Digital Transformation Technology (Digitech) | e-ISSN : 2807-9000
Volume 5, **Number** 2, September 2025
https://doi.org/ 10.47709/digitech.v5i2.6532

Segmentasi Wilayah Digitalisasi di Indonesia dengan DBSCAN dan Validasi menggunakan Random Forest

Tan Azza Nuro Azzikra[1]
[1]Politeknik Statistika STIS, Indonesia
[1]222212894@stis.ac.id

open access
Histori Artikel:
Diajukan:16 Juli 2025
Disetujui:11 Oktober 2025
Dipublikasi:12 Oktober 2025
Kata Kunci:
DBSCAN, Segmentasi, Random Forest, Prediktif, Teknologi Informasi

Digital Transformation Technology (Digitech) *is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).*

Abstrak
Transformasi digital memainkan peran penting dalam pembangunan wilayah di Indonesia, namun masih terdapat ketimpangan antar provinsi yang cukup mencolok, terutama antara wilayah barat dan timur. Penelitian ini bertujuan untuk menganalisis kondisi digitalisasi wilayah Indonesia dengan pendekatan segmentasi spasial dan validasi prediktif berbasis data. Metode yang digunakan adalah algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise) untuk mengelompokkan provinsi berdasarkan indikator digitalisasi yang terdiri dari tiga dimensi utama: akses dan infrastruktur, penggunaan teknologi informasi, serta keahlian digital. Selanjutnya, Random Forest digunakan untuk mengevaluasi ketepatan segmentasi serta mengidentifikasi variabel yang paling berpengaruh dalam membedakan klaster digitalisasi wilayah. Hasil menunjukkan terbentuknya empat klaster digitalisasi: dua provinsi (DKI Jakarta dan DI Yogyakarta) berada pada klaster "tinggi", Papua dalam klaster "rendah", dan 31 provinsi lainnya dalam klaster "menengah". Nilai Silhouette Score sebesar 0,56 dan Davies-Bouldin Index sebesar 0,25 menunjukkan kualitas segmentasi yang baik. Meskipun Random Forest menunjukkan akurasi out-of-bag sebesar 91,18%, ketidakseimbangan kelas menyebabkan keterbatasan dalam memprediksi provinsi ekstrem. Namun, analisis feature importance mengungkap bahwa indikator "akses dan infrastruktur" merupakan penentu utama tingkat digitalisasi wilayah. Temuan ini menegaskan bahwa kebijakan digital di Indonesia memerlukan pendekatan yang diferensiatif, dengan prioritas pembangunan infrastruktur di wilayah tertinggal dan inovasi lanjutan di wilayah digital maju. Pendekatan kombinatif eksploratif dan prediktif ini memberikan kontribusi pada pemetaan digitalisasi wilayah berbasis data serta mendukung penyusunan kebijakan yang lebih adaptif dan responsif.

PENDAHULUAN

Transformasi digital telah menjadi elemen krusial dalam pembangunan nasional Indonesia karena diyakini mampu mempercepat pertumbuhan ekonomi, meningkatkan efisiensi pelayanan publik, serta memperluas akses informasi dan peluang (Kartiasih et al., 2023; Lito et al., 2025). Namun, adopsi dan pemanfaatan teknologi informasi dan komunikasi (TIK) masih belum merata di seluruh wilayah Indonesia. Disparitas digital yang mencolok antara provinsi di Pulau Jawa dan luar Jawa menunjukkan bahwa keberhasilan digitalisasi sangat dipengaruhi oleh kondisi sosial ekonomi, kesiapan infrastruktur, serta tingkat literasi digital penduduk (Susilowati et al., 2024; BPS, 2024).

Indeks Pembangunan TIK (IP-TIK) yang diterbitkan oleh Badan Pusat Statistik (2024) secara konsisten mencatat adanya ketimpangan digital antar wilayah. Provinsi seperti DKI Jakarta, DI Yogyakarta, dan Jawa Barat menunjukkan nilai indeks yang tinggi, sedangkan wilayah timur Indonesia seperti Papua dan Maluku cenderung tertinggal. Ketimpangan ini tidak hanya berimplikasi pada akses terhadap teknologi, tetapi juga memperkuat kesenjangan dalam bidang pendidikan, kesehatan, ekonomi digital, dan pelayanan publik (Kartiasih et al., 2023; Zhang et al., 2023). Sebagian besar ketimpangan ini bersumber dari perbedaan kapasitas fiskal dan kebijakan daerah, yang merupakan konsekuensi dari pelaksanaan desentralisasi fiskal di Indonesia. Penelitian Aritenang (2022) menunjukkan bahwa desentralisasi fiskal memiliki efek spasial yang signifikan terhadap ketimpangan pembangunan antardaerah. Selaras dengan itu, Suwanan dan Sulistiani (2023) menegaskan bahwa disparitas kemampuan fiskal antarwilayah berkontribusi terhadap ketidakseimbangan dalam pembangunan infrastruktur digital dan layanan publik berbasis teknologi.

