JAISI: Journal of Advanced Informatics and Sustainable Intelligent Systems
Vol. 1 No. 2 Tahun 2026 , e-ISSN (23164-2268) https://doi.org/10.23960/jaisi.v1i2.46

Analisis Disparitas Pendidikan Di Indonesia Menggunakan Clustering Dan Classification Berbasis Data Geospasial

(Analysis of Educational Disparity in Indonesia Using Geospatial Data-Based Clustering and Classification)

Trisya Septiana[1*], Rio Ariestia Pradipta[2], Nadjwa Tasya Safira[3], Erlin Sari Ramadhani[4]

[1,2,3,4]*Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung, Jl. Prof. Dr. Sumantri Brojonegoro No. 1, Bandar Lampung 35145, Indonesia*

Riwayat artikel:
Submitted : 03/01/2026
Published : 02/04/2026

*** Email Korespondensi:**
trisya.septiana@eng.unila.ac.id

Abstrak: Pemerataan akses pendidikan di Indonesia masih menghadapi persoalan spasial, demografis, dan kelembagaan yang tidak dapat dijelaskan hanya melalui jumlah sekolah secara agregat. Penelitian ini bertujuan menganalisis pola disparitas pendidikan Indonesia menggunakan pendekatan data mining berbasis data geospasial. Metode penelitian mengikuti tahapan CRISP-DM, mulai dari pemahaman masalah, eksplorasi data, pembersihan data, rekayasa fitur, pemodelan, evaluasi, hingga penyusunan rekomendasi. Dataset yang digunakan adalah Indonesia School Dataset with Province Data yang memuat atribut lokasi sekolah, jenjang pendidikan, status kepemilikan, luas provinsi, jumlah penduduk, dan penduduk usia sekolah. Proses analisis menerapkan K-Means Clustering untuk segmentasi wilayah serta Random Forest Classifier untuk memprediksi status sekolah negeri atau swasta. Hasil penelitian menunjukkan bahwa jumlah cluster optimal adalah empat cluster dengan Silhouette Score sebesar 0,414. Model Random Forest memperoleh akurasi 0,83 dengan precision kelas negeri 0,88, recall 0,91, dan F1-score 0,89. Feature importance menunjukkan longitude, latitude, dan school level sebagai faktor paling dominan. Temuan ini menegaskan bahwa disparitas pendidikan Indonesia memiliki pola geospasial yang kuat dan memerlukan kebijakan pemerataan berbasis data.

Kata kunci: CRISP-DM; data mining pendidikan; disparitas pendidikan; K-Means; Random Forest

Abstract: Equitable access to education in Indonesia remains shaped by spatial, demographic, and institutional disparities that cannot be adequately explained by aggregate school counts alone. This study aims to analyze educational disparity patterns in Indonesia through a geospatial data mining approach. The research follows the CRISP-DM stages, including problem understanding, data exploration, data cleaning, feature engineering, modeling, evaluation, and recommendation formulation. The dataset used is the Indonesia School Dataset with Province Data, which contains school location, education level, ownership status, province area, total population, and school-age population attributes. The analysis applies K-Means Clustering to segment regions and Random Forest Classifier to predict public or private school status. The results show that the optimal number of clusters is four, with a Silhouette Score of 0.414. The Random Forest model achieves 0.83 accuracy, with precision of 0.88, recall of 0.91, and F1-score of 0.89 for the public school class. Feature importance indicates that longitude, latitude, and school level are the most influential variables. These findings confirm that Indonesian educational disparity has a strong geospatial pattern and requires data-driven equity policy.

Keywords: CRISP-DM; educational data mining; educational disparity; K-Means; Random Forest

1. PENDAHULUAN

Pemerataan pendidikan merupakan isu strategis karena kualitas pembangunan manusia sangat dipengaruhi oleh keterjangkauan layanan pendidikan dasar dan menengah. Dalam kerangka SDG 4, pendidikan yang inklusif dan bermutu menuntut tersedianya kesempatan belajar yang merata bagi seluruh penduduk, bukan hanya pada wilayah yang secara ekonomi dan geografis lebih maju [1]. Bagi negara kepulauan seperti Indonesia, pemerataan tersebut tidak sederhana karena sekolah tersebar dalam ruang geografis yang luas, kepadatan penduduk antarwilayah tidak seimbang, dan kapasitas penyelenggaraan sekolah negeri maupun swasta berbeda pada setiap provinsi.

