https://jurnal.pustakagalerimandiri.co.id/index.php/pustakaai
DOI : https://doi.org/10.55382/jurnalpustakaai.v5i.984

JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE

Vol. 5 No. 2 (2025) 204 – 214 E ISSN: 2809-4069

Model Data-Driven untuk Prediksi Digitalisasi UMKM Menggunakan GMM dan XGBoost

Evi Purnamasari[1], Dwi Asa Verano[2]
[1,2]Teknik Informatika, Fakultas Ilmu Komputer dan Sains, Universitas Indo Global Mandiri
[1]evi.ps@uigm.ac.id, [2]dwiasa@uigm.ac.id

Abstract

The uneven adoption of digital technology among Micro, Small, and Medium Enterprises (MSMEs) presents a challenge in achieving inclusive and targeted digital transformation. This gap often weakens the effectiveness of MSME development policies due to the absence of data-driven segmentation. This study aims to build a model that can automatically and accurately map and predict the level of digitalization among MSMEs. A hybrid approach is employed, integrating the Gaussian Mixture Model (GMM) for segmentation and XGBoost as the main classification algorithm. GMM effectively groups MSMEs into four segments: (1) Traditional, (2) Early Semi-Digital, (3) Advanced Semi-Digital, and (4) Fully Digital. The model achieved optimal clustering performance with a Silhouette Score of 0.3712, Davies-Bouldin Index of 0.9120, and Calinski-Harabasz Index of 129.75.. For classification, XGBoost outperformed other models with an accuracy of 98.63%, precision of 0.99, recall of 0.98, and F1-score of 0.99. The model accurately predicts the digital segment of new MSMEs based on input features such as revenue, profit, and digitalization score. The proposed hybrid model is effective for data-driven segmentation and prediction, supporting the development of more precise and adaptive MSME digitalization policie

Keywords: *MSMEs, Digitalization, Data-Driven, Gaussian Mixture Model, XGBoost, Segmentation, Classification, Machine Learning, Prediction, Digital Transformation*

Abstrak

Tingkat adopsi teknologi digital yang belum merata di kalangan Usaha Mikro, Kecil, dan Menengah (UMKM) menjadi tantangan dalam mewujudkan transformasi digital yang inklusif dan terarah. Kesenjangan ini sering kali melemahkan efektivitas kebijakan pengembangan UMKM akibat belum adanya segmentasi berbasis data. Penelitian ini bertujuan untuk membangun model yang mampu memetakan dan memprediksi tingkat digitalisasi UMKM secara otomatis dan akurat. Pendekatan hybrid digunakan dengan mengintegrasikan Gaussian Mixture Model (GMM) untuk segmentasi dan XGBoost sebagai algoritma klasifikasi utama. GMM berhasil mengelompokkan UMKM ke dalam empat segmen, yaitu: (1) Tradisional, (2) Semi-Digital Awal, (3) Semi-Digital Lanjut, dan (4) Digital Penuh. Hasil evaluasi menunjukkan performa optimal dengan Silhouette Score sebesar 0,3712, Davies-Bouldin Index 0,9120, dan Calinski-Harabasz Index 129,75. Pada tahap klasifikasi, XGBoost menunjukkan performa terbaik dengan akurasi 98,63%, precision 0,99, recall 0,98, dan F1-score 0,99. Model mampu memprediksi segmen digitalisasi UMKM baru berdasarkan fitur-fitur seperti pendapatan, profit, dan skor digitalisasi. Model hybrid ini efektif dalam mendukung segmentasi dan prediksi berbasis data, serta dapat menjadi dasar dalam perumusan kebijakan digitalisasi UMKM yang lebih presisi dan adaptif.

Kata Kunci: UMKM, Digitalisasi, Data-Driven, Gaussian Mixture Model, XGBoost, Segmentasi, Klasifikasi, Machine Learning, Prediksi, Transformasi Digital.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

1. Pendahuluan

Usaha Mikro, Kecil, dan Menengah (UMKM) memainkan peran yang sangat penting dalam struktur perekonomian nasional dan global. Di Indonesia, sektor ini menjadi tulang punggung ekonomi karena kontribusinya yang besar terhadap Produk Domestik Bruto (PDB) dan penyerapan tenaga kerja [1], [2]. Namun demikian, seiring dengan semakin cepatnya laju digitalisasi, banyak UMKM menghadapi tantangan serius dalam meningkatkan daya saingnya secara berkelanjutan [3]. Ketimpangan dalam adopsi teknologi digital masih menjadi isu utama, di mana sebagian UMKM telah mulai mengimplementasikan sistem pembayaran digital dan manajemen berbasis aplikasi, sementara sebagian lainnya masih menggunakan metode tradisional [4].

Perbedaan dalam tingkat adopsi teknologi tersebut menimbulkan kesenjangan kompetitif antar UMKM. Oleh karena itu, pemetaan tingkat digitalisasi menjadi hal penting untuk membantu para pemangku kepentingan dalam merancang intervensi yang tepat sasaran [5]. Sayangnya, pendekatan yang selama ini digunakan dalam mengelompokkan UMKM masih banyak bergantung pada metode konvensional seperti klasifikasi manual atau segmentasi deskriptif sederhana, yang kurang akurat dan sulit diandalkan dalam skala besar [6]. Permasalahan lainnya terletak pada struktur data UMKM yang cenderung tidak terorganisir dan bersifat heterogen, sehingga sulit untuk dilakukan analisis mendalam secara otomatis [7].

