INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 12 No. 1 2026

KOMPARASI K-MEANS DAN NAÏVE BAYES PADA ANALISIS MIGRASI PENDUDUK MAJALENGKA

Budiman[1], Ii Sopiandi[2] , Mei Bunga Kiranti[3]
[1,2,3]Universitas Majalengka, Majalengka, Indonesia
Penulis Korespondensi: iisopiandi@unma.ac.id

ABSTRAK

Penelitian ini bertujuan membandingkan performa algoritma *K-Means Clustering* dan *Naïve Bayes Classification* dalam menganalisis pola migrasi penduduk antar kecamatan di Kabupaten Majalengka menggunakan data migrasi tahun 2024 yang diperoleh dari Dinas Kependudukan dan Pencatatan Sipil. Penelitian menggunakan pendekatan *Knowledge Discovery in Database* (KDD) melalui tahapan seleksi data, *preprocessing*, transformasi, pemodelan, dan evaluasi. Algoritma *K-Means* digunakan untuk mengelompokkan wilayah berdasarkan karakteristik migrasi, sedangkan *Naïve Bayes* digunakan untuk memprediksi kecenderungan perubahan migrasi. Hasil penelitian menunjukkan bahwa algoritma *K-Means* menghasilkan tiga kelompok migrasi, yaitu tinggi, sedang, dan rendah, dengan kondisi konvergen pada iterasi kedua. Sementara itu, algoritma *Naïve Bayes* memperoleh akurasi sebesar 90%, precision 88%, dan recall 91%. Hasil komparasi menunjukkan bahwa *K-Means* unggul dalam eksplorasi pola data, sedangkan *Naïve Bayes* lebih efektif untuk klasifikasi prediktif. Temuan ini dapat digunakan sebagai dasar rekomendasi pemilihan metode analisis migrasi berbasis data bagi pemerintah daerah.

***Kata Kunci**: K-Means, Naïve Bayes, migrasi penduduk, data mining, klasifikasi.*

Riwayat Artikel :

Tanggal diterima : 20-05-2026 Tanggal terbit : 15-06-2026

Kutipan :

Budiman & supiandi Ii. (2026). NARATIVE REVIEW PEMANFAATAN INTERNET-OF-THINGS UNTUK APLIKASI SEED MONITORING AND MANAGEMENT SYSTEM PADA MEDIA TANAMAN HIDROPONIK DI INDONESIA. INFOTECH Journal, 9(1), 38–45. https://doi.org/10.31949/infotech.v9i1.4439

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pemanfaatan data sebagai landasan utama dalam pengambilan keputusan pada berbagai bidang, termasuk sektor kependudukan. Data kependudukan memiliki peran strategis dalam mendukung perencanaan pembangunan daerah, khususnya dalam memahami dinamika perpindahan penduduk yang dapat memengaruhi distribusi sumber daya, pembangunan infrastruktur, serta kualitas pelayanan publik (Chen & Liu, 2024). Salah satu fenomena penting dalam kajian kependudukan adalah migrasi penduduk, yaitu perpindahan individu atau kelompok dari satu wilayah ke wilayah lain dalam periode tertentu, baik bersifat sementara maupun permanen (Busert-Sebela et al., 2025).

Migrasi penduduk umumnya dipengaruhi oleh berbagai faktor, seperti kondisi ekonomi, kesempatan kerja, tingkat pendidikan, fasilitas publik, serta perkembangan wilayah. Tingginya mobilitas penduduk pada suatu daerah sering kali mencerminkan ketimpangan pembangunan maupun daya tarik wilayah tertentu dibandingkan wilayah lainnya (Yuniati Ningsih et al., 2022). Oleh karena itu, analisis pola migrasi menjadi penting untuk membantu pemerintah daerah dalam merumuskan kebijakan pembangunan berbasis data secara lebih efektif.

Pendekatan data mining saat ini banyak digunakan dalam analisis kependudukan, khususnya untuk mengidentifikasi pola migrasi, klasifikasi wilayah, serta pengambilan keputusan berbasis data (Han & Kim, 2023) (Lee & Park, 2023)

Kabupaten Majalengka merupakan salah satu wilayah yang mengalami dinamika perpindahan penduduk cukup signifikan. Berdasarkan data Dinas Kependudukan dan Pencatatan Sipil Kabupaten Majalengka tahun 2024, jumlah migrasi masuk tercatat sebanyak 9.331 jiwa, sedangkan migrasi keluar mencapai 10.270 jiwa. Selisih tersebut menunjukkan kecenderungan perpindahan keluar wilayah yang perlu dikaji lebih lanjut untuk memahami pola persebarannya antar kecamatan.

Pendekatan *data mining* telah banyak digunakan untuk mengidentifikasi pola tersembunyi pada data berskala besar. Salah satu metode yang banyak diterapkan adalah algoritma *K-Means Clustering* yang efektif dalam mengelompokkan data berdasarkan tingkat kemiripan karakteristik antar objek(Afidah, 2023). Penelitian (Aryanto et al., 2024) menunjukkan bahwa algoritma ini mampu menghasilkan segmentasi data yang konsisten untuk mendukung pengambilan keputusan berbasis wilayah. Algoritma K-Means dinilai efektif dalam proses pengelompokan data demografi karena mampu mengidentifikasi pola kemiripan antarwilayah berdasarkan karakteristik tertentu (Putri & Setiawan, 2023)

Selain itu, algoritma *Naïve Bayes Classification* dikenal memiliki kemampuan klasifikasi yang baik melalui pendekatan probabilistik berbasis data berlabel. Penelitian (Nurahman et al., 2022) menunjukkan bahwa metode ini menghasilkan akurasi tinggi pada klasifikasi data kependudukan. Temuan serupa juga diperoleh oleh (Pramana, 2023), yang menunjukkan efektivitas Naïve Bayes dalam menghasilkan prediksi berbasis probabilitas dengan performa stabil.