Ketimpangan pendidikan tidak hanya dapat dibaca dari banyak atau sedikitnya jumlah sekolah. Pada wilayah urban padat, jumlah sekolah dapat terlihat tinggi, tetapi akses terhadap sekolah negeri dapat terbatas karena kompetisi ruang, biaya, dan dominasi penyelenggara swasta. Sebaliknya, wilayah rural dan Indonesia bagian timur dapat memiliki proporsi sekolah negeri lebih tinggi, tetapi sekolahnya tersebar jarang sehingga hambatan jarak tetap menjadi persoalan. Beberapa studi pendidikan di Indonesia menegaskan bahwa kondisi geografis, infrastruktur, kualitas layanan,

dan sumber daya wilayah berkaitan erat dengan kesenjangan akses pendidikan [2]-[7].

Perkembangan data mining membuka peluang untuk membaca persoalan pemerataan pendidikan secara lebih terukur. Data sekolah yang mengandung atribut lokasi, status kepemilikan, jenjang pendidikan, luas wilayah, dan jumlah penduduk dapat diolah untuk menemukan pola tersembunyi yang tidak mudah ditangkap melalui tabulasi biasa. Dengan pendekatan clustering, wilayah dapat dikelompokkan berdasarkan kemiripan karakteristik pendidikan. Dengan pendekatan classification, status sekolah dapat diprediksi berdasarkan kombinasi fitur geospasial dan demografis.

Penelitian ini mengubah laporan proyek data mining menjadi artikel ilmiah dengan fokus pada analisis disparitas pendidikan di Indonesia menggunakan K-Means Clustering dan Random Forest Classification. Kontribusi utama artikel ini terletak pada integrasi dua pendekatan pembelajaran mesin: K-Means digunakan untuk memetakan karakteristik wilayah pendidikan secara tidak terawasi, sedangkan Random Forest digunakan untuk menjelaskan faktor prediktif status sekolah. Tujuan penelitian ini adalah mengidentifikasi pola spasial disparitas pendidikan, mengevaluasi performa model clustering dan classification, serta merumuskan implikasi kebijakan pemerataan pendidikan berbasis data geospasial.

2. BAHAN DAN METODE PENELITIAN

2.1 Data Penelitian

Data yang digunakan berasal dari Indonesia School Dataset with Province Data yang tersedia pada Kaggle dan disusun dari data sekolah Indonesia yang diperkaya dengan data provinsi dari Badan Pusat Statistik [8], [9]. Dataset memuat atribut province_name, city_name, district_name, school_name, stage, status, lat, long, province_area, total_population, dan total_education_age_population. Atribut tersebut merepresentasikan dimensi administratif, kelembagaan, geospasial, dan demografis yang relevan untuk analisis pemerataan pendidikan.

Laporan proyek mencatat bahwa data awal berjumlah 215.371 baris, sedangkan data analitik setelah preprocessing berjumlah 214.896 baris. Perbedaan ini diperlakukan sebagai hasil penyelarasan teknis terhadap data yang tidak lengkap, sementara missing value koordinat sebanyak 1.150 baris ditangani melalui imputasi rata-rata koordinat per provinsi. Fokus analisis tidak diarahkan pada evaluasi mutu pembelajaran, tetapi pada pola distribusi sekolah dan kecenderungan status kepemilikan sekolah berdasarkan indikator spasial dan demografis.

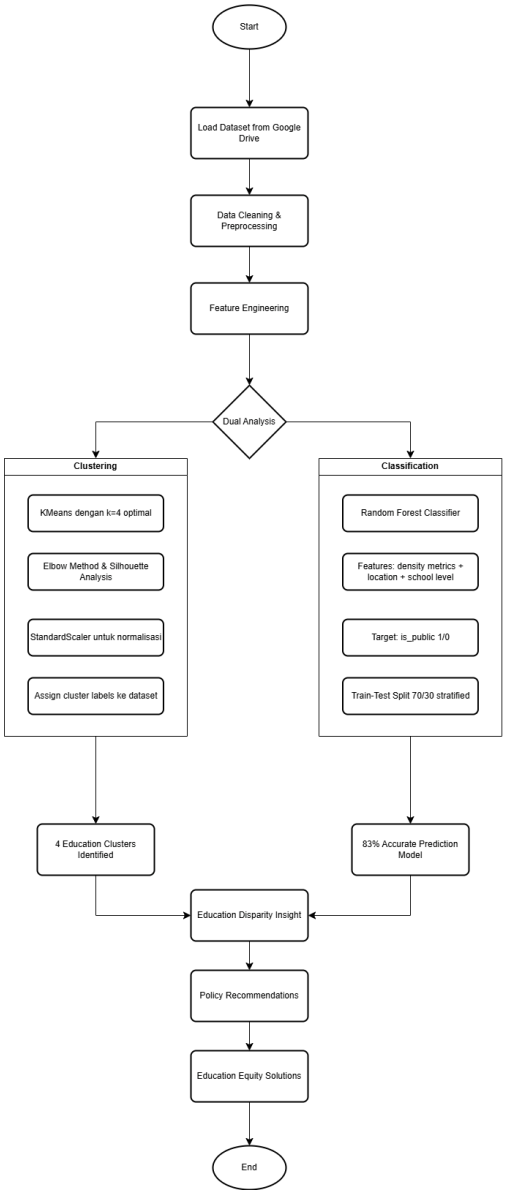
Tabel 1. Atribut utama dataset penelitian