Dalam konteks ini, diperlukan pendekatan berbasis teknologi yang lebih canggih dan presisi. Model analitik berbasis data (*data-driven*) yang mengintegrasikan metode pembelajaran mesin, baik dalam bentuk pembelajaran tanpa pengawasan (*unsupervised learning*) maupun terawasi (*supervised learning*), menjadi solusi yang menjanjikan. Teknik seperti Gaussian Mixture Model (GMM) memungkinkan pengelompokan entitas UMKM berdasarkan distribusi probabilistik dan pola tersembunyi yang kompleks [8]. Selanjutnya, untuk memprediksi tingkat digitalisasi UMKM baru secara akurat, algoritma Random Forest dipilih karena kemampuannya yang tinggi dalam menangani variabel yang saling berinteraksi serta ketahanannya terhadap overfitting [9].

Penelitian ini bertujuan untuk merancang sebuah model *data-driven* yang mampu melakukan pemetaan dan prediksi tingkat digitalisasi UMKM secara efektif. Tujuan khusus dari penelitian ini meliputi:

1. Mengelompokkan UMKM berdasarkan karakteristik digital dan operasional menggunakan empat algoritma clustering, yaitu K-Means, Agglomerative, Gaussian Mixture Model (GMM), dan HDBSCAN.
2. Mengevaluasi performa masing-masing algoritma menggunakan metrik Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index untuk menentukan algoritma terbaik.
3. Mengembangkan model klasifikasi menggunakan enam algoritma supervised learning untuk memprediksi klaster digitalisasi, dengan fokus utama pada Random Forest.
4. Melakukan analisis fitur dominan yang mempengaruhi digitalisasi UMKM untuk mendukung strategi intervensi yang lebih presisi [10].

Kontribusi utama dari penelitian ini adalah pengembangan kerangka kerja analitik berbasis *machine learning* untuk memetakan dan memprediksi tingkat digitalisasi UMKM. Pendekatan ini tidak hanya memungkinkan pengelompokan otomatis berdasarkan pola yang tidak terlihat secara eksplisit, tetapi juga menyediakan kemampuan prediktif yang dapat digunakan dalam pengambilan keputusan real-time oleh pemerintah, lembaga pendamping UMKM, dan sektor industri digital. Dengan demikian, penelitian ini diharapkan memberikan nilai praktis dalam mempercepat transformasi digital UMKM secara nasional.

Seluruh proses analisis, pelatihan model, dan evaluasi dilakukan menggunakan platform Google Colab berbasis cloud, guna memastikan efisiensi pemrosesan dan fleksibilitas kolaboratif dalam eksperimen data

2. Metode Penelitian

2.1. Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode analisis data eksploratif untuk mengidentifikasi pola dan karakteristik digitalisasi UMKM berdasarkan data operasional dan adopsi teknologi. Proses analisis dilakukan dalam dua tahap utama, yaitu:

1. Pemetaan (mapping) menggunakan pendekatan unsupervised learning melalui algoritma clustering, dan
2. Prediksi menggunakan supervised learning.

Empat algoritma clustering diterapkan untuk mengeksplorasi struktur alami dari data UMKM, yaitu K-Means, Agglomerative Clustering, Gaussian Mixture Model (GMM), dan HDBSCAN. Setiap algoritma mewakili pendekatan yang berbeda: partisi, hierarki, probabilistik, dan densitas. Hasil clustering dievaluasi menggunakan tiga metrik performa: Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index [6], [7].

Dari keempat algoritma, hasil terbaik digunakan sebagai dasar pelabelan (label klaster) dalam tahap selanjutnya. Pada tahap kedua, dilakukan klasifikasi prediktif untuk mengidentifikasi tingkat digitalisasi UMKM baru berdasarkan karakteristik inputnya. Proses ini dilakukan menggunakan enam algoritma supervised learning, yaitu: Random Forest, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Neural Network, dan Logistic Regression. Namun, fokus utama dalam penelitian ini adalah menilai performa dari Random Forest sebagai kandidat utama model prediksi [8], [9].

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Kerangka kerja ini diimplementasikan menggunakan platform Google Colab untuk mengelola proses komputasi, pelatihan model, dan evaluasi performa secara efisien dan kolaboratif [10].

2.2. Data dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh melalui penyebaran kuesioner secara daring kepada pelaku UMKM di wilayah Kota Palembang, dengan dukungan dan fasilitasi dari Dinas Koperasi dan UMKM Kota Palembang [1], [2].

Kuesioner dirancang untuk menggali informasi terkait kondisi operasional usaha serta tingkat pemanfaatan teknologi digital dalam aktivitas sehari-hari. Tujuannya adalah agar data yang dikumpulkan dapat mencerminkan kondisi nyata serta mendukung pemodelan berbasis data secara komprehensif.