Analisis berbasis data menjadi sangat penting untuk memahami struktur dan distribusi digitalisasi antar wilayah. Dalam konteks ini, algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN) digunakan karena kemampuannya untuk mengelompokkan data berdasarkan kepadatan, tanpa memerlukan

jumlah klaster awal, dan mampu mengidentifikasi wilayah yang menyimpang dari pola umum (*outlier*) (Wijaya et al., 2025; Tsabita et al., 2025). Dengan pendekatan ini, penelitian ini bertujuan untuk mengelompokkan provinsi di Indonesia berdasarkan indikator digitalisasi untuk mengidentifikasi pola kesamaan dan provinsi-provinsi outlier secara natural.

Namun, pengelompokan semata tidak cukup. Evaluasi terhadap kualitas dan ketepatan hasil klaster juga perlu dilakukan. Oleh karena itu, algoritma Random Forest digunakan sebagai metode prediktif untuk menilai keakuratan klasifikasi serta mengukur kontribusi relatif setiap indikator digitalisasi dalam membedakan klaster wilayah (DeLise, 2021; Fajri et al., 2025). Pendekatan ini memungkinkan interpretasi yang lebih mendalam mengenai variabel mana yang paling berperan dalam menentukan karakter digital suatu wilayah.

Dengan menggabungkan metode segmentasi spasial dan validasi prediktif, penelitian ini diharapkan dapat memberikan gambaran menyeluruh mengenai struktur digitalisasi wilayah di Indonesia. Lebih jauh, hasilnya dapat berkontribusi pada perumusan kebijakan pembangunan digital yang lebih adil, adaptif, dan berbasis bukti, sekaligus memperkuat upaya mengurangi kesenjangan digital antar daerah.

STUDI LITERATUR

Perkembangan digitalisasi wilayah telah menjadi fokus utama dalam berbagai studi di Indonesia, terutama karena implikasinya terhadap pembangunan ekonomi, sosial, dan tata kelola publik. Kartiasih et al. (2023) menyoroti ketimpangan digital antarprovinsi yang berkaitan erat dengan karakteristik sosial ekonomi dan ketersediaan infrastruktur. Studi ini menggunakan pendekatan multivariat untuk memetakan disparitas digital, namun belum mengeksplorasi metode segmentasi berbasis machine learning. Sementara itu, Zhang et al. (2023) mengembangkan kerangka pemantauan indeks masyarakat digital Indonesia menggunakan analisis spasial dan temporal, tetapi lebih menekankan pada pemetaan deskriptif dibanding segmentasi struktural.

Dari sisi metodologi, algoritma DBSCAN telah digunakan dalam beberapa penelitian untuk mengelompokkan data spasial dan ekonomi. Wijaya et al. (2025) berhasil menerapkan DBSCAN dalam mengidentifikasi klaster industri mikro dan kecil, menunjukkan bahwa algoritma ini efektif dalam mendeteksi pola alami tanpa harus menentukan jumlah klaster awal. Pendekatan serupa juga diadopsi oleh Wororomi et al. (2023), yang membandingkan performa DBSCAN dan K-Means dalam segmentasi PDRB regional. Namun, kedua studi tersebut belum mengintegrasikan pendekatan prediktif untuk memvalidasi klaster yang terbentuk.

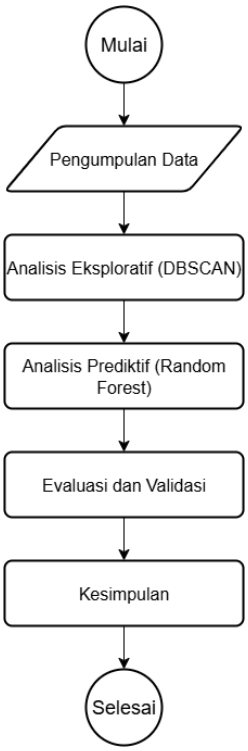
Penggunaan Random Forest dalam penelitian berbasis spasial juga telah dikaji oleh DeLise (2021), yang menggabungkan DBSCAN dan Random Forest untuk mengevaluasi hasil klasterisasi dan mengukur *feature importance*. Meskipun demikian, penelitian tersebut dilakukan dalam konteks non-Indonesia dan belum menggunakan indikator digitalisasi. Sementara itu, penelitian Fajri et al. (2025) membahas hubungan antara transformasi digital dan ketahanan pangan di Sumatra, tetapi tidak menggunakan teknik segmentasi klaster.