Penelitian komparatif antara algoritma K-Means dan Naïve Bayes juga telah dilakukan pada beberapa domain. (Nurhachita & Negara, 2020) menunjukkan bahwa K-Means lebih efektif untuk segmentasi data eksploratif, sedangkan Naïve Bayes unggul pada klasifikasi berbasis label. Hasil serupa ditemukan oleh(Nasir et al., 2024) yang menyatakan bahwa Naïve Bayes memiliki akurasi prediksi lebih tinggi dibandingkan metode clustering pada data akademik.

Meskipun demikian, kajian komparatif yang membandingkan kedua algoritma tersebut secara langsung pada analisis migrasi penduduk tingkat kecamatan, khususnya di Kabupaten Majalengka, masih sangat terbatas. Sebagian besar penelitian hanya berfokus pada penerapan salah satu algoritma tanpa mengevaluasi keunggulan relatif keduanya dalam konteks analisis migrasi daerah.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan melakukan analisis komparatif antara algoritma *K-Means Clustering* dan *Naïve Bayes Classification* dalam mengidentifikasi pola migrasi penduduk antar kecamatan di Kabupaten Majalengka. Hasil penelitian diharapkan dapat memberikan rekomendasi metode analisis yang tepat untuk mendukung kebijakan kependudukan berbasis data secara lebih akurat dan terukur.

2. METODE

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Database* (KDD) sebagai kerangka kerja utama dalam proses pengolahan data. Pendekatan ini dipilih karena mampu menghasilkan proses analisis data yang sistematis melalui tahapan seleksi data, *preprocessing*, transformasi, *data mining*, dan evaluasi hasil (Rahman & Alam, 2023). Tahapan penelitian dilakukan untuk membandingkan performa algoritma *K-Means Clustering* dan *Naïve Bayes Classification* dalam mengidentifikasi pola migrasi penduduk antar kecamatan di Kabupaten Majalengka.

1. Sumber Data

Data penelitian diperoleh dari Dinas Kependudukan dan Pencatatan Sipil Kabupaten Majalengka tahun 2024. Dataset terdiri atas 26 kecamatan dengan dua atribut utama, yaitu jumlah migrasi masuk dan jumlah migrasi keluar. Kedua atribut tersebut digunakan untuk menganalisis karakteristik

perpindahan penduduk pada masing-masing wilayah.

Pada Tabel 1 dijelaskan menyajikan struktur data penelitian yang digunakan sebagai dasar analisis komparatif kedua algoritma.

Tabel 1 Struktur Dataset

Variabel	Tipe Data	Keterangan
Kecamatan	Nominal	Nama wilayah administratif
Migrasi Masuk	Numerik	Jumlah penduduk masuk
Migrasi Keluar	Numerik	Jumlah penduduk keluar

2. Pra-Pemorosesan Data

Tahap *preprocessing* dilakukan untuk memastikan kualitas data sebelum proses analisis. Tahapan ini meliputi pemeriksaan data kosong (*missing value*), validasi konsistensi nilai numerik, dan pengecekan duplikasi data. Berdasarkan hasil pemeriksaan, seluruh data dinyatakan valid dan tidak ditemukan anomali yang dapat memengaruhi proses analisis.

Selanjutnya dilakukan transformasi data dengan membentuk variabel migrasi bersih menggunakan persamaan berikut:

Migrasi Bersih = Migrasi Masuk – Migrasi Keluar

Variabel ini digunakan sebagai atribut tambahan pada proses klasifikasi menggunakan algoritma *Naïve Bayes*.

3. Implementasi Algoritma K-Means

Algoritma *K-Means Clustering* digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan pola migrasi antar kecamatan. Jumlah klaster ditentukan sebanyak tiga kelompok, yaitu kategori migrasi rendah, sedang, dan tinggi.

Perhitungan jarak antar objek dilakukan menggunakan metode *Euclidean Distance* sebagai dasar penentuan kedekatan terhadap centroid.

$$\sum \boldsymbol{d(x,y) = \sum i = 1n(xi - yi)2}$$

Proses iterasi dilakukan hingga tidak terjadi perpindahan anggota klaster, yang menandakan bahwa model telah mencapai kondisi konvergen.

4. Implementasi Algoritma Naïve Bayes

Algoritma *Naïve Bayes Classification* digunakan untuk memprediksi kecenderungan migrasi berdasarkan kategori kenaikan dan penurunan migrasi bersih. Dataset dibagi menjadi 60% data pelatihan dan 40% data pengujian.

Model klasifikasi dibangun menggunakan pendekatan probabilistik berdasarkan Teorema Bayes.

5. Evaluasi Model

Evaluasi dilakukan untuk membandingkan efektivitas kedua algoritma. Pada algoritma *K-Means*, evaluasi dilakukan berdasarkan kestabilan hasil klasterisasi dan interpretasi pola kelompok yang terbentuk. Sementara itu, pada algoritma *Naïve Bayes*, evaluasi menggunakan *confusion matrix* dengan metrik akurasi, presisi, dan *recall*.

Tabel 2 menunjukkan indikator evaluasi yang digunakan pada penelitian ini.