Jumlah total data yang berhasil dikumpulkan dan digunakan dalam penelitian ini adalah 362 entitas UMKM, yang mencakup berbagai sektor usaha dan skala operasi [3].

2.3. Variabel Penelitian

Variabel dalam penelitian ini dibagi ke dalam dua kategori utama, yaitu variabel operasional dan variabel digitalisasi [4], [5].

a. Variabel Operasional
   Variabel ini mencerminkan informasi demografis, struktur usaha, serta performa keuangan UMKM. Variabel-variabel ini digunakan untuk memahami kondisi usaha secara menyeluruh dan meliputi: Education_Level, Marital Status, Business Ownership Status, Subdistrict, Business Scale, Business Type, Monthly Revenue, Operating Costs, Profit, Avg Monthly Production, Products Sold Per Month
b. Variabel Digitalisasi
   Variabel ini merepresentasikan tingkat adopsi teknologi digital dalam operasional UMKM. Pengukuran dilakukan melalui satu indikator komposit yaitu Digitalization_Score, yang diperoleh berdasarkan respons terhadap beberapa aspek penggunaan teknologi, seperti penggunaan media sosial, sistem kasir digital, platform marketplace, dan layanan keuangan digital.
c. Kedua kategori variabel tersebut digunakan secara bersama-sama dalam proses analisis untuk memetakan profil UMKM serta membangun model prediktif digitalisasi yang lebih akurat dan presisi [6].

2.2. Analisis Data Eksploratif

Tahap eksplorasi data dilakukan sebagai bagian dari pendekatan data-driven untuk memperoleh pemahaman awal mengenai distribusi, karakteristik, serta hubungan antar variabel dalam dataset UMKM yang menjadi dasar pemetaan dan prediksi tingkat digitalisasi [11]. Analisis statistik deskriptif terhadap variabel numerik disajikan guna memberikan gambaran menyeluruh mengenai nilai rata-rata (mean), standar deviasi, nilai minimum, kuartil, dan nilai maksimum dari setiap atribut yang digunakan dalam proses pemodelan [12]. Hasil statistik ini disajikan pada Tabel 1 sebagai dasar awal untuk seleksi fitur dan pengembangan model menggunakan Gaussian Mixture Model (GMM) dan Random Forest.

Tabel 1. Hasil Statistik Deskriptif data Penelitian

Variabel	Mean	Std Dev	Min	25%	50 %	75%	Max
Education Level	3	1	2	2	3	5	5
Marital Status	1	0	1	1	1	1	2
Business Ownership Status	1	1	1	1	1	1	3
Subdistrict	11	6	1	5	10	16	20
Business Scale	1	0	1	1	1	1	3
Business Type	3	3	0	1	2	4	15
Monthly Revenue	6,139*	7,515*	40*	1,637*	4**	8**	75**
Operating Costs	2,859*	7,762*	0	500*	1**	3**	130**
Profit	3,144*	4,440*	0	1**	2**	4**	50**
AvgMonthly Production	9,050	157,7	1	10	30	300	3**
Products Sold Per_Month	6,070	105,1	1	10	30	300	2**
Digitalization_Score	3	1	2	2	3	5	5

* dalam ribuan
** dalam jutaan

Hasil identifikasi awal terhadap dataset menunjukkan bahwa tidak terdapat outlier, nilai yang hilang (*missing values*), maupun inkonsistensi data. Dengan demikian, seluruh variabel dinyatakan valid dan dapat digunakan secara utuh dalam proses analisis dan pemodelan selanjutnya.

2.3. Proses Analisa Data

1) Pra-pemrosesan Data

Tahap pra-pemrosesan data diawali dengan pemeriksaan nilai kosong (*missing values*) pada seluruh atribut dalam dataset. Hasil analisis menunjukkan bahwa tidak terdapat nilai kosong, sehingga tidak diperlukan proses imputasi maupun penghapusan data [13].

Selanjutnya,dila kukan deteksi outlier menggunakan pendekatan statistik deskriptif serta visualisasi melalui boxplot. Berdasarkan hasil tersebut, tidak ditemukan outlier yang signifikan, sehingga seluruh data dapat digunakan secara utuh dalam proses pemodelan [14].

Mengingat fitur dalam dataset memiliki rentang nilai yang sangat bervariasi, dilakukan proses normalisasi menggunakan metode Min-Max Scaler untuk menyamakan skala antar atribut. Teknik ini mengubah nilai setiap fitur ke dalam rentang 0 hingga 1, sehingga mendukung kinerja optimal dari algoritma Gaussian Mixture Model (GMM) dalam tahap pemetaan serta Random Forest dalam tahap prediksi [15].