Dari studi-studi terdahulu, tampak bahwa terdapat kekosongan dalam riset yang mengintegrasikan DBSCAN dan Random Forest secara komprehensif dalam konteks segmentasi digitalisasi wilayah di Indonesia. Sebagian besar literatur hanya berfokus pada pemetaan atau korelasi antar indikator, tanpa pendekatan segmentatif dan validasi prediktif secara bersamaan. Oleh karena itu, penelitian ini hadir untuk mengisi gap tersebut dengan menerapkan model analitik berbasis data yang tidak hanya memetakan kondisi digitalisasi, tetapi juga mengevaluasi kontribusi masing-masing indikator terhadap pembentukan klaster wilayah digital.

METODE

Penelitian ini menggunakan pendekatan kuantitatif eksploratif dan prediktif. Pendekatan eksploratif bertujuan untuk mengidentifikasi pola pengelompokan alami (natural grouping) antar provinsi di Indonesia berdasarkan indikator digitalisasi, menggunakan algoritma DBSCAN. Sementara pendekatan prediktif digunakan untuk memvalidasi segmentasi tersebut melalui pemodelan klasifikasi menggunakan algoritma Random Forest.

Pendekatan kombinatif ini memungkinkan analisis dua arah: (1) memahami pola regional digitalisasi secara alami tanpa asumsi jumlah klaster tertentu, dan (2) mengevaluasi apakah segmentasi tersebut dapat diprediksi secara andal dari variabel input yang tersedia. Pendekatan serupa telah digunakan oleh DeLise[1] (2020) dalam konteks segmentasi dan validasi perilaku pelanggan. Adapun alur penelitian dapat dilihat dari gambar 1.



Gambar 1. Bagan Alir Tahapan Penelitian

Sumber dan Variabel Data

Data yang digunakan adalah data sekunder yang dikumpulkan dan diolah oleh Badan Pusat Statistik (BPS), yang mencerminkan kondisi digitalisasi tiap provinsi. Tiga dimensi dan indikator yang digunakan dalam penelitian ini merujuk pada metodologi penyusunan IP-TIK yang dipublikasikan oleh Badan Pusat Statistik (2024), yang telah menjadi rujukan standar dalam mengukur tingkat perkembangan TIK di tingkat provinsi, yaitu:

a. Akses & Infrastruktur (5 Indikator)

Dimensi akses dan infrastruktur menilai kesiapan fisik serta ketersediaan fasilitas teknologi di tingkat rumah tangga. Indikator yang digunakan mencakup tingkat kepemilikan dan akses terhadap layanan telekomunikasi serta internet, seperti penlanggan telepon tetap per 100 penduduk, pelanggan telepon seluler per 100 penduduk, *bandwith* internet internasional per pengguna, persentase rumah tangga dengan komputer, serta peresntase rumah tangga dengan akses internet di rumah.

b. Penggunaan (3 Indikator)

Dimensi penggunaan berfungsi untuk mengukur intensitas penggunaan teknologi informasi dan komunikasi oleh masyarakat. Tiga indikator yang digunakan meliputi persentase individu yang menggunakan internet, pelanggan *fixed broadband* internet per 100 penduduk, dan pelanggan *mobile broadband* aktif per 100 penduduk.

c. Keahlian (3 Indikator)

Dimensi keahlian mengukur tingkat literasi dan kesiapan sumber daya manusia dalam memanfaatkan teknologi. Indikator yang digunakan terdiri dari rata-rata lama sekolah, angka partisipasi kasar pendidikan menengah (SLTP & SLTA), dan angka partisipasi kasar pendidikan tinggi (D1-S1).

Setiap dimensi dihitung sebagai indeks rata-rata tertimbang berdasarkan bobot indikator yang telah ditentukan. Semua nilai dinormalisasi ke skala [0,1] menggunakan Min-Max Scaling agar setara dalam pembobotan (Wijaya[3] dkk., 2025).

Metode Segmentasi: DBSCAN

Tahap segmentasi dalam penelitian menggunakan algoritma DBSCAN (*Density-Based Spatial Clustering of Application with Noise*), yaitu algoritma klasterisasi bebasis kepadatan yang mampu membentuk klaster tanpa perlu menentukan jumlah klaster secara eksplisit. Pendekatan ini memungkinkan identifikasi pola pengelompokan alami (*natural grouping*) antarprovinsi berdasarkan kemiripan indikator digitalisasi yang dimiliki, serta mendeteksi *outlier*, yaitu provinsi dengan kondisi digitalisasi yang ekstrem atau menyimpang dari pola mayoritas.