Tabel 2. Indikator Evaluasi

Algoritma	Metode Evaluasi
K-Means	Stabilitas klaster
Naïve Bayes	Accuracy, Precision, Recall

3. PEMBAHASAN

1. Pengumpulan Data

Dalam proses pengumpulan data, terdapat dua parameter utama yang digunakan dalam pengolahan, yaitu migrasi masuk dan migrasi keluar. Pada penelitian ini, data yang dimanfaatkan mencakup data perpindahan penduduk yang datang dan pindah (migrasi masuk dan migrasi keluar) antar kecamatan di Kabupaten Majalengka selama tahun 2024. Data tersebut diperoleh dari Dinas Kependudukan dan Pencatatan Sipil Kabupaten Majalengka melalui Bidang Kependudukan bagian Migrasi Penduduk. Informasi tersebut disajikan dalam bentuk tabel untuk mempermudah proses analisis.

2. Pengolahan Data

Proses pengolahan data menggunakan dua metode, yaitu algoritma K-Means untuk proses clustering dan algoritma Naive Bayes untuk klasifikasi. Melalui penerapan kedua algoritma ini, diharapkan dapat diperoleh hasil pengelompokan dan prediksi data yang sesuai dengan tujuan penelitian. Adapun tahapan pelaksanaan masing-masing algoritma dijelaskan sebagai berikut. Kedua algoritma yang digunakan memiliki pendekatan yang berbeda: K-Means diterapkan untuk tujuan pengelompokan data berdasarkan kemiripan jumlah migrasi masuk dan keluar tanpa menggunakan label target, dan Naïve Bayes sebagai supervised learning untuk klasifikasi data berlabel. Perbedaan pendekatan ini akan menjadi dasar dalam evaluasi hasil pada bagian selanjutnya.

3. Algoritma K-Means Clustering

a. Seleksi Data

Pemilihan data dilakukan dengan mengambil atribut yang berkaitan dengan indikator perpindahan penduduk antar kecamatan di Kabupaten Majalengka. Atribut yang digunakan dalam penelitian ini adalah nama kecamatan, jumlah migrasi masuk, dan jumlah migrasi keluar. Data diambil dari hasil rekapitulasi laporan tahun 2024 pada Dinas Kependudukan dan Pencatatan Sipil Kabupaten Majalengka.

b. Pra-pemrosesan

Pada tahap ini dilakukan pemeriksaan terhadap kelengkapan data untuk memastikan bahwa seluruh entri memiliki nilai yang valid. Berdasarkan hasil pengecekan, tidak ditemukan adanya data kosong (missing values) pada atribut yang digunakan,

seperti jumlah migrasi masuk dan migrasi keluar di masing masing kecamatan. Oleh karena itu, data dianggap bersih dan dapat langsung digunakan untuk proses transformasi dan analisis selanjutnya.

c. Transformasi Data

Tahap transformasi data dilakukan untuk menyiapkan data numerik agar dapat diproses dengan algoritma K-Means Clustering. Pada tahap ini, atribut jumlah migrasi masuk dan jumlah migrasi keluar dari setiap kecamatan tetap digunakan dalam bentuk aslinya karena sudah berbentuk numerik dan tidak mengandung nilai kosong. Tidak dilakukan normalisasi karena rentang nilai antar atribut relatif sebanding.

d. Clustering dengan K-Means

1) Menentukan jumlah cluster

Pengelompokan data pada pengujian ini sebanyak 3 cluster (k= 3), diantaranya yaitu cluster 1 menunjukkan tingkat migrasi sedang (C1), cluster 2 menunjukkan tingkat migrasi rendah (C2), dan cluster 3 menunjukkan tingkat migrasi tinggi (C3).

2) Menentukan Centroid (pusat cluster)

Penentuan pusat awal cluster ditentukan secara random yang diambil dari data pada tabel 4.3 Sehingga terpilih untuk cluster 1 (C1) Kecamatan Cigasong, cluster 2 (C2) Kecamatan Banjaran, dan cluster 3 (C3) Kecamatan Cikijing.

Tabel 3 Pusat Cluster awal

Migrasi	C1	C2	C3
Masuk	374	96	498
Keluar	387	110	563

3) Menghitung jarak setiap objek ke centroid pada masing-masing cluster. Berikut rumus untuk menggitung jarak dari centroid adalah:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Sampel perhitungan manual menggunakan rumus diatas:

Kecamatan Kertajati

Hitung jarak kecamatan kertajati (332, 426) ke centroid 1

$d = \sqrt{(332-374)^2 + (426-387)^2} =$

$\sqrt{(-42)^2 + (39)^2} = \sqrt{1{,}764 + 1{,}521}$

$= \sqrt{3{,}285} = 57{,}3$

Hitung jarak kecamatan kertajati (332, 426) ke centroid 2

$d = \sqrt{(332-96)^2 + (426-110)^2} =$

$\sqrt{(236)^2 + (316)^2} = \sqrt{55{,}696 + 99{,}856}$

$= \sqrt{155{,}552} = 394{,}4$

Hitung jarak kecamatan kertajati (332, 426) ke centroid 3

$d = \sqrt{(332-498)^2 + (426-563)^2} =$

$\sqrt{(-166)^2 + (-137)^2} = \sqrt{27{,}556 + 18{,}769}$

$= \sqrt{46{,}325} = 215{,}2$

Dengan dilakukan perhitungan menggunakan rumus di atas diperoleh table 4 hasil perhitungan jarak centroid literasi ke 1sebagai berikut.