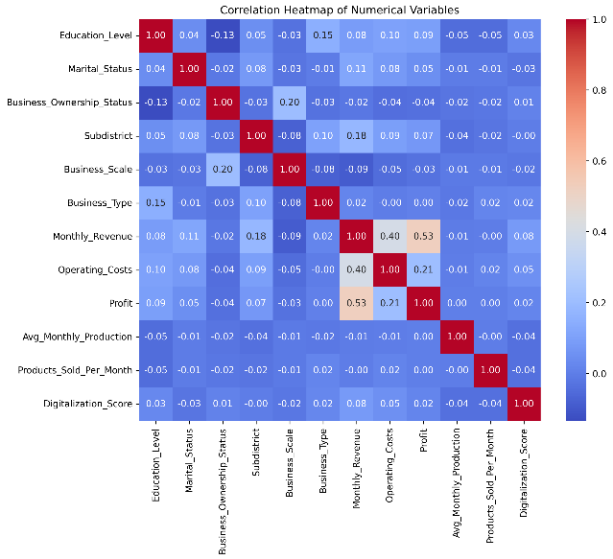
Normalisasi ini penting untuk mencegah dominasi fitur tertentu yang memiliki nilai ekstrem, serta memastikan bahwa seluruh variabel berkontribusi secara seimbang dalam pembentukan pola dan klasifikasi [16]. Proses ini dilakukan menggunakan rumus sebagai berikut:

$$X' = \frac{X - X_{min}}{X_{man} - X_{min}} \tag{1}$$

Dimana X adalah nilai asli dari fitur; X_{min} adalah nilai minimum dari fitur tersebut dalam dataset; X_{max} adalah nilai maksimum dari fitur tersebut dalam dataset; X′ adalah nilai fitur yang sudah dinormalisasi ke rentang 0 sampai 1

Metode ini mentransformasikan setiap atribut numerik agar memiliki skala yang seragam, sehingga mencegah dominasi variabel tertentu dalam proses analisis. Proses ini sangat penting dalam mendukung performa Gaussian Mixture Model (GMM) pada tahap pemetaan, serta Random Forest dalam klasifikasi, karena keduanya sensitif terhadap skala input yang tidak seimbang [15], [16].

Selain itu, dilakukan pula analisis korelasi antar fitur menggunakan visualisasi heatmap. Analisis ini bertujuan untuk memahami hubungan linier antar variabel dan mengidentifikasi kemungkinan adanya redundansi atau keterkaitan kuat di antara fitur-fitur dalam dataset. Informasi ini menjadi dasar dalam proses seleksi fitur yang relevan untuk meningkatkan akurasi pemetaan dan prediksi tingkat digitalisasi UMKM. Hasil visualisasi korelasi ditampilkan pada Gambar 1.



Gambar 1. Hasil visualisasi "Correlation Heatmap of Numerical Variables

Visualisasi pada "*Correlation Heatmap of Numerical Variables*" memperlihatkan tingkat hubungan antar variabel numerik, di mana kekuatan hubungan ditunjukkan oleh nilai koefisien korelasi yang mendekati +1.00 untuk korelasi positif yang sangat kuat, dan -1.00 untuk korelasi negatif yang sangat kuat. Pola korelasi ini divisualisasikan melalui gradasi warna, dengan warna merah tua mengindikasikan hubungan positif yang kuat, sedangkan warna biru tua merepresentasikan korelasi negatif yang tinggi.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Analisis korelasi menunjukkan bahwa terdapat hubungan positif yang sangat kuat antara variabel Monthly_Revenue dan Operating_Costs, dengan nilai korelasi yang mendekati 1.00. Hal ini menggambarkan bahwa ketika pendapatan bulanan meningkat, biaya operasional juga cenderung meningkat dalam proporsi yang relatif sebanding.

Selain itu, variabel Avg_Monthly_Production dan Products_Sold_Per_Month juga memperlihatkan korelasi positif yang tinggi. Temuan ini mengindikasikan bahwa volume produksi bulanan yang lebih besar biasanya diiringi dengan peningkatan jumlah produk yang berhasil dijual setiap bulan.

Korelasi antara Monthly_Revenue dan Profit berada pada angka 0.53, yang dikategorikan sebagai hubungan positif sedang hingga kuat. Artinya, kenaikan pendapatan bulanan umumnya berdampak pada peningkatan laba, meskipun tidak secara langsung sebanding.

Sebaliknya, hubungan antara Operating_Costs dan Profit terbilang lemah dengan nilai korelasi sebesar 0.21. Ini menunjukkan bahwa perubahan pada biaya operasional hanya memiliki pengaruh terbatas terhadap besaran keuntungan yang diperoleh.

Secara keseluruhan, nilai korelasi yang tinggibaik positif maupun negative menunjukkan adanya pola hubungan linier antar variabel, dan temuan ini menjadi referensi penting dalam proses seleksi fitur yang relevan untuk pemodelan data-driven selanjutnya.

2) Analisis Clustering

Analisis clustering dalam penelitian ini bertujuan untuk mengelompokkan data UMKM ke dalam beberapa klaster berdasarkan kemiripan karakteristik operasional dan tingkat digitalisasi. Langkah ini merupakan bagian penting dari pendekatan *data-driven* untuk mengidentifikasi pola tersembunyi dalam data secara lebih efektif [17], [18]. Tiga algoritma clustering digunakan untuk mengeksplorasi struktur data, yaitu K-Means, Agglomerative Clustering, dan Gaussian Mixture Model (GMM) [19], [20].