Kelompok atribut	Nama atribut	Fungsi analisis
Administratif	province_name, city_name, district_name	Menjelaskan lokasi administratif sekolah
Identitas sekolah	school_name, stage, status	Menjelaskan jenjang dan kepemilikan sekolah
Geospasial	lat, long	Mewakili posisi geografis sekolah
Demografis	province_area, total_population	Menghitung kepadatan penduduk
Pendidikan	total_education_ age_population	Menghitung rasio dan kepadatan penduduk usia sekolah

2.2 Metodologi CRISP-DM

Penelitian mengikuti Cross Industry Standard Process for Data Mining (CRISP-DM), yaitu kerangka kerja data mining yang menempatkan pemahaman masalah sebagai tahap awal sebelum pemodelan teknis dilakukan [10]. Pemilihan CRISP-DM relevan karena penelitian ini tidak hanya mengejar akurasi model, tetapi juga menafsirkan hasil analisis dalam konteks pemerataan pendidikan. Enam fase yang digunakan adalah business understanding, data understanding, data preparation, modeling, evaluation, dan deployment.

Pada fase business understanding, masalah didefinisikan sebagai ketimpangan distribusi sekolah antarwilayah Indonesia. Pada fase data understanding, struktur dataset, missing value, distribusi jenjang sekolah, dan variasi status kepemilikan dianalisis. Pada fase data preparation, dilakukan pembersihan data, imputasi koordinat, rekayasa fitur, encoding, dan standardisasi. Pada fase modeling, dua algoritma diterapkan, yaitu K-Means dan Random Forest. Pada fase evaluation, kualitas cluster dan performa classification diuji. Pada fase deployment, hasil model diterjemahkan menjadi insight dan rekomendasi kebijakan.



Gambar 1. Alur metodologi CRISP-DM yang digunakan dalam penelitian

2.3 Preprocessing dan Rekayasa Fitur

Tahap preprocessing diawali dengan imputasi nilai kosong pada lat dan long menggunakan rata-rata koordinat setiap provinsi. Strategi ini dipilih agar observasi tetap dapat dipertahankan tanpa menghilangkan representasi wilayah. Kolom kategorikal seperti stage, status, province_name, dan city_name yang kosong diberi label Unknown untuk mencegah error saat proses encoding. Data duplikat dan ketidakkonsistenan format nama wilayah diperiksa sebelum model dijalankan.

Rekayasa fitur dilakukan untuk menghasilkan indikator yang lebih informatif daripada variabel mentah.

a. Population density (menggambarkan tingkar kepadatan penduduk dalam suatu provinsi:
b. Education Density (mengukur Jumlah Penduduk usia sekola per luas wilayah provinsi
c. Education ratio (menunjukkan proporsi penduduk usia sekolah terhadap populasi total
d. School level encoding (Jenjang sekolah dikodekan dengan skema
   SD = 1, SMP = 2, SMA/SMK = 3, dan SLB = 4.
e. Status sekolah dikodekan menjadi is_public, yaitu 1 untuk negeri dan 0 untuk swasta.

2.4 Pemodelan dan Evaluasi

2.4.1 Implementasi K-Means Clustering

Implementasi K-Means Clustering dilakukan setelah seluruh fitur numerik berada pada skala yang seragam. Fitur yang digunakan meliputi population_density, education_density, education_ratio, school_level, is_public, dan province_encoded. Keenam fitur tersebut dipilih karena mewakili tekanan demografis, proporsi penduduk usia sekolah, jenjang pendidikan, status kepemilikan sekolah, serta konteks wilayah provinsi. Sebelum proses clustering, data ditransformasi menggunakan StandardScaler agar variabel dengan nilai besar, seperti kepadatan penduduk, tidak mendominasi perhitungan jarak Euclidean [11], [12].

Penentuan jumlah cluster dilakukan secara bertahap dengan membandingkan nilai inertia pada beberapa kandidat k dan menilai koherensi antarcluster menggunakan Silhouette Score. Berdasarkan laporan UTS, kombinasi Elbow Method dan Silhouette Analysis menunjukkan bahwa k = 4 merupakan pilihan yang paling sesuai, dengan Silhouette Score sebesar 0,414. Nilai tersebut diperlakukan sebagai indikasi bahwa data memiliki pemisahan cluster yang cukup stabil, meskipun karakteristik sosial pendidikan antarwilayah tetap memiliki variasi internal yang tinggi [5], [13].