DBSCAN bekerja dengan dua parameter utama. Paramater pertama adalah ε (epsilon) yang menunjukkan jarak maksimum antar titik agar dapat dianggap sebagai tetangga dan parameter kedua adalah MinPts, yaitu jumlah minimum tentanga dalam radius ε agar sebuah titik dapat dikategorikan sebagai puast klaster (*core point*). Setelah proses klasterisasi dilakukan, hasil yang diperoleh dianalissi secara deskriptif untuk menafsirkan

karakteristik setiap klaster. Selanjutnya, hasil klasterisasi ini dikonversi menjadi label klasifikasi berdasarkan nilai rerata atau skor digitalisasi provinsi per klaster (DeLise, 2020).

Metode Validasi Prediktif: Random Forest

Untuk mengevaluasi hasil segmentasi yang diperoleh melalui DBSCAN, penelitian ini menggunakan pendekatan Random Forest sebagai metode klasifikasi prediktif. Random Forest merupakan algoritma *ensemble* yang menggabungkan sejumlah *decision tree* dan menentukan hasil akhir berdasarkan voting mayoritas. Metode ini dipilih karena memiliki kemampuan untuk menangani hubungan non-linear antarvariabel, than terhadap *overfitting*, serta mampu memberikan informasi mengenai *feature importance*, yaitu tingkat pengaruh relatif dari masing-masing indikator digitalisasi terhadap pembentukan klaster provinsi (DeLise, 2020).

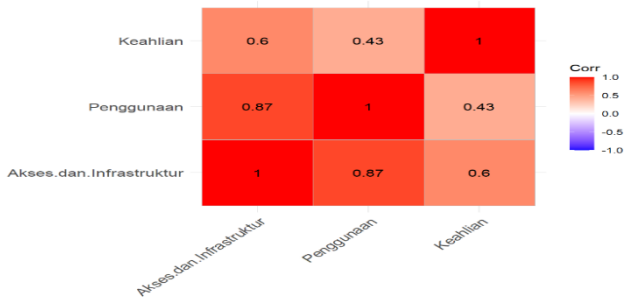
Model Random Forest yang dibangun dalam penelitian ini dievaluasi menggunakan akurasi klasifikasi dan *confusion matrix* untuk menilai kesesuaian antara hasil prediksi dengan segementasi DBSCAN. Selain itu, digunakan juga metode *Leave-One-Out Cross Validation* (LOOCV) sebagai teknik validasi yang sesuai dengan jumlah observasi yang relatif kecil, yaitu 34 provinsi. Pendekatan ini memastikan bahwa hasil segmentasi yang terbentuk bersifat deskriptif dan memiliki struktur yang dapat diprediksi secara baik dari variabel-variabel yang digunakan.

HASIL

Persiapan dan Transformasi Data Digitalisasi Wilayah

Tahapan awal dalam analisis ini adalah proses pra pemrosesan data untuk memastikan kesiapan data dalam proses segmentasi dan pemodelan. Data mentah yang diperoleh berasal dari indikator-indikator digitalisasi wilayah yang terdiri atas tiga kelompok besar: Akses dan Infrastruktur, Penggunaan, serta Keahlian (Skill). Masing-masing kelompok terdiri dari indikator yang memiliki satuan dan rentang nilai berbeda, meskipun umumnya sudah berada pada skala agregat 0–10.

Agar tidak terjadi distorsi dalam penghitungan jarak pada algoritma DBSCAN yang berbasis *Euclidean distance*, Seluruh variabel ditransformasikan menggunakan Min-Max Scaling ke dalam rentang [0,1]. Normalisasi ini memastikan setiap variabel memiliki kontribusi yang seimbang dalam perhitungan jarak. Hasil transformasi menunjukkan bahwa seluruh variabel kini berada dalam skala seragam, tanpa kehilangan pola distribusi dasarnya.