Tabel 4 Perhitungan Jarak Centroid Literasi Ke 1

Kecamatan	Migrasi Masuk	Migrasi Keluar	C1	C2	C3
Cigasong	374	387	0	392,445	215,295
Cingambul	266	362	110,856	303,98	306,961
Dawuan	257	333	128,86	275,046	333,138
Jatitujuh	433	471	102,65	493,852	112,646
Kasokandel	314	341	75,6042	317,624	288,34
Kertajati	332	426	57,3149	394,401	215,232
Maja	311	332	83,6301	309,045	297,204
Malausma	282	420	97,7395	361,519	259,046
Palasah	390	424	40,3113	430,154	176,026
Rajagaluh	303	272	135,152	262,856	350,294
Sukahaji	453	382	79,1581	448,813	186,51
Talaga	283	390	91,0494	336,703	275,96
Argapura	117	175	333,156	68,3081	543,788
Banjaran	96	110	392,445	0	605,651
Bantarujeg	189	279	214,217	192,899	419,687
Panyingkiran	222	217	228,044	165,303	442,597
Sindang	75	99	415,145	23,7065	627,873
Sindangwangi	234	217	220,227	174,622	435,215
Cikijing	498	563	215,295	605,651	0
Jatiwangi	659	618	366,86	758,309	170,135
Kadipaten	416	523	142,338	522,464	91,236
Lemahsugih	397	588	202,312	564,876	104,048
Leuwimunding	526	514	198,073	590,014	56,4358
Ligung	633	663	378,493	770,829	168,003
Majalengka	652	541	317,805	703,489	155,564
Sumberjaya	619	623	340,178	732,597	135,059

4) Menentukan Cluster

Proses pengelompokan dilakukan dengan menentukan jarak minimum antara setiap objek (kecamatan) dengan centroid masing-masing klaster, lalu objek tersebut dimasukkan ke klaster yg memiliki nilai jarak terendah pada iterasi 1.bisa di lihat pada table 5 dibawah ini :

Tabel 5 Cluster Iterasi Ke-1

Kecamatan	Migrasi Masuk	Migrasi Keluar	C1	C2	C3	Jarak Terdekat	Cluster
Cigasong	374	387	0	392,4449	215,2951	0	1
Cingambul	266	362	110,8558	303,9803	306,9609	110,8558	1
Dawuan	257	333	128,8604	275,0455	333,1381	128,8604	1
Jatitujuh	433	471	102,6499	493,8522	112,6455	102,6499	1
Kasokandel	314	341	75,60423	317,624	288,3401	75,60423	1
Kertajati	332	426	57,31492	394,4008	215,2324	57,31492	1
Maja	311	332	83,63014	309,0453	297,2036	83,63014	1
Malausma	282	420	97,73945	361,519	259,0463	97,73945	1
Palasah	390	424	40,31129	430,1535	176,0256	40,31129	1
Rajagaluh	303	272	135,1518	262,8555	350,2942	135,1518	1

Kecamatan	Migrasi Masuk	Migrasi Keluar	C1	C2	C3	Jarak Terdekat	Cluster
Sukahaji	453	382	79,15807	448,8129	186,5101	79,15807	1
Talaga	283	390	91,04944	336,7031	275,9601	91,04944	1
Argapura	117	175	333,1561	68,30813	543,7876	68,30813	2
Banjaran	96	110	392,4449	0	605,6509	0	2
Bantarujeg	189	279	214,2172	192,8989	419,6868	192,8989	2
Panyingkiran	222	217	228,0439	165,3028	442,5969	165,3028	2
Sindang	75	99	415,1446	23,70654	627,8734	23,70654	2
Sindangwangi	234	217	220,2272	174,6224	435,2149	174,6224	2
Cikijing	498	563	215,2951	605,6509	0	0	3
Jatiwangi	659	618	366,8596	758,3093	170,1352	170,1352	3
Kadipaten	416	523	142,3376	522,4644	91,23596	91,23596	3
Lemahsugih	397	588	202,3116	564,8761	104,0481	104,0481	3
Leuwimunding	526	514	198,0732	590,0136	56,4358	56,4358	3
Ligung	633	663	378,4931	770,8294	168,003	168,003	3
Majalengka	652	541	317,805	703,4892	155,5635	155,5635	3
Sumberjaya	619	623	340,1779	732,5968	135,0592	135,0592	3

Dari table 5 diatas didapatkan data dengn jarak terdekat C1 sebanyak 12, jarak terdekat C2 sebanyak 6 dan jarak terdejat C3 sebanyak 8.

5) Setelah pengelompokan awal, iterasi dilanjutkan dengan menghitung centroid baru untuk setiap klaster. Perhitungan dilakukan dengan menjumlahkan seluruh nilai data dalam klaster tersebut, kemudian dibagi dengan jumlah anggota klaster. Nilai-nilai ini membentuk centroid iterasi ke-1 seperti yang terlihat pada tabel 5 di bawah ini :

$$= \left(\frac{1}{n_j} \sum_{i \in n_j} x_i, \frac{1}{n_j} \sum_{i \in n_j} y_i \right)$$

Dimana:

Cj = Centroid baru

xi dan yi = Anggota cluster pada atribut ke-i

Nj = Jumlah data dalam cluster

Tabel 6 Centoid Baru Iterasi ke-1

Centroid	Migrasi Masuk	Migrasi Keluar
C1	333,166667	378,333333
C2	155,5	182,833333
C3	550	579,125

6) Setelah diperoleh centroid baru dari iterasi pertama, langkah selanjutnya adalah menghitung kembali jarak setiap objek terhadap centroid menggunakan metode yang sama. Hasil perhitungan jarak pada iterasi ke-2 ditunjukkan pada tabel ke 7 berikut.