Setelah model K-Means dilatih, label cluster 0 sampai 3 dikembalikan ke dataset utama untuk dianalisis bersama atribut wilayah. Tahap ini penting karena angka cluster tidak memiliki makna substantif sebelum diprofilkan. Oleh karena itu, setiap cluster kemudian dibaca melalui rata-rata population_density, rasio sekolah negeri, contoh provinsi dominan, dan karakteristik distribusi sekolah. Dengan cara ini, implementasi clustering tidak hanya menghasilkan kelompok statistik, tetapi juga membentuk dasar interpretasi tentang pola disparitas pendidikan.

2.4.2 Implementasi Random Forest Classification

Implementasi classification menggunakan Random Forest Classifier dengan target is_public, yaitu 1 untuk sekolah negeri dan 0 untuk sekolah swasta. Fitur input yang digunakan adalah

population_density, education_density, education_ratio, school_level, province_encoded, lat, dan long. Penggunaan lat dan long dimaksudkan untuk menangkap pola spasial yang tidak selalu tampak melalui nama provinsi saja, sedangkan school_level digunakan untuk melihat perbedaan status kepemilikan pada jenjang SD, SMP, SMA/SMK, dan SLB.

Dataset dibagi menjadi 70% data latih dan 30% data uji dengan random_state = 42. Pembagian data dilakukan secara stratified agar proporsi sekolah negeri dan swasta tetap relatif seimbang pada data latih dan data uji. Model dibangun dengan n_estimators = 100 sehingga prediksi akhir diperoleh melalui voting dari sejumlah decision tree. Pendekatan ensemble ini dipilih karena lebih stabil dibandingkan satu pohon keputusan tunggal dan mampu menangkap hubungan nonlinier antara faktor geospasial, demografis, dan status sekolah [14].

Evaluasi classification dilakukan menggunakan accuracy, precision, recall, F1-score, dan confusion matrix. Accuracy digunakan untuk melihat ketepatan model secara umum, sedangkan precision dan recall diperlukan untuk membaca apakah model terlalu banyak salah mengklasifikasikan sekolah negeri atau gagal mengenali sekolah negeri. Selain itu, feature importance digunakan untuk menafsirkan variabel yang paling berpengaruh terhadap prediksi. Dengan demikian, model tidak hanya dipakai sebagai alat prediksi, tetapi juga sebagai alat interpretasi untuk memahami struktur ketimpangan status sekolah.

Tabel 2. Rancangan model dan metrik evaluasi

Model	Tujuan	Input utama	Metrik / parameter
K-Means	Segmentasi wilayah berdasarkan kemiripan indikator pendidikan dan demografi	population_density; education_density; education_ratio; school_level; is_public; province_encoded	StandardScaler; k = 4; Elbow; Silhouette Score
Random Forest	Prediksi status sekolah negeri atau swasta	population_density; education_density; education_ratio; school_level; province_encoded; lat; long	70:30 split; n_estimators = 100; accuracy; precision; recall; F1-score

3. HASIL DAN PEMBAHASAN

3.1 Hasil Preprocessing

Tahap preprocessing menghasilkan dataset analitik yang siap digunakan untuk pemodelan. Nilai kosong pada koordinat lat dan long berhasil dikurangi dari 1.150 menjadi 0 melalui imputasi rata-rata per provinsi. Proses ini penting karena koordinat geografis menjadi variabel utama dalam classification dan visualisasi spasial. Kolom kategorikal yang dibutuhkan model juga telah distandarisasi sehingga seluruh data dapat dikonversi menjadi representasi numerik.

Hasil rekayasa fitur memperlihatkan bahwa setiap sekolah tidak lagi hanya direpresentasikan sebagai entitas administratif, tetapi juga sebagai observasi yang membawa informasi wilayah. Population density menunjukkan tekanan demografis, education density menunjukkan tekanan penduduk usia sekolah terhadap ruang wilayah, dan education ratio menunjukkan proporsi populasi usia sekolah. Dengan fitur tersebut, data menjadi lebih sesuai untuk membaca disparitas pendidikan sebagai fenomena spasial dan struktural.

3.2 Hasil K-Means Clustering

Hasil Elbow Method dan Silhouette Analysis menunjukkan bahwa jumlah cluster optimal adalah k = 4. Nilai Silhouette Score sebesar 0,414 menunjukkan bahwa pemisahan cluster cukup stabil untuk data sosial pendidikan yang bersifat heterogen. Nilai tersebut tidak menunjukkan pemisahan yang sempurna, tetapi cukup untuk membaca kecenderungan wilayah karena data pendidikan nasional memang memiliki variasi internal yang tinggi.