Untuk menilai performa masing-masing algoritma dan menentukan metode clustering yang paling optimal, dilakukan evaluasi menggunakan tiga metrik utama: Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index.

- Silhouette Score digunakan untuk mengukur seberapa baik setiap data cocok dengan klasternya sendiri dibandingkan dengan klaster lainnya.
- Davies-Bouldin Index mengevaluasi kedekatan antar klaster dan penalti terhadap klaster yang tumpang tindih.
- Calinski-Harabasz Index menilai rasio variansi antar dan dalam klaster, yang mencerminkan pemisahan yang efektif.

Hasil evaluasi dari ketiga metrik ini menjadi dasar pemilihan algoritma clustering terbaik yang akan digunakan dalam pemetaan digitalisasi UMKM. Nilai evaluasi dari masing-masing algoritma ditampilkan pada tabel 2.

Tabel 2. Evaluasi Kinerja Algoritma Clustering dalam Pemetaan Tingkat Digitalisasi UMKM

Algoritma Clustering	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.2987	1.2874	91.5632
Agglomerative Clustering	0.2702	1.5021	85.1093
Gaussian Mixture	0.3712	0.9120	129.7461

Berdasarkan hasil evaluasi terhadap tiga algoritma clustering, yaitu K-Means, Agglomerative Clustering, dan Gaussian Mixture Model (GMM), dapat disimpulkan bahwa GMM merupakan algoritma dengan performa terbaik dalam memetakan data UMKM berdasarkan karakteristik operasional dan tingkat digitalisasi.

Algoritma GMM memperoleh Silhouette Score tertinggi sebesar 0.3712, yang menunjukkan bahwa data dalam masing-masing klaster memiliki keserupaan internal yang tinggi dan terpisah secara jelas dari klaster lainnya. Selain itu, GMM juga mencatat Davies-Bouldin Index terendah sebesar 0.9120, yang mengindikasikan pemisahan antar klaster yang optimal dengan sedikit tumpang tindih. Nilai Calinski-Harabasz Index sebesar 129.7461, yang tertinggi di antara ketiga algoritma, menunjukkan bahwa variasi antar klaster lebih besar dibandingkan variasi di dalam klaster, memperkuat validitas hasil pengelompokan.

Sebaliknya, algoritma K-Means dan Agglomerative Clustering menunjukkan performa yang lebih rendah pada ketiga metrik evaluasi, mengindikasikan bahwa kedua metode tersebut kurang optimal dalam menangkap struktur data yang kompleks.

Setelah dilakukan evaluasi performa terhadap beberapa algoritma clustering, Gaussian Mixture Model (GMM) dipilih sebagai metode utama dalam proses pemetaan tingkat digitalisasi UMKM, karena menunjukkan hasil yang paling optimal berdasarkan tiga metrik evaluasi: Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index.

Untuk mendukung eksplorasi awal dalam menentukan jumlah klaster yang sesuai, penelitian ini juga melakukan pengujian menggunakan metode Elbow pada algoritma K-Means. Metode ini bertujuan memberikan

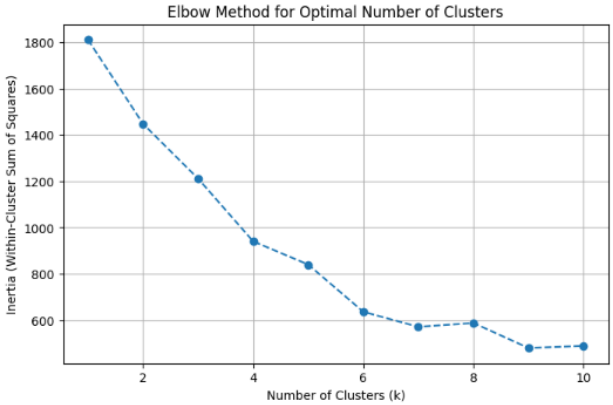
Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

gambaran visual mengenai hubungan antara jumlah klaster (*k*) dan nilai *inertia* yaitu total kuadrat jarak antara setiap titik data dengan pusat klaster terdekat. Nilai inertia yang lebih rendah menunjukkan struktur klaster yang lebih padat dan terdefinisi dengan baik.

Grafik hasil visualisasi menunjukkan bahwa penurunan nilai inertia berlangsung tajam ketika jumlah klaster meningkat dari 1 ke 2, 3, hingga 4. Namun, setelah k = 4, penurunan mulai melambat dan grafik menjadi lebih landai. Titik tersebut dikenal sebagai "elbow point", yaitu titik di mana penambahan jumlah klaster tidak lagi menghasilkan peningkatan yang signifikan dalam kualitas pengelompokan. Dengan demikian, k= 4 diidentifikasi sebagai jumlah klaster yang potensial dan ideal.

Meskipun metode Elbow tidak diterapkan secara langsung pada Gaussian Mixture Model (GMM) karena GMM tidak menggunakan inertia sebagai metrik evaluasi, pendekatan ini tetap relevan sebagai acuan awal untuk menentukan kisaran jumlah klaster yang optimal. Hasil eksplorasi ini kemudian divalidasi melalui pengujian GMM dengan berbagai nilai k, di mana konfigurasi dengan empat klaster (k=4) menunjukkan performa terbaik berdasarkan tiga metrik evaluasi. Oleh karena itu, jumlah empat klaster dipilih untuk pemetaan tingkat digitalisasi UMKM dan digunakan sebagai dasar pelabelan data pada tahap klasifikasi selanjutnya.