Gambar 2. Matriks Korelasi Antar Dimensi Digitalisasi

Gambar 2 menunjukkan hasil korelasi antar tiga dimensi digitalisasi, yaitu Akses dan Infrastruktur, Penggunaan, dan Keahlian. Terlihat bahwa dimensi Akses dan Infrastruktur memiliki korelasi yang sangat kuat dengan Penggunaan (r = 0,87), yang mengindikasikan bahwa provinsi dengan infrastruktur digital yang baik cenderung memiliki tingkat penggunaan teknologi yang tinggi. Sementara itu, korelasi antara Akses dan Infrastruktur dengan Keahlian (r = 0,60) serta antara Penggunaan dengan Keahlian (r = 0,43) menunjukkan hubungan yang cukup moderat. Pola ini menegaskan bawha aspek keahlian digital masyarakat belum sepenuhnya sebanding dengan tingkat ketersediaan infrastruktur atau tingkat penggunaan teknologi di suatu provinsi.

Selanjutnya, dilakukan analisis korelasi antar variabel untuk memeriksa kemungkinan multikolinearitas. Meski indikator digitalisasi disusun dari kelompok berbeda, beberapa variabel menunjukkan korelasi cukup kuat yang dapat mengindikasikan redundansi informasi. Kondisi ini menunjukkan adanya korelasi antar indikator yang cukup kuat, yang secara teoritis dapat mempengaruhi hasil klasterisasi DBSCAN.

Mengingat jumlah variabel tidak terlalu banyak, pendekatan reduksi dimensi seperti PCA tidak diterapkan pada tahap ini karena dimensi sudah dilakukan semacam PCA. Sebagai bentuk evaluasi tambahan, pemodelan Random Forest dilakukan pada tahap selanjutnya untuk menilai relevansi relatif setiap variabel terhadap hasil segmentasi wilayah, sehingga dapat memberikan gambaran mengenai kontribusi masing-masing fitur terhadap pemisahan klaster yang dihasilkan DBSCAN.

Kombinasi antara normalisasi dan validasi berbasis supervised learning ini memberikan fondasi analisis yang kuat, terutama dalam mengurangi efek ketidakseimbangan skala serta memastikan bahwa pola klasterisasi yang terbentuk benar-benar mencerminkan perbedaan substantif dalam kondisi digitalisasi antarwilayah.

Penentuan Parameter Optimal untuk Segmentasi Wilayah

Pemilihan parameter inti dalam DBSCAN, yaitu nilai epsilon (ε) dan minimum points (MinPts), sangat berperan dalam menentukan kualitas klaster yang terbentuk. Karena data telah dinormalisasi menggunakan Min-Max Scaler, seluruh variabel berada dalam rentang [0,1], sehingga pemilihan nilai ε dibatasi antara 0 hingga 1. Nilai ε yang melebihi 1 tidak relevan karena akan mencakup seluruh ruang data dan menghilangkan kepekaan terhadap kepadatan lokal. Pemilihan parameter ε dan MinPts sangat menentukan kualitas klaster dan penentuan nilai epsilon dan MinPts dilakukan secara eksperimental, oleh karena itu akan dilakukan simulasi untuk menentukan parameter optimal.

Tabel 1. Simulasi Epsilon dan Sampel Minimum Optimal

eps	minPts	clusters	noise	silhouette	davis_bouldin
0.1	1	11	0	0.263	0.414
0.2	1	4	0	0.560	0.252
0.3	1	4	0	0.560	0.252
0.4	1	3	0	0.568	0.275
0.5	1	2	0	0.692	0.242
0.6	1	2	0	0.692	0.242
0.7	1	2	0	0.692	0.242
0.8	1	1	0	-	-
0.9	1	1	0	-	-
1	1	1	0	-	-

Eksperimen menunjukkan bahwa nilai ε = 0,2 dan MinPts = 1 menghasilkan konfigurasi klaster terbaik. Kombinasi ini membentuk empat klaster yang jelas tanpa adanya noise, yang menghasilkan segmentasi empat klaster tanpa outlier, dengan semua provinsi terkelompok.

Meskipun nilai Silhouette Score yang dihasilkan tidak terlalu tinggi (0,560), hal ini dapat dimaklumi mengingat adanya heterogenitas internal antarprovinsi. Perolehan nilai Davies-Bouldin Index sebesar 0,25, yang secara umum digunakan untuk menilai pemisahan klaster yang baik.

Dibandingkan konfigurasi lainnya, kombinasi ini juga memiliki keunggulan signifikan. Misalnya, pada epsilon yang sebesar 0,2, dengan jumlah minimum titik (MinPts) sebesar 1, meskipun Silhouette Score sedikit meningkat, jumlah klaster berkurang dan muncul beberapa outlier. Konfigurasi dengan jumlah minimum titik (MinPts) lebih besar atau sama dengan 4 menghasilkan lebih sedikit klaster dan lebih banyak noise. Oleh karena itu, epsilon yang sebesar 0,2, dengan jumlah minimum titik (MinPts) sebesar 1 dipilih sebagai parameter akhir karena memberikan keseimbangan terbaik antara sensitivitas segmentasi dan kualitas evaluasi klaster.