Dari sisi implementasi, pemilihan k = 4 tidak dilakukan dengan membaca angka inertia saja. Inertia cenderung selalu turun ketika nilai k bertambah, sehingga perlu dikombinasikan dengan Silhouette Score agar jumlah cluster tidak terlalu banyak dan tetap dapat ditafsirkan secara substantif. Dalam penelitian ini, empat cluster dipandang cukup karena mampu membedakan pola urban padat, semi-urban, wilayah dominan negeri, dan metropolitan dengan kepadatan ekstrem.

Empat cluster yang terbentuk menunjukkan pola pendidikan yang berbeda. Cluster 0 merepresentasikan wilayah urban padat dengan dominasi sekolah swasta, terutama Jawa Barat dan Banten. Cluster 1 merepresentasikan wilayah semi-urban dengan distribusi negeri dan swasta yang lebih seimbang, seperti Sumatera Utara dan Kalimantan Barat. Cluster 2 menunjukkan wilayah yang relatif homogen dengan dominasi sekolah negeri, seperti Jawa Tengah. Cluster 3 menggambarkan wilayah metropolitan dengan kepadatan ekstrem, terutama DKI Jakarta, dengan komposisi status sekolah yang lebih campuran.

Label cluster kemudian digunakan sebagai variabel baru untuk membaca distribusi sekolah pada level wilayah. Cluster 0 dan cluster 3 menunjukkan bahwa wilayah padat tidak selalu identik dengan pemerataan akses, karena kepadatan sekolah dapat berjalan bersama dominasi sekolah swasta atau komposisi

kepemilikan yang campuran. Sebaliknya, cluster 1 dan cluster 2 menunjukkan bahwa dominasi sekolah negeri tetap perlu dibaca bersama kepadatan penduduk dan luas wilayah, sebab wilayah dengan rasio negeri tinggi belum tentu memiliki keterjangkauan geografis yang baik.

Tabel 3. Ringkasan hasil K-Means Clustering

Cluster	Contoh wilayah	Kepadatan	Rasio negeri	Interpretasi
0	Jawa Barat, Banten	722	0,01	Urban padat, dominan swasta
1	Sumatera Utara, Kalbar	147	0,94	Semi-urban, relatif seimbang
2	Jawa Tengah	838	1,00	Homogen, dominan negeri
3	DKI Jakarta	16.084	0,47	Metropolitan, kepadatan ekstrem

Pola clustering memperlihatkan bahwa disparitas pendidikan tidak identik dengan kekurangan jumlah sekolah saja. Pada wilayah urban padat, sekolah lebih banyak, tetapi komposisinya cenderung didorong oleh sekolah swasta. Hal ini menunjukkan adanya ketimpangan akses terhadap sekolah negeri. Pada wilayah yang lebih luas dan kurang padat, sekolah negeri menjadi tulang punggung layanan pendidikan, tetapi tantangan utamanya adalah keterjangkauan fisik dan sebaran sekolah yang jarang. Dengan demikian, interpretasi cluster harus selalu dikaitkan dengan struktur ruang dan kependudukan wilayah.

Hasil implementasi clustering juga menghasilkan empat luaran analitis. Pertama, label cluster untuk setiap observasi sekolah. Kedua, ringkasan karakteristik tiap cluster. Ketiga, dasar visualisasi spasial untuk melihat konsentrasi wilayah. Keempat, dasar rekomendasi kebijakan untuk membedakan kebutuhan wilayah urban, semi-urban, rural, dan metropolitan. Keempat luaran tersebut membuat clustering berfungsi sebagai instrumen pemetaan ketimpangan, bukan hanya sebagai proses pengelompokan otomatis.

Secara implementatif, proses K-Means dijalankan pada data yang telah distandarisasi agar setiap atribut memiliki kontribusi yang seimbang terhadap perhitungan jarak. Tanpa standardisasi, variabel dengan skala besar seperti kepadatan penduduk berpotensi mendominasi proses pengelompokan dan membuat variabel lain, seperti rasio usia sekolah atau status kepemilikan, kurang terbaca oleh model.

Fitur population_density dan education_density digunakan untuk menangkap tekanan demografis terhadap ketersediaan layanan pendidikan, sedangkan education_ratio menggambarkan proporsi penduduk usia sekolah terhadap total populasi. Fitur school_level dan is_public membantu model membaca komposisi jenjang serta kepemilikan sekolah. Dengan kombinasi tersebut, cluster yang terbentuk tidak hanya mencerminkan wilayah padat atau tidak padat, tetapi juga memperlihatkan struktur penyediaan layanan pendidikan pada masing-masing kelompok wilayah.