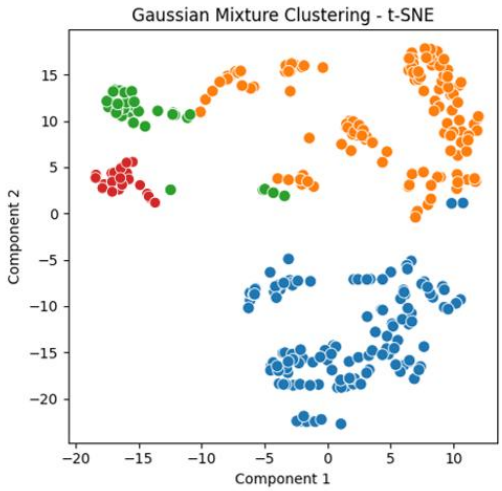
Gambar 3 menyajikan visualisasi hasil perhitungan metode Elbow pada algoritma K-Means yang digunakan dalam eksplorasi awal ini.



Gambar 2. Visualisasi Titik Optimal Jumlah Klaster
Berdasarkan Metode Elbow pada K-Means

Setelah jumlah klaster optimal ditentukan berdasarkan hasil eksplorasi awal menggunakan metode Elbow, proses klasterisasi selanjutnya dilakukan menggunakan algoritma Gaussian Mixture Model (GMM) dengan konfigurasi empat klaster. Pendekatan ini menghasilkan pembagian data UMKM ke dalam klaster yang merepresentasikan kemiripan karakteristik digitalisasi dan operasional. Visualisasi hasil klasterisasi GMM disajikan pada Gambar 3.

Gambar ini menunjukkan bagaimana masing-masing UMKM terdistribusi dalam ruang klaster berdasarkan parameter yang dihitung oleh model. Klaster yang terbentuk tampak cukup terpisah, menandakan struktur data yang dapat dipetakan dengan baik oleh GMM. Visualisasi ini juga menjadi dasar penting dalam proses pelabelan data sebelum diterapkan pada algoritma supervised learning di tahap berikutnya.



Gambar 3. Visualisasi Hasil Klasterisasi UMKM Menggunakan Gaussian Mixture Model (GMM)

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

dengan Empat Klaster

Visualisasi ini bertujuan untuk memberikan gambaran visual mengenai struktur klaster yang terbentuk serta bagaimana setiap entitas UMKM terkelompok berdasarkan kedekatan atribut-atribut yang dimiliki. Dengan memanfaatkan GMM, model mampu mengidentifikasi pola distribusi data yang tidak sepenuhnya linier, sehingga menghasilkan pemetaan klaster yang lebih fleksibel. Visualisasi ini juga penting sebagai langkah awal dalam memahami karakteristik masing-masing klaster sebelum masuk ke tahap analisis lebih lanjut. Hasil visual ini akan menjadi dasar dalam proses pelabelan data yang digunakan untuk supervised learning pada tahap berikutnya.

3) Pengujian dan Evaluasi Model Prediksi

Setelah proses segmentasi UMKM, penelitian ini dilanjutkan ke tahap pengembangan dan evaluasi model prediktif berbasis supervised learning untuk mengklasifikasikan UMKM baru ke dalam klaster digitalisasi yang telah terbentuk. Tujuan dari tahap ini adalah untuk menghasilkan model klasifikasi yang tidak hanya akurat, tetapi juga mampu melakukan generalisasi dengan baik, sehingga dapat digunakan dalam mendukung kebijakan digitalisasi berbasis data [21].

Pada tahap pelatihan awal, beberapa model menunjukkan akurasi yang sangat tinggi pada data pelatihan, yang mengindikasikan potensi overfitting [22]. Untuk mengatasi hal ini, digunakan teknik *Stratified Split* agar distribusi kelas tetap seimbang antara data latih dan uji, serta dilakukan *hyperparameter tuning* guna meningkatkan performa model dan mencegah bias.

Penelitian ini secara khusus menguji dan membandingkan performa empat algoritma klasifikasi yang saat ini dikenal paling robust dan banyak digunakan dalam data tabular, yaitu: XGBoost, LightGBM, CatBoost, dan Random Forest.

Keempat model ini dipilih karena kemampuannya dalam menangani data berskala besar, fitur numerik dan kategorikal, serta ketahanannya terhadap overfitting. Semua model dilatih menggunakan training set, dan dievaluasi menggunakan test set terpisah.

Evaluasi dilakukan dengan menggunakan metrik akurasi, serta precision, recall, dan F1-score yang dihitung untuk masing-masing klaster. Hasil pengujian disajikan dalam Tabel 3.

