009	CIANGSANA	0	2019
3201190002	GUNUNG PUTRI	4	2019
3201190004	BOJONG NANGKA	1	2019
…	…	…	…
3201280004	KALONGSAWAH	2	2021
3201280005	SIPAK	1	2021
3201280011	JASINGA	0	2021
3201280010	KOLEANG	0	2021
3201280013	CIKOPOMAYAK	0	2021

5. Metode utama yang digunakan untuk mengekstrak informasi yang berguna dan tersembunyi dari data adalah Data Mining. Berdasarkan kelompok data mining, berbagai metode dapat diterapkan.
6. Penilaian Pola untuk menemukan pola yang menarik dalam distribusi pengetahuan yang ada dalam database.
7. *Knowledge Presentation* adalah visualisasi dan penyebaran data yang berkaitan dengan teknik yang digunakan pengguna untuk mendapatkan informasi.

4. HASIL DAN PEMBAHASAN

Dalam penelitian ini, penulis menggunakan dua metode pertama dari program Microsoft Excel dan metode kedua dari aplikasi RapidMiner.

Penerapan *Algoritma K-Means Clustering* pada Microsoft Excel mempunyai tahapan sebagai berikut:

1. Hitung nilai jumlah k-cluster.
   Karena klaster 2 adalah klaster terbaik, hanya dua klaster yang digunakan dalam penyelidikan ini. Cluster tinggi (C0) dan cluster rendah adalah jenis cluster yang terbentuk (C1).
2. Temukan Nilai Centroid 2. (Pusat Cluster)
   Perhitungan acak digunakan untuk memilih pusat kluster awal, yang didasarkan pada data dari rentang. Tabel data sentroid yang memiliki tabel dengan data yang terlihat tercantum di bawah ini.

Tabel 2. Centroid Awal

centroid	x	y	z
C0	3201197116	1	2019
C1	3201035565	0	2020

3. Mengukur jarak antara setiap titik data dan sentroid (Cluster Center)
   Menghitung jarak dari setiap data ke pusat cluster datang setelah menentukan data nilai puat cluster awal. Tabel berikut menunjukkan cara kerja prosedur pencarian jarak terdekat iterasi 1:

Tabel 3. Hasil Perhitungan Jarak Pusat Cluster Iterasi 1

Nama kelurahan	jumlah kejadian	Tahun	C0	C1	Jarak Terdekat	cluster
WANAHERANG	1	2019	7110	154441	7110	0
BOJONG KULUR	0	2019	7106	154445	7106	0
CIANGSANA	0	2019	7107	154444	7107	0
GUNUNG PUTRI	4	2019	7114	154437	7114	0
BOJONG NANGKA	1	2019	7112	154439	7112	0
TLAJUNG UDIK	0	2019	7113	154438	7113	0
CICADAS	0	2019	7111	154440	7111	0
….	….	….	….	….	….	….
CIBENING	0	2021	167108	5557	5557	1
GUNUNGBUNDER I	1	2021	167109	5558	5558	1
CIBITUNG KULON	0	2021	167106	5555	5555	1
GUNUNG PICUNG	0	2021	167107	5556	5556	1
CIASIHAN	1	2021	167112	5561	5561	1
GUNUNGSARI	0	2021	167111	5560	5560	1
CIASMARA	0	2021	167113	5562	5562	1

Berdasarkan hasil perhitungan Microsoft Excel dalam iterasi 1 ditemukan 439 sata sebagai cluster tinggi dan 162 data sebagai cluster rendah selama tiga tahun terakhir.

4. Gunakan temuan pada setiap cluster untuk menghitung sentroid baru. Lanjutkan ke iterasi kedua setelah mendapatkan hasil jarak dari setiap objek dalam iterasi pertama.

Tabel 4. Centoid Baru

centriod	x	y	z
C0	3201190006	1	2020
C1	3201020018	0	2020

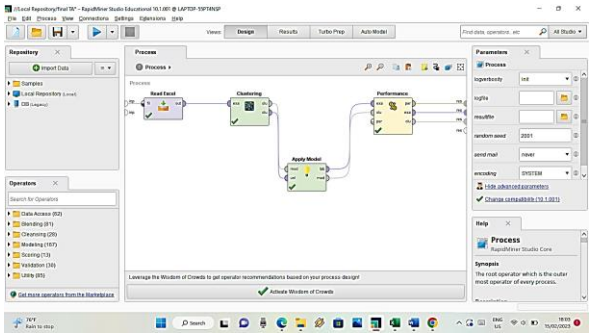
5. Selain itu, prosedur identik yang digunakan untuk iterasi 1 digunakan untuk menghitung hasil iterasi kedua. Proses iterasi berakhir jika posisi cluster tidak berubah dan nilai sentroid iterasi 2 sama dengan nilai sentroid sebelumnya. Prosedur iterasi, bagaimanapun, berlanjut dalam iterasi berikutnya jika nilai sentroid berbeda dan posisi data masih bergeser.