Interpretasi cluster dilakukan setelah label 0 sampai 3 diperoleh dari model. Angka cluster tidak diperlakukan sebagai urutan kualitas, melainkan sebagai tanda kelompok karakteristik. Oleh karena itu, analisis difokuskan pada profil centroid, dominasi wilayah, rasio sekolah negeri, dan tekanan kepadatan. Cara baca ini penting agar hasil clustering tidak disederhanakan menjadi peringkat baik-buruk, tetapi dipahami sebagai pola spasial yang membutuhkan strategi kebijakan berbeda.

3.3 Hasil Random Forest Classification

Model Random Forest digunakan untuk memprediksi status sekolah negeri atau swasta. Hasil evaluasi menunjukkan akurasi sebesar 0,83. Untuk kelas sekolah negeri, model memperoleh precision 0,88, recall 0,91, dan F1-score 0,89. Nilai tersebut menunjukkan bahwa model cukup baik dalam mengenali sekolah negeri, sekaligus mampu menjaga keseimbangan prediksi pada data yang heterogen.

Pada tahap training, model mempelajari hubungan antara fitur geospasial-demografis dan label status sekolah. Prediksi yang dihasilkan bukan sekadar menebak kategori negeri atau swasta, tetapi membaca pola sistematis dalam data. Misalnya, koordinat geografis membantu model mengenali konsentrasi sekolah swasta pada wilayah perkotaan padat, sedangkan school_level membantu membedakan kecenderungan status kepemilikan antarjenjang pendidikan.

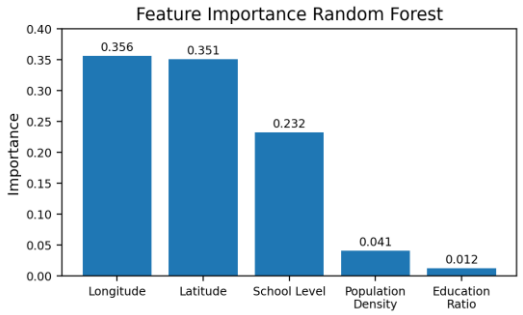
Feature importance memberikan informasi yang lebih substantif dibandingkan akurasi semata. Longitude memiliki importance tertinggi sebesar 0,356, diikuti latitude sebesar 0,351 dan school_level sebesar 0,232. Population_density dan education_ratio memiliki pengaruh yang lebih kecil. Temuan ini menunjukkan bahwa status sekolah sangat dipengaruhi oleh posisi geografis dan jenjang pendidikan. Dengan kata lain, distribusi sekolah negeri dan swasta di Indonesia memiliki pola spasial yang kuat.

Dominasi longitude dan latitude menunjukkan bahwa status sekolah memiliki pola keruangan yang kuat. Artinya, peluang suatu sekolah berstatus negeri atau swasta tidak berdiri sendiri, tetapi berkaitan dengan lokasi wilayah, kepadatan penduduk, dan sejarah perkembangan layanan pendidikan. Sementara itu, pengaruh school_level menunjukkan bahwa distribusi sekolah negeri dan swasta tidak sama pada setiap jenjang. Temuan ini memperkuat bahwa implementasi classification

perlu dibaca sebagai bagian dari analisis disparitas, bukan hanya sebagai pengujian akurasi model.

Tabel 4. Evaluasi model Random Forest

Metrik	Nilai	Makna
Accuracy	0,83	Model mampu mengklasifikasikan status sekolah secara baik
Precision negeri	0,88	Sebagian besar prediksi negeri benar
Recall negeri	0,91	Sebagian besar sekolah negeri berhasil dikenali
F1-score negeri	0,89	Keseimbangan precision dan recall baik



Gambar 2. Feature importance pada model Random Forest

Pada classification, target prediksi adalah status sekolah yang telah dikodekan menjadi is_public. Nilai 1 menunjukkan sekolah negeri, sedangkan nilai 0 menunjukkan sekolah swasta. Model tidak menggunakan status asli sebagai fitur masukan, tetapi mempelajari hubungan antara karakteristik wilayah, koordinat geografis, jenjang sekolah, dan kecenderungan status kepemilikan sekolah.

Random Forest dipilih karena mampu menangkap hubungan nonlinier dan interaksi antarfaktor yang sulit dijelaskan oleh model tunggal. Dalam konteks data pendidikan, pengaruh lokasi, kepadatan penduduk, dan jenjang sekolah tidak selalu bekerja secara linear. Suatu wilayah dapat memiliki kepadatan tinggi, tetapi pola kepemilikan sekolahnya tetap berbeda karena faktor urbanisasi, sejarah pertumbuhan sekolah, dan kapasitas layanan publik.

Pembacaan confusion matrix dilakukan untuk melihat apakah model cenderung bias terhadap salah satu kelas. Accuracy sebesar 0,83 menunjukkan performa umum yang baik, tetapi precision, recall, dan F1-score tetap diperlukan agar penilaian tidak hanya bergantung pada satu angka. Precision negeri sebesar 0,88 menunjukkan ketepatan prediksi sekolah negeri, sedangkan recall negeri sebesar 0,91 menunjukkan kemampuan model mengenali sebagian besar sekolah negeri dalam data uji.