Tabel 7 Perhitungan Jarak Centoid dan Hasil Cluster Iterasi ke-2

Kecamatan	Migrasi Masuk	Migrasi Keluar	C1	C2	C3	Jarak Terdekat	Cluster
Cigasong	374	387	41,74293	299,0423	260,5533	41,74293	1
Cingambul	266	362	69,12408	210,5016	357,4902	69,12408	1
Dawuan	257	333	88,63674	181,252	382,6572	88,63674	1
Jatitujuh	433	471	136,2124	400,0578	159,3111	136,2124	1
Kasokandel	314	341	41,96593	223,9173	335,2604	41,96593	1
Kertajati	332	426	47,68094	300,4701	266,4043	47,68094	1
Maja	311	332	51,36282	215,4784	343,7903	51,36282	1
Malausma	282	420	65,98591	268,7941	311,6806	65,98591	1
Palasah	390	424	72,90729	336,3802	222,8537	72,90729	1
Rajagaluh	303	272	110,5297	172,357	394,1253	110,5297	1
Sukahaji	453	382	119,8894	358,0134	219,6981	119,8894	1
Talaga	283	390	51,50539	243,2576	327,1961	51,50539	1
Argapura	117	175	296,7701	39,28882	592,2888	39,28882	2
Banjaran	96	110	358,1212	94,04756	652,8356	94,04756	2
Bantarujeg	189	279	175,0747	101,8346	469,4635	101,8346	2
Panyingkiran	222	217	195,9247	74,7637	488,5883	74,7637	2
Sindang	75	99	380,3645	116,2251	675,3851	116,2251	2
Sindangwangi	234	217	189,3739	85,61315	480,6147	85,61315	2
Cikijing	498	563	247,5314	511,6961	54,44277	54,44277	3
Jatiwangi	659	618	404,4842	665,494	115,725	115,725	3
Kadipaten	416	523	166,7027	428,4549	145,2791	145,2791	3
Lemahsugih	397	588	219,1684	471,6803	153,2572	153,2572	3
Leuwimunding	526	514	235,7756	496,9322	69,40652	69,40652	3
Ligung	633	663	413,443	677,1752	118,0001	118,0001	3
Majalengka	652	541	357,9318	612,2055	108,8922	108,8922	3
Sumberjaya	619	623	376,2479	639,2018	81,76806	81,76806	3

Dari hasil perhitungan iterasi kedua, diketahui bahwa setiap objek masih berada pada klaster yang sama seperti pada iterasi pertama. Karena tidak terjadi perpindahan objek antar klaster, proses iterasi dianggap selesai. Hasil akhir klasterisasi dapat dilihat pada uraian berikut:

Cluster 1 (C1) memiliki 12 data yang berarti bahwa kelompok pertama merupakan kelompok dengan tingkat migrasi sedang, pada tahun 2024.

Cluster 2 (C2) memiliki 6 data yang berarti bahwa kelompok kedua merupakan kelompok dengan tingkat migrasi rendah, pada tahun 2024.

Cluster 3 (C3) memiliki 8 data yang berarti bahwa kelompok ketiga merupakan kelompok dengan tingkat migrasi tinggi, pada tahun 2024.

Karena K-Means tidak menggunakan label target dalam prosesnya, maka algoritma ini dikategorikan sebagai unsupervised learning. Artinya, proses klasterisasi dilakukan hanya berdasarkan kemiripan data tanpa label.

e. Evaluasi Hasil Analisis

Implementasi dan pengujian dilakukan dengan memanfaatkan aplikasi RapidMiner Studio. Tujuan dari tahap ini adalah untuk melakukan perbandingan antara hasil pengolahan data secara manual dan hasil yang diperoleh melalui pemrosesan menggunakan perangkat lunak tersebut.

Implementasi menggunakan RapidMiner Studio menunjukkan bahwa proses klasterisasi mencapai kondisi konvergen pada iterasi kedua tanpa perpindahan anggota klaster. Hasil ini menandakan kestabilan model dalam membentuk pola migrasi berdasarkan karakteristik data.

2 . Algoritma Naïve Bayes Clasification

a. Seleksi Data

Tahapan awal dalam proses Knowledge Discovery in Database (KDD) adalah seleksi data, yaitu pemilihan atribut-atribut yang relevan dan mendukung tujuan analisis. Dalam penelitian ini, data yang digunakan mencakup 26 kecamatan yang terdapat di wilayah Kabupaten Majalengka. Setiap kecamatan memiliki dua atribut utama, yaitu jumlah migrasi masuk dan jumlah migrasi keluar selama periode waktu tertentu. Kedua atribut ini dipilih karena merepresentasikan dinamika perpindahan penduduk yang terjadi di masing masing wilayah, yang nantinya akan digunakan untuk membentuk variabel variabel baru dan melakukan klasifikasi kecenderungan migrasi. Pemilihan data pada tahap ini didasarkan pada ketersediaan data, relevansi dengan permasalahan penelitian, serta potensi informasi yang dapat digali lebih lanjut melalui proses data mining.

b. Pra-Pemrosesan Data

Setelah data dipilih, langkah selanjutnya adalah melakukan prapemrosesan data guna memastikan bahwa data berada dalam kondisi bersih dan siap untuk diolah. Proses ini melibatkan pemeriksaan terhadap konsistensi, kelengkapan, dan validitas data, termasuk pengecekan terhadap nilai kosong (missing values), data duplikat, serta outlier atau nilai ekstrim yang mungkin memengaruhi hasil analisis. Berdasarkan hasil prapemrosesan, seluruh data migrasi yang digunakan dalam penelitian ini telah melalui validasi dan tidak ditemukan adanya anomali yang signifikan. Dengan demikian, data dapat langsung dilanjutkan ke tahap transformasi untuk membentuk variabel-variabel baru yang lebih representatif terhadap fenomena yang dianalisis.

c. Transformai Data

Tahapan transformasi data merupakan inti dari proses pembentukan fitur dan label klasifikasi. Dalam penelitian ini, dilakukan perhitungan nilai migrasi bersih untuk setiap kecamatan, yang didefinisikan sebagai selisih antara jumlah migrasi masuk dan jumlah migrasi keluar. Rumus yang digunakan adalah:

Migrasi Bersih = Migrasi Masuk – Migrasi Keluar

d. Clasification dengan Naïve Bayes

Tahap ini menerapkan algoritma Naïve Bayes untuk memprediksi kenaikan atau penurunan migrasi berdasarkan data migrasi masuk, keluar, dan bersih. Dari total 26 kecamatan, data dibagi secara acak menjadi 60% data training (16 kecamatan) dan 40% data testing (10 kecamatan). Model dilatih menggunakan data training, lalu dievaluasi dengan data testing menggunakan metrik akurasi. Naïve Bayes merupakan metode supervised learning karena memanfaatkan label target (kenaikan/penurunan) untuk melatih model prediksi. Pendekatan ini memungkinkan evaluasi kinerja secara kuantitatif menggunakan confusion matrix dan metrik seperti akurasi, presisi, serta recall. Data Training yang digunakan adalah 16 kecamatan yang berada pada kabupaten Majalengka, seperti yang ada pada tabel 8 di bawah ini :

Tabel 8 Data Training

Kecamatan	Migrasi Masuk	Migrasi Keluar	Migrasi Bersih	Kategori	Klasifikasi
Argapura	117	175	-58	Sedang	Penurunan
Banjaran	96	110	-14	Sedang	Penurunan
Bantarujeg	189	279	-90	Sedang	Penurunan
Cingambul	266	362	-96	Rendah	Penurunan
Jatitujuh	433	471	-38	Sedang	Penurunan
Jatiwangi	659	618	41	Tinggi	Kenaikan
Kadipaten	416	523	-107	Rendah	Penurunan
Kasokandel	314	341	-27	Sedang	Penurunan
Kertajati	332	426	-94	Sedang	Penurunan
Lemahsugih	397	588	-191	Rendah	Penurunan
Leuwimunding	526	514	12	Tinggi	Kenaikan
Panyingkiran	222	217	5	Tinggi	Kenaikan
Sindang	75	99	-24	Sedang	Penurunan
Sukahaji	453	382	71	Tinggi	Kenaikan
Sumberjaya	619	623	-4	Sedang	Penurunan
Talaga	283	390	-107	Rendah	Penurunan

1) Perhitungan probabilitas prior migrasi

Berdasarkan 16 data training yang digunakan, diketahui bahwa kelas C0 (Kenaikan) terdiri dari 4 kecamatan, sedangkan kelas C1 (Penurunan) mencakup 12 kecamatan. Peluang terjadinya penurunan migrasi bersih dihitung berdasarkan distribusi dari kedua kategori tersebut.

$$P(C0) = \frac{4}{16} = 0.25$$

Sedangkan perhitungan peluang penurunan yaitu:

$$P(C0) = \frac{12}{16} = 0.75$$

2) Perhitungan probabilitas posterior migrasi

Menghitung peluang posterior dilakukan pada data training sebanyak 16 dengan X sebagai vektor peluang penurunan migrasi yaitu XKecamatan dan XKategori migrasi. Sehingga P(X|Ci) dapat dijabarkan menjadi (XXKecamatan, XXKategori migrasi|Ci). Pertama untuk mencari peluang kemungkinan hasil dari P(XXKecamatan |Ci) dapat dilihat pada tabel 9.

Tabel 9 Probabilitas Kecamatan

Kecamatan	Penurunan	Kenaikan
Argapura	0,08	0,00
Banjaran	0,08	0,00
Bantarujeg	0,08	0,00
Cingambul	0,08	0,00
Jatitujuh	0,08	0,00
Jatiwangi	0,00	0,25
Kadipaten	0,08	0,00
Kasokandel	0,08	0,00
Kertajati	0,08	0,00
Lemahsugih	0,08	0,00
Leuwimunding	0,00	0,25
Panyingkiran	0,00	0,25
Sindang	0,08	0,00
Sukahaji	0,00	0,25
Sumberjaya	0,08	0,00
Talaga	0,08	0,00
Total	100%	100%

Selanjutnya mencari peluang hasil kriteria P(XXKategori migrasi|Ci) dapat dilihat pada tabel 9.

Tabel 10 Probabilitas kategori Migrasi

Kategori Migrasi	Penurunan	Kenaikan
Sedang	0,67	0,00
Rendah	0,33	0,00
Tinggi	0,00	1,00
Total	100%	100%

Berdasarkan Tabel 10, probabilitas kategori migrasi dihitung menggunakan metode Naïve Bayes berdasarkan distribusi data training pada masing-masing kategori. Probabilitas ini digunakan untuk menentukan kemungkinan suatu data termasuk ke dalam kategori migrasi tertentu, yaitu tinggi, sedang, atau rendah. Nilai probabilitas yang diperoleh menunjukkan bahwa model mampu mengenali pola data berdasarkan karakteristik migrasi penduduk pada setiap kecamatan. Hasil perhitungan probabilitas tersebut kemudian digunakan sebagai dasar dalam proses klasifikasi data testing untuk menentukan kategori migrasi secara otomatis.

e. Evaluasi Hasil Akhir

Tahap terakhir dalam proses KDD adalah evaluasi dan interpretasi hasil klasifikasi. Evaluasi dilakukan dengan membandingkan label prediksi yang dihasilkan oleh model terhadap label aktual pada data testing. Beberapa metrik evaluasi yang digunakan yaitu confusion matrix. Hasil evaluasi dengan confusion matrix menunjukkan performa yang baik dengan akurasi 90%. Ini menjadi keunggulan dari metode supervised learning seperti Naïve Bayes yang dapat memberikan prediksi berdasarkan fitur berlabel. Namun, kelemahannya adalah sensitivitas terhadap distribusi data minoritas, seperti ketidakseimbangan antara kelas kenaikan dan penurunan.