Pola Klasterisasi Digitalisasi Provinsi di Indonesia

Hasil segmentasi wilayah menggunakan algoritma DBSCAN dengan parameter optimal epsilon yang digunakan adalah 0,2, dengan jumlah minimum titik (MinPts) sebesar 1. menghasilkan empat klaster utama tanpa noise atau outlier, yang secara umum mencerminkan pola persebaran tingkat digitalisasi antar provinsi di Indonesia. Visualisasi ini disusun berdasarkan data yang telah dinormalisasi dan diproyeksikan ke dalam klaster DBSCAN.

Klaster 1 dan Klaster 2 masing-masing hanya terdiri dari satu provinsi: DI Yogyakarta dan DKI Jakarta, yang memiliki skor tertinggi secara konsisten dalam ketiga dimensi digitalisasi—akses & infrastruktur, penggunaan, dan keahlian. Kedua provinsi ini diberi label “Tinggi”, merepresentasikan wilayah dengan kesiapan digital yang sangat matang, baik dari sisi ketersediaan teknologi maupun kapasitas manusianya dalam memanfaatkannya. Letaknya yang berada di Pulau Jawa semakin memperkuat karakteristik unik digitalisasi mereka.

Klaster 3, yang merupakan klaster terbesar, terdiri dari 31 provinsi lainnya yang tersebar di seluruh Indonesia. Klaster ini mencerminkan kelompok wilayah dengan tingkat digitalisasi “Menengah”. Meskipun terdapat variasi kecil antar provinsi, keseluruhan indikator menunjukkan nilai-nilai yang relatif seragam dan merata. Sebagian besar provinsi tergabung dalam satu klaster menengah, dengan sedikit provinsi berada di klaster berbeda.

Klaster 4 hanya mencakup satu provinsi, yaitu Papua, yang menempati posisi terendah dalam hampir seluruh indikator digitalisasi. Oleh karena itu, klaster ini diberi label “Rendah”, merefleksikan keterbatasan akses terhadap infrastruktur TIK, penggunaan internet, serta tantangan dalam literasi dan keahlian digital di wilayah tersebut.

DBSCAN menghasilkan empat klaster wilayah berdasarkan digitalisasi. Klaster 1 dan 2 masing-masing terdiri dari satu provinsi, sedangkan Klaster 3 mencakup 31 provinsi, dan Klaster 4 terdiri dari satu provinsi. Pola ini menunjukkan variasi tingkat digitalisasi antarprovinsi.

Evaluasi Keakuratan dan Faktor Penentu Digitalisasi

Sebagai langkah validasi terhadap hasil segmentasi DBSCAN, algoritma Random Forest digunakan untuk memprediksi label klaster berdasarkan tiga fitur utama: Akses dan Infrastruktur, Penggunaan, dan Keahlian. Namun, hasil validasi ini mengungkapkan bahwa klasifikasi dengan Random Forest memiliki keterbatasan dalam konteks distribusi klaster yang sangat tidak seimbang.

Dari total 34 provinsi yang dianalisis, 31 provinsi tergolong dalam klaster "Menengah", sementara hanya 2 provinsi berada dalam klaster "Tinggi" (DKI Jakarta dan DI Yogyakarta), dan 1 provinsi dalam klaster "Rendah" (Papua). Ketimpangan distribusi ini menyebabkan model Random Forest cenderung memprediksi seluruh data ke dalam kategori "Menengah", sebagaimana terlihat dari confusion matrix yang menunjukkan tidak adanya prediksi yang benar pada klaster "Tinggi" dan "Rendah". Meskipun akurasi out-of-bag (OOB) yang dicapai mencapai 91,18%, angka ini perlu ditafsirkan dengan hati-hati karena kemungkinan bias akibat dominasi klaster mayoritas.

Secara keseluruhan, hasil prediksi Random Forest justru memperkuat temuan dari klasterisasi DBSCAN. Model prediktif yang secara konsisten memetakan wilayah ke dalam klaster menengah mencerminkan tingkat keseragaman digitalisasi yang tinggi di sebagian besar wilayah Indonesia, dengan hanya sedikit provinsi yang menyimpang secara signifikan. Dengan demikian, kombinasi pendekatan eksploratif dan prediktif ini memberikan gambaran yang konsisten bahwa ketimpangan digital saat ini hanya terjadi di provinsi-provinsi ekstrem, sementara sebagian besar wilayah menunjukkan pola digitalisasi yang seragam. Meskipun klasifikasi tidak optimal akibat class imbalance, informasi dari feature importance tetap relevan sebagai alat interpretasi dalam segmentasi wilayah (Santos et al., 2020; DeLise, 2021).