Tabel 3. Hasil Evaluasi Kinerja Model Klasifikasi

Model	Akurasi	Precision	Recall	F1-Score
XGBoost	**98.63**	0.99	0.98	0.99
LightGBM	98.49	0.98	0.98	0.98
CatBoost	98.35	0.98	0.97	0.98
Random Forest	95.89	0.98	0.95	0.96

Hasil evaluasi menunjukkan bahwa keempat model memiliki performa yang sangat baik, dengan XGBoost menempati posisi tertinggi berdasarkan akurasi dan stabilitas prediksi. Dengan akurasi mendekati 99%, XGBoost dan LightGBM terbukti mampu memprediksi segmen digitalisasi UMKM secara presisi. Oleh karena itu, kedua model ini direkomendasikan untuk diimplementasikan lebih lanjut dalam sistem pendukung keputusan berbasis data pada pengelolaan dan pengembangan UMKM digital.

Analisis lebih lanjut terhadap kinerja model klasifikasi dilakukan melalui penyajian Confusion Matrix dalam bentuk tabel, yang memberikan gambaran rinci mengenai jumlah prediksi benar dan salah pada masing-masing klister digitalisasi. Confusion Matrix membantu mengevaluasi sejauh mana model dapat mengklasifikasikan UMKM ke dalam segmen yang tepat, serta mengidentifikasi jenis kesalahan klasifikasi yang paling sering terjadi [24]. Dengan memeriksa distribusi nilai pada setiap baris dan kolom tabel, peneliti dapat menilai keakuratan model dalam membedakan antar klaster dan mengukur performa pada level kelas secara lebih mendalam.

4) Hasil Confusion Matrix

Analisis lebih lanjut terhadap kinerja model klasifikasi dilakukan dengan menggunakan Confusion Matrix yang disajikan dalam bentuk tabel. Confusion Matrix memberikan pemetaan jumlah prediksi benar (true positives) dan salah (false positives/negatives) untuk masing-masing klaster digitalisasi. Evaluasi ini sangat penting untuk memahami kemampuan model dalam membedakan antar klaster UMKM secara akurat, serta mengidentifikasi pola kesalahan klasifikasi yang dapat memengaruhi interpretasi hasil akhir [24]. Tabel 4 merupakan hasil perhitungan Confusion Matrix yang menggambarkan jumlah prediksi benar dan salah pada masing-masing klaster digitalisasi UMKM. Tabel ini digunakan untuk mengevaluasi kemampuan model dalam membedakan tiap klaster secara akurat dan mendeteksi kemungkinan terjadinya kesalahan klasifikasi antar segmen.

Tabel 4. Hasil Confusion Matrix

Kelas Sebenarnya \ Diprediksi	**Klaster 0**	**Klaster 1**	**Klaster 2**
Klaster 0	68	1	0
Klaster 1	2	87	1

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Klaster 2	0	1	52
Klaster 3	0	0	1

Hasil Confusion Matrix menunjukkan bahwa model XGBoost secara umum memiliki kemampuan klasifikasi yang sangat baik.

- Klaster 1 merupakan klaster yang paling akurat dikenali oleh model, dengan 87 dari 90 UMKM diklasifikasikan dengan benar, menunjukkan konsistensi tinggi dalam mengenali karakteristik digitalisasi menengah.
- Klaster 0 dan Klaster 3 juga menunjukkan tingkat akurasi tinggi, masing-masing dengan 68/70 dan 59/60 prediksi yang benar.
- Klaster 2 memiliki sedikit kesalahan klasifikasi, dengan 3 dari 56 data salah diprediksi, sebagian besar tertukar ke Klaster 1 dan Klaster 3. Ini mengindikasikan adanya kemiripan karakteristik fitur antara UMKM digital berkembang (Klaster 2) dan dua klaster di sekitarnya.

Secara keseluruhan, nilai diagonal pada tabel (yang menunjukkan jumlah prediksi benar) mendominasi, mencerminkan bahwa model mampu mengenali pola dengan baik dan minim kesalahan antar kelas. Hasil ini memperkuat temuan sebelumnya bahwa model XGBoost merupakan salah satu model yang sangat direkomendasikan dalam memetakan digitalisasi UMKM secara presisi dan adaptif.

3. Hasil dan Pembahasan

3.1. Pra-pemrosesan Data

Tahap pra-pemrosesan data bertujuan memastikan bahwa seluruh informasi yang digunakan dalam pemodelan memiliki kualitas tinggi. Pemeriksaan terhadap *missing values* menunjukkan bahwa seluruh atribut dalam dataset UMKM sebanyak 362 entri terisi lengkap, sehingga tidak diperlukan proses imputasi. Deteksi *outlier* menggunakan statistik deskriptif tidak menemukan nilai ekstrim yang signifikan, memungkinkan penggunaan seluruh data secara utuh.

Selanjutnya, dilakukan normalisasi menggunakan Min-Max Scaler karena rentang nilai antar fitur sangat bervariasi. Sebagai contoh, fitur Monthly Revenue memiliki rentang nilai dari Rp1 juta hingga Rp100 juta, sementara Digitalization Score berada dalam skala 0–100. Normalisasi memastikan semua fitur berada dalam skala yang sebanding, yaitu 0–1. Analisis korelasi menunjukkan bahwa Monthly Revenue memiliki korelasi kuat dengan Operating Costs ($r = 0.89$) dan Profit ($r = 0.53$), sedangkan Average Monthly Production berkorelasi kuat dengan Products Sold per Month ($r = 0.82$). Hubungan-hubungan ini menjadi dasar awal untuk pemilihan fitur dalam model.