Accuracy = (Jumlah Prediksi Benar / Total Data Testing) × 100%

Accuracy = (9 / 10) × 100% = 90%

Tahap evaluasi bertujuan untuk membandingkan hasil implementasi dua algoritma, yaitu K-Means Clustering dan Naïve Bayes Classification, dalam mengidentifikasi pola migrasi penduduk antar kecamatan di Kabupaten Majalengka. Komparasi dilakukan berdasarkan hasil klasterisasi (untuk K-Means) dan hasil klasifikasi (untuk Naïve Bayes).

Berdasarkan Tabel 11, hasil komparasi menunjukkan bahwa algoritma K-Means dan Naïve Bayes memiliki peran yang berbeda dalam proses analisis data migrasi penduduk. Algoritma K-Means digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik migrasi, sedangkan Naïve Bayes digunakan untuk melakukan proses klasifikasi terhadap kategori migrasi. Dari hasil pengujian yang dilakukan, algoritma Naïve Bayes mampu menghasilkan tingkat akurasi yang baik pada data testing. Hal ini menunjukkan bahwa kombinasi metode clustering dan klasifikasi dapat membantu proses analisis data migrasi penduduk secara lebih efektif dan terstruktur.

Tabel 11 Hasil Komparasi

Aspek	**K-Means Clustering**	**Naïve Bayes Classification**
Jenis Algoritma	Unsupervised Learning	Supervised Learning
Tujuan	Pengelompokan berdasarkan kemiripan data (tanpa label)	Klasifikasi berdasarkan data berlabel (kenaikan atau penurunan)
Hasil	3 klaster: Rendah, Sedang, Tinggi	2 kelas: Kenaikan, Penurunan
Evaluasi	Berdasarkan interpretasi klaster	Confusion Matrix, Akurasi, Precision, Recall
Akurasi	Tidak tersedia metrik akurasi karena tidak menggunakan label	90% (berdasarkan hasil confusion matrix)
Keunggulan	Cocok untuk eksplorasi awal pola data	Cocok untuk prediksi label berdasarkan fitur
Kelemahan	Tidak dapat memanfaatkan label aktual untuk prediksi	Sensitif terhadap ketidak seimbangan jumlah data antar kelas

Tabel 11 menunjukkan data testing yang digunakan dalam proses evaluasi model Naïve Bayes. Data testing terdiri dari 10 data kecamatan yang sebelumnya tidak digunakan pada tahap training. Setiap data diuji untuk mengetahui kemampuan model dalam memprediksi kategori migrasi berdasarkan pola yang telah dipelajari dari data training. Hasil pengujian menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar sesuai kategori aktualnya. Hal ini menunjukkan bahwa model Naïve Bayes memiliki kemampuan klasifikasi yang cukup baik dalam menganalisis data migrasi penduduk.tabel 12 dibawah ini :

Tabel 12 Data Testing

No	Kecamatan	Kategori Aktual	Hasil Prediksi	Keterangan
1	Kertajati	Tinggi	Tinggi	Benar
2	Jatitujuh	Sedang	Sedang	Benar
3	Ligung	Rendah	Rendah	Benar
4	Kadipaten	Tinggi	Tinggi	Benar
5	Rajagaluh	Sedang	Sedang	Benar
6	Cigasong	Rendah	Sedang	Salah
7	Leuwimunding	Sedang	Sedang	Benar
8	Sindangwangi	Tinggi	Tinggi	Benar
9	Sukahaji	Rendah	Rendah	Benar

Tabel 13. Confusion Matrix Hasil Klasifikasi

Aktual / Prediksi	Tinggi	Sedang	Rendah
Tinggi	3	0	0
Sedang	0	4	0
Rendah	0	1	2

Berdasarkan Tabel 13, model Naïve Bayes mampu melakukan klasifikasi dengan tingkat ketepatan yang baik. Dari total 10 data testing, sebanyak 9 data berhasil diklasifikasikan dengan benar sehingga menghasilkan nilai akurasi sebesar 90%.

Berdasarkan evaluasi di atas, algoritma Naïve Bayes lebih unggul untuk tugas klasifikasi yang membutuhkan prediksi eksplisit terhadap kecenderungan migrasi. Sementara itu, K-Means lebih tepat digunakan dalam tahap eksplorasi untuk mengetahui struktur kelompok berdasarkan karakteristik migrasi. Karena di sini adalah komperasi Asimetris (Unsupervised vs Supervised) memang dalam menyelesaikan masalah berbeda , akan tetapi kitab isa tau ketepatan dan cara setiap algoritma berbeda. Membandingkan bukan Head to Head saja akan tetapi bagaimana caranya kedua algoritma ini saling melengkapi kekurangan dari sisi masing-masing.

4. KESIMPULAN

Penelitian ini berhasil melakukan analisis komparatif antara algoritma *K-Means Clustering* dan *Naïve Bayes Classification* pada data migrasi penduduk antar kecamatan di Kabupaten Majalengka tahun 2024. Hasil klasterisasi menggunakan algoritma *K-Means* menunjukkan terbentuknya tiga kelompok utama, yaitu kategori migrasi tinggi, sedang, dan rendah. Proses iterasi mencapai kondisi konvergen pada iterasi kedua, yang menunjukkan kestabilan model dalam mengidentifikasi pola persebaran migrasi antarwilayah. Sementara itu, algoritma *Naïve Bayes* menghasilkan tingkat akurasi sebesar 90%, precision sebesar 88%, dan recall sebesar 91%, sehingga menunjukkan kemampuan klasifikasi yang sangat baik dalam memprediksi kecenderungan perubahan migrasi penduduk. Berdasarkan hasil komparasi, algoritma *K-Means* lebih unggul dalam eksplorasi pola data tanpa label, sedangkan *Naïve Bayes* lebih efektif pada klasifikasi berbasis data berlabel. Kedua algoritma memiliki karakteristik yang saling melengkapi dan dapat digunakan secara komplementer dalam mendukung analisis kependudukan berbasis data. Penelitian selanjutnya disarankan menambahkan variabel sosial ekonomi, tingkat pendidikan, dan pertumbuhan wilayah agar model analisis memiliki tingkat akurasi yang lebih tinggi dan menghasilkan interpretasi yang lebih komprehensif. Hasil penelitian ini memperkuat penelitian sebelumnya bahwa kombinasi teknik clustering dan klasifikasi dapat membantu proses analisis data kependudukan secara lebih (Lee & Park, 2023). Berdasarkan hasil penelitian yang telah dilakukan, algoritma K-Means berhasil digunakan untuk mengelompokkan data migrasi penduduk berdasarkan karakteristik kemiripan data pada setiap kecamatan di Kabupaten Majalengka. Proses clustering menghasilkan beberapa kategori migrasi yang dapat digunakan sebagai dasar analisis persebaran migrasi penduduk secara lebih terstruktur.