Feature importance memperlihatkan bahwa longitude, latitude, dan school_level menjadi kontributor utama. Temuan ini menegaskan bahwa status sekolah memiliki dimensi spasial yang kuat. Dalam artikel ini, koordinat tidak hanya dipahami sebagai titik lokasi, tetapi sebagai representasi kedekatan wilayah dengan pusat pertumbuhan, akses layanan, serta pola sebaran lembaga pendidikan negeri dan swasta.

3.4 Integrasi Hasil Clustering dan Classification

K-Means dan Random Forest memberikan dua lapisan pembacaan yang saling melengkapi. K-Means menjelaskan segmentasi wilayah pada level makro, sedangkan Random Forest menjelaskan faktor prediktif pada level observasi sekolah. Konsistensi keduanya terlihat pada dominasi variabel lokasi. Cluster wilayah urban dan metropolitan sangat berkaitan dengan kepadatan dan koordinat geografis, sementara model classification juga menempatkan longitude dan latitude sebagai faktor paling berpengaruh.

Secara metodologis, clustering berperan sebagai alat eksploratif untuk menemukan struktur kelompok tanpa label awal, sedangkan classification berperan sebagai alat prediktif untuk menguji apakah fitur wilayah mampu menjelaskan status sekolah. Perbedaan fungsi ini membuat keduanya saling melengkapi: clustering menjawab pertanyaan “wilayah seperti apa yang memiliki karakteristik serupa?”, sedangkan classification menjawab pertanyaan “fitur apa yang paling menentukan status sekolah?”.

Hasil tersebut memperlihatkan bahwa kebijakan pemerataan pendidikan perlu berpindah dari pendekatan administratif umum menuju pendekatan spasial berbasis data. Pembangunan sekolah baru tidak cukup diarahkan berdasarkan jumlah penduduk agregat, tetapi perlu mempertimbangkan jarak antar sekolah, dominasi status sekolah, kepadatan penduduk usia sekolah, serta pola pertumbuhan wilayah. Pada wilayah barat Indonesia yang padat dan lebih kompetitif, isu pentingnya adalah keseimbangan akses negeri dan swasta. Pada wilayah timur dan rural, isu pentingnya adalah perluasan infrastruktur, keterjangkauan transportasi, dan penguatan sekolah negeri sebagai penyedia layanan utama.

Untuk memperjelas hubungan antara dua model yang digunakan, integrasi hasil clustering dan classification dirangkum pada Tabel 5. Tabel ini menunjukkan bagaimana cluster wilayah, prediksi status sekolah, dan feature importance dapat digunakan bersama untuk menyusun rekomendasi berbasis data.

Tabel 5. Integrasi hasil clustering dan classification

Aspek	Temuan gabungan	Makna
Segmentasi	K-Means membentuk 4 cluster wilayah.	Menentukan tipe wilayah prioritas.
Prediksi	Random Forest memprediksi status negeri/swasta.	Membaca faktor pembeda status sekolah.
Spasial	Latitude dan longitude dominan dalam model.	Disparitas memiliki pola keruangan.
Kebijakan	Cluster dan prediksi dibaca bersama.	Rekomendasi lebih terarah dan operasional.

Berdasarkan integrasi tersebut, hasil clustering berfungsi sebagai peta segmentasi awal, sedangkan classification memperkuat pembacaan faktor penentu. Kombinasi keduanya membuat rekomendasi tidak berhenti pada deskripsi cluster, tetapi bergerak menuju penentuan prioritas intervensi pendidikan yang lebih operasional.

3.5 Implikasi Kebijakan

Implikasi pertama adalah perlunya dashboard pemerataan pendidikan nasional yang mengintegrasikan data sekolah, data penduduk, dan koordinat spasial. Dashboard semacam ini dapat membantu pemerintah pusat dan daerah memantau wilayah dengan tekanan akses tinggi secara berkala. Implikasi kedua adalah perlunya prioritisasi pembangunan dan revitalisasi sekolah berdasarkan indikator geospasial, bukan hanya berdasarkan usulan administratif. Implikasi ketiga adalah perlunya evaluasi keseimbangan peran sekolah negeri dan swasta, terutama pada wilayah urban yang memiliki kepadatan tinggi.

Hasil penelitian juga mengindikasikan pentingnya memisahkan strategi wilayah urban dan rural. Wilayah urban memerlukan kebijakan afirmatif agar akses terhadap sekolah negeri tetap tersedia bagi kelompok sosial ekonomi yang lebih rentan. Wilayah rural memerlukan kebijakan konektivitas, penguatan sekolah kecil, dan pengadaan fasilitas dasar agar jarak tidak menjadi hambatan utama. Dengan demikian, model data mining tidak menggantikan analisis kebijakan, tetapi menyediakan dasar kuantitatif untuk memperbaiki ketepatan intervensi pendidikan.