Tabel 2. Feature Importance Random Forest

Variabel	Mean Decrease Accuracy	MeanDecreaseGini
Akses dan Infrastruktur	10.52	2.21
Penggunaan	-1.82	1.68
Keahlian	-0.56	1.71

Meski demikian, analisis feature importance tetap memberikan wawasan penting. Variabel Akses dan Infrastruktur memiliki kontribusi paling tinggi terhadap pemisahan klaster (MeanDecreaseAccuracy = 10.52), jauh melampaui variabel Penggunaan dan Keahlian. Jadi, nilai feature importance tertinggi terdapat pada variabel Akses dan Infrastruktur.

PEMBAHASAN

Hasil segmentasi menggunakan algoritma DBSCAN menunjukkan bahwa sebagian besar provinsi di Indonesia tergolong dalam satu klaster digitalisasi menengah, sedangkan hanya tiga provinsi berada di klaster ekstrem, yaitu DKI Jakarta, DI Yogyakarta (tinggi), dan Papua (rendah). Hal ini menegaskan bahwa ketimpangan digital antarwilayah tidak sepenuhnya menyebar merata, melainkan terkonsentrasi pada daerah tertentu. Temuan ini memperluas pemahaman dari penelitian Kartiasih et al. (2023) dan Susilowati et al. (2024), yang menunjukkan adanya hubungan kuat antara disparitas digital dan kondisi sosial-ekonomi, terutama di luar wilayah Jawa.

Pendekatan DBSCAN memungkinkan segmentasi tanpa asumsi jumlah klaster dan terbukti mampu mengidentifikasi provinsi outlier berdasarkan pola kepadatan data. Metode ini terbukti efektif dalam mengidentifikasi provinsi dengan karakteristik digitalisasi ekstrem. Nilai evaluasi Silhouette Score sebesar 0,56 dan Davies-Bouldin Index sebesar 0,25 mengindikasikan segmentasi yang cukup baik, meskipun terdapat heterogenitas internal dalam klaster menengah. Hasil ini juga konsisten dengan temuan Wororomi et al. (2023) yang menyatakan bahwa DBSCAN lebih fleksibel dalam menangani data spasial dibanding K-Means.

Untuk mengevaluasi hasil segmentasi secara prediktif, algoritma Random Forest digunakan sebagai pendekatan supervised learning. Meskipun model menunjukkan akurasi OOB yang tinggi (91,18%), distribusi klaster yang tidak seimbang menyebabkan performa klasifikasi menjadi bias terhadap klaster mayoritas. Kondisi ini merupakan contoh umum dari ketidakseimbangan kelas pada data klasifikasi. Namun demikian, analisis *feature importance* tetap memberikan informasi berharga, di mana indikator "Akses dan Infrastruktur" memiliki kontribusi tertinggi terhadap pemisahan klaster, mendahului indikator penggunaan dan keahlian.

Temuan ini menggarisbawahi pentingnya konektivitas digital dan ketersediaan infrastruktur sebagai prasyarat utama dalam pengembangan wilayah digital. Hasil ini sejalan dengan temuan Lito et al. (2025) dan Susanti et al. (2023), yang menunjukkan bahwa ketimpangan akses dan perangkat teknologi berperan besar dalam memperlebar jurang digital antarwilayah. Selain itu, pengaruh geografis yang tidak secara eksplisit digunakan dalam model ternyata tetap tercermin dalam hasil klasterisasi, mengindikasikan bahwa faktor lokal dan institusional masih sangat relevan. Studi oleh Aritenang (2022) dan Pratama & Nugroho (2023) mendukung pandangan ini, dengan menekankan bahwa desentralisasi fiskal dan kapasitas tata kelola turut mempengaruhi keberhasilan transformasi digital.

KESIMPULAN

Penelitan ini berhasil mengelompokkan provinsi-provinsi di indonesia ke dalam empat klaster digitalisasi menggunakan algoritma DBSCAN, dengan sebagian besar provinsi berada di kategori menengah, dua provinsi (DKI Jakarta dan DI Yogyakarta) berada di dua klaster berbeda tetapi masih berkategori tinggi, dan satu provinsi (Papua) pada kategori rendah. Hasil validasi melalui meodel Random Forest menunjukkan bawha indikator akses dan infrastruktur merupakan faktor paling berpengaruh dalam membentuk klaster digitalisasi wilayah. Temuan ini menegaskan bahwa tingkat digitalisasi di Indonesia masih terkonsentrasi pada wilayah tetentu dan dipengaruhi kuat oleh ketersediaan infrastruktur. Oleh karena itu, diperlukan kebijakan dalam pembangunan digital yang lebih merata, terutama dalam peningkatan akses dan kapasitas infrastruktur di daerah yang tertinggal.