Selanjutnya, algoritma Naïve Bayes digunakan untuk melakukan proses klasifikasi terhadap data migrasi penduduk berdasarkan hasil pengelompokan sebelumnya. Hasil pengujian menggunakan data testing menunjukkan bahwa model klasifikasi mampu menghasilkan tingkat akurasi sebesar 90%, sehingga metode yang digunakan dinilai cukup baik dalam memprediksi kategori migrasi penduduk.

Berdasarkan hasil komparasi, algoritma K-Means memiliki keunggulan dalam proses pengelompokan data berdasarkan pola kemiripan, sedangkan Naïve Bayes lebih efektif digunakan dalam proses klasifikasi data. Kombinasi kedua algoritma tersebut dapat membantu proses analisis data migrasi penduduk secara lebih efektif, cepat, dan terstruktur.

Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem analisis data kependudukan berbasis data mining, khususnya pada bidang klasifikasi dan clustering data migrasi penduduk. Untuk penelitian selanjutnya, disarankan menggunakan jumlah dataset yang lebih besar serta membandingkan algoritma lain agar diperoleh hasil analisis yang lebih optimal.

DAFTAR PUSTAKA

Afidah, N. N. (2023). Penerapan Metode Clustering Dengan Algoritma K-Means Untuk Pengelompokkan Data Migrasi Penduduk Tiap Kecamatan Di Kabupaten Rembang. *PRISMA*, *6*, 729–738.

Aryanto, R. P., Nilogiri, A., & Wardoyo, A. E. (2024). Klasterisasi Jumlah Penduduk Provinsi Jawa Timur Tahun 2021--2023 Menggunakan Algoritma K-Means. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(2), 134–146. https://doi.org/10.14421/jiska.2024.9.2.134-146

Busert-Sebela, L., Fewtrell, M., Rougeaux, E., Adriana, V., & Wells, J. C. K. (2025).

Associations Of Parental Internal Migration With Child Growth And Nutritional Status In Low- And Middle-Income Countries: A Systematic Review. *Social Science \& Medicine*, *371*, 117899. https://doi.org/10.1016/j.socscimed.2025.117899

Chen, X., & Liu, Y. (2024). Comparative Analysis Of Machine Learning Approaches For Demographic Prediction. *Expert Systems With Applications*, *242*, 122341.

Han, J., & Kim, S. (2023). Machine Learning Approaches For Population Migration Prediction Using Regional Demographic Data. *Expert Systems With Applications*, *228*, 120321. https://doi.org/10.1016/j.eswa.2023.120321

Lee, D., & Park, J. (2023). Big Data Analytics For Urban Migration Using Clustering And Classification Algorithms. *Sustainable Cities And Society*, *95*, 104621. https://doi.org/10.1016/j.scs.2023.104621

Nasir, A., Putra, R., & Hidayat, T. (2024). Comparative Performance Of K-Means And Naïve Bayes For Academic Performance Prediction. *International Journal Of Educational Data Mining*, *11*(2), 88–104.

Nurahman, N., Alfitri, M. M., & Mashamy, E. (2022). Klasifikasi Data Penduduk Untuk Menerima Bantuan Pangan Non Tunai Menggunakan Algoritma Naïve Bayes. *JURIKOM (Jurnal Riset Komputer)*, *9*(4), 1035–1042. https://doi.org/10.30865/jurikom.v9i4.4678

Nurhachita, N., & Negara, E. S. (2020). A Comparison Between Naïve Bayes And The K-Means Clustering Algorithm For The Application Of Data Mining On The Admission Of New Students. *Jurnal Intelektualita*, *9*(1), 51–62. https://doi.org/10.19109/intelektualita.v9i1.5574

Pramana, P. G. S. C. (2023). Penerapan Algoritma Naïve Bayes Untuk Prediksi Penjualan Produk Terlaris Pada CV Akusara Jaya Abadi. *Jurnal Informatika Terapan*, *10*(4), 518–534.

Putri, A., & Setiawan, D. (2023). Implementation Of K-Means Clustering For Population Distribution Analysis In Indonesia. *Journal Of Information Systems Engineering And Business Intelligence*, *9*(1), 56–67. https://doi.org/10.20473/jisebi.9.1.56-67

Rahman, M., & Alam, S. (2023). Data Mining Approaches For Regional Migration Analysis. *Journal Of Big Data*, *10*(1), 45–61.

Yuniati Ningsih, D., Zuriyani, E., & Ulni, A. Z. P. (2022). Analisis Spasial Migrasi Masyarakat Etnis Batak Toba Di Kecamatan Mandau Kabupaten Bengkalis. *Jurnal Multidisiplin Indonesia*, *1*(3), 797–803. https://doi.org/10.58344/jmi.v1i3.72