Tabel 6. Rekomendasi berbasis hasil analisis

Masalah utama	Rekomendasi kebijakan
Dominasi swasta dan keterbatasan akses negeri.	Menambah kapasitas sekolah negeri dan mengatur distribusi fasilitas.
Distribusi relatif seimbang tetapi rawan tekanan penduduk.	Monitoring pertumbuhan penduduk usia sekolah secara berkala.
Sebaran sekolah jarang dan akses fisik sulit.	Prioritas infrastruktur, transportasi, dan sekolah layanan dasar.
Data pendidikan belum selalu dibaca secara spasial.	Mengembangkan dashboard pemerataan pendidikan berbasis geospasial.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan data mining berbasis CRISP-DM dapat digunakan untuk membaca disparitas pendidikan Indonesia secara lebih sistematis. K-Means Clustering menghasilkan empat cluster wilayah dengan karakteristik berbeda, mulai dari urban padat yang didominasi sekolah swasta hingga wilayah metropolitan dengan kepadatan ekstrem. Nilai Silhouette Score sebesar 0,414 menunjukkan bahwa struktur cluster cukup stabil untuk menggambarkan pola pendidikan nasional yang heterogen.

Random Forest Classifier menghasilkan akurasi 0,83 dengan performa kelas negeri yang baik, yaitu precision 0,88, recall 0,91, dan F1-score 0,89. Feature importance menunjukkan bahwa longitude, latitude, dan school_level merupakan faktor paling dominan dalam memprediksi status sekolah. Temuan ini menegaskan bahwa kesenjangan pendidikan memiliki dimensi geospasial yang kuat. Oleh karena itu, pemerataan pendidikan sebaiknya dirancang melalui kebijakan berbasis data yang mempertimbangkan lokasi, kepadatan, komposisi status sekolah, dan keterjangkauan layanan pendidikan.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Program Studi Teknik Informatika Universitas Lampung serta seluruh pihak yang telah mendukung penyusunan proyek data mining ini.

DAFTAR PUSTAKA

[1] UNESCO, Education 2030: Incheon Declaration and Framework for Action for the Implementation of Sustainable Development Goal 4. Paris, France: UNESCO, 2016.

[2] T. Muttaqin, “Determinants of unequal access to and quality of education in Indonesia,” The Indonesian Journal of Development Planning, vol. 2, no. 1, pp. 1-23, 2018.

[3] A. Hakim, “Pemerataan akses pendidikan bagi rakyat sesuai UU Sisdiknas No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional,” Jurnal EduTech, vol. 2, no. 1, pp. 45-54, 2016.

[4] Z. Mustakim and R. Kamal, “K-Means clustering for classifying the quality management of secondary education in Indonesia,” Jurnal Cakrawala Pendidikan, vol. 40, no. 3, pp. 725-743, 2021.

[5] M. R. Putri, G. S. Nugraha, and R. Dwiyansaputra, “Pengelompokan provinsi di Indonesia berdasarkan indikator pendidikan menggunakan metode K-Means clustering,” J-COSINE: Journal of Computer Science and Informatics Engineering, vol. 7, no. 2, pp. 45-58, 2023.

[6] N. Putri, P. S. Dewi, and N. Pradnyawathi, “Pengaruh kondisi geografis terhadap pendidikan dasar di Indonesia: Studi literatur pada SD/MI,” Jurnal Widya Balina, vol. 9, no. 2, pp. 5-8, 2024.

[7] R. Wijayati, D. Damanik, and A. Prawirosastro, “Kesenjangan akses pendidikan di daerah terpencil: Analisis kebijakan dan alternatif solusi,” Jurnal Cahaya Mandalika, vol. 5, no. 1, pp. 671-678, 2025.

[8] M. Tridyo, “Indonesia School Dataset with Province Data,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/marchotridyo/indonesia-school-dataset-with-province-data

[9] Badan Pusat Statistik, “Penduduk, laju pertumbuhan penduduk, distribusi persentase penduduk, kepadatan

penduduk, rasio jenis kelamin penduduk menurut provinsi, 2024," BPS, 2025. [Online]. Available: https://www.bps.go.id/

[10] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 2000, pp. 29-39.

[11] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, Berkeley, CA, USA, 1967, pp. 281-297.

[13] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987.

[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[15] C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, pp. 357-362, 2020.

[16] W. McKinney, "Data structures for statistical computing in Python," in Proc. 9th Python in Science Conf., Austin, TX, USA, 2010, pp. 56-61.

[17] J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.