REFERENSI

Aritenang, A. F. (2022). The spatial effect of fiscal decentralisation on regional disparities: The case from Indonesia. *Indonesian Journal of Geography, 54*(2), 180–192. https://doi.org/10.22146/ijg.4985

Badan Pusat Statistik. (2024). *Indeks pembangunan teknologi informasi dan komunikasi (IP-TIK) Indonesia 2023.* Jakarta, Indonesia: BPS.

DeLise, T. (2021). Data segmentation via t-SNE, DBSCAN, and random forest. *arXiv preprint arXiv:2010.13682.* https://doi.org/10.48550/arXiv.2010.13682

Fajri, M. N., Nurdin, M. H., & Yulianingsih, R. (2025). Spatial analysis of digital transformation on food security in Sumatra. *Jurnal Ilmu Ekonomi Terapan, 10*(1), 77–90. Retrieved from https://e-journal.unair.ac.id/JIET/article/view/60891

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). San Francisco, CA: Morgan Kaufmann.

Kartiasih, F., Nachrowi, N. D., Wisana, I. D. G. K., & Handayani, D. (2023). Inequalities of Indonesia's regional digital development and its association with socioeconomic characteristics: A spatial and multivariate analysis. *Information Technology for Development, 29*(2–3), 299–328.

Lito, L. S. J., Samudro, B. R., & Soesilo, A. M. (2025). Digital transformation and regional development disparities in Indonesia. *Jurnal Transformasi Digital dan Kebijakan Publik, 3*(1), 45–59.

Pratama, M. I., & Nugroho, Y. (2023). Digital infrastructure, governance, and regional inequality in Indonesia: A multidimensional approach. *Asian Journal of Comparative Politics, 8*(1), 62–80. https://doi.org/10.1177/20578911221134421

Santos, J. M., & Embrechts, M. (2020). On the use of random forest for imbalanced data classification. *Neurocomputing, 286*, 168–179.

Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems, 42*(3), 1–21.

Susanti, I., Nurhayati, E., & Azzahra, R. (2023). Pengaruh akses dan literasi digital terhadap pembangunan desa di Indonesia. *Jurnal Ekonomi dan Kebijakan Pembangunan, 12*(1), 66–78.

Susilowati, A. P. E., Rachmawati, R., & Rijanta, R. (2024). Does the digital divide matter? Comparison between Java and the rest of Indonesia in developing smart villages. *Human Geographies – Journal of Studies and Research in Human Geography, 18*(2). https://doi.org/10.5719/hgeo.2024.182.6

Suwanan, A., & Sulistiani, E. H. (2023). Fiscal decentralization and regional disparities in Indonesia. *Jurnal Ilmu Ekonomi dan Pembangunan, 24*(1), 41–52. https://doi.org/10.20961/jiep.v24i1.6310

Tsabita, H., Ahmad, I. S., & Adipradana, D. W. (2025). Klasterisasi provinsi di Indonesia menggunakan algoritma K-means dan DBSCAN berdasarkan indikator makroekonomi dan sosial. *Jurnal Sains dan Seni ITS, 11*(2), F184–F189. https://doi.org/10.12962/j23373520.v11i2.149903

Wijaya, F. B., Budiaji, W., & Wicaksono, A. S. (2025). Applied machine learning DBSCAN for identifying clusters of micro and small industries. *RIGGS, 4*(2), 380–386. https://doi.org/10.31004/riggs.v4i2.515

Wororomi, J. K., Allo, C. B. G., & Paranoan, N. R. (2023). Performance of K-means and DBSCAN algorithm in clustering gross regional domestic product. *Jurnal Ilmiah Citra Prasasti (JICP), 6*(5), 179–193. https://doi.org/10.32535/jicp.v6i5.2710

Zhang, W., Nasution, A., & Nugroho, A. (2023). A framework for monitoring the spatiotemporal distribution and clustering of the digital society index of Indonesia. *Sustainability, 16*(24), 11258. https://doi.org/10.3390/su162411258

Zhang, Y., & Démurger, S. (2022). Bridging the digital divide in developing countries: The role of spatial inequality and policy design. *World Development, 158*, 106004. https://doi.org/10.1016/j.worlddev.2022.106004