Tabel 5. Hasil Perhitungan Jarak Pusat Cluster Iterasi 2

Nama kelurahan	jumlah kejadian	tahun	C0	C1	Jarak Terdekat	cluster
WANAHERANG	1	2019	1,034758	169988	1,034758	0
BOJONG KULUR	0	2019	4,186171	169992	4,186171	0
CIANGSANA	0	2019	3,244075	169991	3,244075	0
GUNUNG PUTRI	4	2019	5,264107	169984	5,264107	0
BOJONG NANGKA	1	2019	2,251827	169986	2,251827	0
TLAJUNG UDIK	0	2019	3,244075	169985	3,244075	0
…	…	…	…	…	…	…
GUNUNGBUNDER I	1	2021	159999	9989	9989	1
CIBITUNG KULON	0	2021	159996	9992	9992	1
GUNUNG PICUNG	0	2021	159997	9991	9991	1
CIASIHAN	1	2021	160002	9986	9986	1
GUNUNGSARI	0	2021	160001	9987	9987	1
CIASMARA	0	2021	160003	9985	9985	1

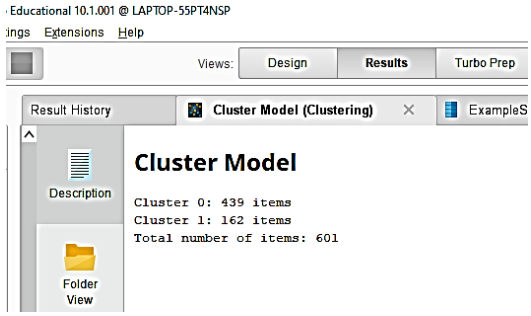
Berdasarkan hasil perhitungan Microsoft Excel dijelaskan bahwa hasil perhitungan jarak cluster dari iterasi 1 ke iterasi 2 menghasilkan nilai yang sama yaitu cluster 0 sebanyak 439 dan low cluster 162, dan memasukkan perhitungan Microsoft Excel ke dalam RapidMiner juga memiliki nilai yang sama.

Implementasi Algoritma K-Means Clustering di RapidMiner Studio 10.1. Desain operator RapidMiner yang digunakan dalam penelitian, ditunjukkan pada Gambar 1, melewati tahap-tahap berikut:



Gambar 2. Desain K-Means Pada Rapid Miner

Ketika pemodelan ini selesai, model cluster yang mirip dengan yang ada pada gambar akan diproduksi.



Gambar 3. Cluster Model

Dalam cluster model ini menghasilkan 2 cluster yaitu cluster 0 dan cluster 1. cluster 0 cluster 0 diidentifikasi sebanyak 439 items dan cluster 1 diidentifikasi sebanyak 162 items. Dan pada hasil perhitungan rapid miner menghasilkan Davies Bouldin Indeks seperti pada gambar berikut:

Gambar 4. Nilai Davies Bouldin Indeks

Dari hasil pengklusteran ini menghasilkan nilai Davies Bouldin Indeks sebesar 0.101, rata-rata pada jarak pusat: 511721137.495, rata-rata pada jarak pusat 0 : 686582206.698 dan rata-rata pada jarak pusat 1 : 37869227.741.

5. KESIMPULAN DAN SARAN

Berdasarkan temuan penelitian dan perdebatan, ditentukan bahwa dua klaster, klaster tinggi dan klaster rendah, menghasilkan hasil klasterisasi terbaik. Indeks nilai Davies Bouldin dihitung dengan pengelompokan angka kematian bayi menggunakan metode k means clustering menggunakan dua klaster, dengan klaster tinggi terdiri dari 439 desa dan klaster rendah terdiri dari 162 desa. Dan mengasilkan Davies Bouldin Indeks sebesar 0.101, rata-rata pada jarak pusat : 511721137.495, rata-rata pada jarak pusat 0 : 686582206.698 dan rata-rata pada jarak pusat 1 : 37869227.741. Selain itu, penulis mendesak agar lebih banyak analisis dilakukan pada pengelompokan K-Means untuk memungkinkan perbandingan yang lebih tepat dan unggul.

DAFTAR PUSTAKA

[1] N. Qisthi, Y. Andriyana, and Y. Hidayat, "Pemodelan Angka Kematian Bayi di Jawa Barat Menggunakan Analisis Regresi Kuantil," *Pros. Semin. Nas. Stat. | Dep. Stat. FMIPA Univ. Padjadjaran*, vol. 10, pp. 55–55, 2021, [Online]. Available: http://prosiding.statistics.unpad.ac.id/index.php/prosidingnasional/article/view/122

[2] P. M. S. Tarigan, J. T. Hardinata, H. Qurniawan, and ..., "Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang: Studi Kasus: Toko Sinar Harahap," *... Janitra Inform. dan ...*, vol. 12, no. 2, pp. 51–61, 2022, [Online]. Available: http://www.janitra.org/index.php/home/article/view/142

[3] F. Yunita, "Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru," *Sistemasi*, vol. 7, no. 3, p. 238, 2018, doi: 10.32520/stmsi.v7i3.388.

[4] K. Fatmawati and A. Perdana Wirdanto, "DATA MINING : PENERAPAN RAPIDMINER DENGAN K-MEANS CLUSTER PADA DAERAH TERJANGKIT DEMAM BERDARAH DENGUE (DBD) BERDASARKAN PROVINSI," vol. 3, no. 2, pp. 173–178, 2018.

[5] R. Supardi and I. Kanedi, "Implementasi Metode Algoritma K-Means Clustering pada Toko Eidelweis," *J. Teknol. Inf.*, vol. 4, no. 2, pp. 270–277, 2020, doi: 10.36294/jurti.v4i2.1444.

[6] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, "Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 1, pp. 8–17, 2021, doi: 10.25077/teknosi.v7i1.2021.8-17.

[7] D. Andriyani, E. Harahap, F. H. Badruzzaman, M. Y. Fajar, and D. Darmawan, "Aplikasi Microsoft Excel Dalam Penyelesaian Masalah Rata-rata Data Berkelompok," *Matematika*, vol. 18, no. 1, pp. 41–46, 2019, doi: 10.29313/jmtm.v18i1.er5078.

[8] A. H. Ardiansyah, W. Nugroho, N. H. Alfiyah, R. A. Handoko, and M. A. Bakhtiar, "Penerapan Data Mining Menggunakan Metode Clustering untuk Menentukan Status Provinsi di Indonesia 2020," *Semin. Nas. Inov. Teknol.*, vol. 4, no. 3, pp. 329–333, 2020.