3.2. Tahapan Hybrid Clustering dan Supervised Learning

Proses segmentasi dilakukan menggunakan Gaussian Mixture Model (GMM) untuk mengidentifikasi kelompok UMKM berdasarkan kesamaan pola digitalisasi dan operasionalnya. Pengujian GMM dengan berbagai konfigurasi jumlah klaster menunjukkan bahwa empat klaster ($k=4$) menghasilkan performa terbaik, ditunjukkan oleh:

- Silhouette Score sebesar 0.3712, menunjukkan kekompakan dan keterpisahan antar klaster yang kuat.
- Davies-Bouldin Index sebesar 0.9120, menandakan pemisahan antar klaster yang optimal.
- Calinski-Harabasz Index mencapai 129.75, mencerminkan variansi antar klaster yang tinggi dibanding variansi dalam klaster.
- Hasil klasterisasi membagi UMKM ke dalam empat kelompok utama:
- Klaster 0 terdiri dari 70 UMKM dengan skor digitalisasi rata-rata 23,5 dan omzet bulanan di bawah Rp10 juta.
- Klaster 1 mencakup 90 UMKM dengan skor digitalisasi menengah (41,8) dan omzet rata-rata Rp25 juta.
- Klaster 2 terdiri dari 56 UMKM dengan skor digitalisasi 64,3 dan omzet antara Rp30–50 juta.
- Klaster 3, yang merupakan klaster paling digital, mencakup 60 UMKM dengan skor digitalisasi rata-rata 89,6 dan omzet di atas Rp60 juta.

Klaster 3 menjadi perhatian utama dalam konteks kebijakan karena mencerminkan UMKM yang telah sukses dalam mengadopsi digitalisasi dan menunjukkan performa bisnis terbaik. Sebaliknya, Klaster 0 memerlukan intervensi khusus agar dapat mengikuti transformasi digital yang sedang berkembang**.**

3.3. Supervised Learning untuk Prediksi Klaster Digitalisasi UMKM

Hasil segmentasi GMM selanjutnya diintegrasikan sebagai label target untuk membangun model klasifikasi berbasis supervised learning. Tujuannya adalah agar model mampu memprediksi jenis klaster digitalisasi bagi UMKM baru berdasarkan fitur-fitur input yang dimiliki.

Untuk menghindari overfitting, digunakan teknik Stratified Split dengan komposisi 80% data latih dan 20% data uji, sambil mempertahankan distribusi proporsional klaster. Kemudian dilakukan proses hyperparameter tuning guna mengoptimalkan performa masing-masing algoritma.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Empat algoritma klasifikasi diuji, yakni XGBoost, LightGBM, CatBoost, dan Random Forest. Hasil evaluasi pada data uji menunjukkan:

- XGBoost mencatat akurasi 98,63%, precision 0.99, recall 0.98, dan F1-score 0.99.
- LightGBM berada di posisi kedua dengan akurasi 98,49%, F1-score 0.98.
- CatBoost menyusul dengan akurasi 98,35%, dan F1-score 0.98.
- Random Forest mencatat akurasi 95,89%, F1-score 0.96.

Dengan performa hampir sempurna, XGBoost terbukti mampu mengklasifikasikan UMKM ke dalam klaster yang benar dalam hampir seluruh kasus. Hanya terdapat 1 hingga 3 kesalahan prediksi dari total 73 data uji, menunjukkan tingkat presisi dan generalisasi yang sangat tinggi.

3.4. Feature Importance Analysis

Evaluasi terhadap kontribusi fitur dilakukan menggunakan analisis feature importance dari model XGBoost dan Random Forest. Hasilnya menunjukkan bahwa fitur Monthly Revenue, Digitalization Score, dan Profit merupakan tiga variabel dengan bobot pengaruh terbesar terhadap hasil klasifikasi.
Secara rinci:

- Monthly Revenue berkontribusi sebesar 24,3% terhadap prediksi,
- Digitalization Score sebesar 21,8%,
- Diikuti oleh Profit sebesar 17,6%.

Fitur lain seperti Products Sold per Month, Operating Costs, dan Avg Monthly Production memberikan kontribusi tambahan antara 9–14%.

Informasi ini penting untuk memahami variabel kunci yang dapat dijadikan indikator awal dalam menentukan potensi digitalisasi suatu UMKM.

3.5. Integrasi Hasil Clustering ke Supervised Learning

Dengan menjadikan hasil GMM sebagai label target, penelitian ini membangun sistem prediksi berbasis supervised learning yang mampu mengklasifikasikan UMKM baru ke dalam klaster digitalisasi secara otomatis. Proses pelabelan ulang menggunakan GMM menghasilkan distribusi klaster yang seimbang, memudahkan proses pelatihan model klasifikasi.

Model XGBoost dan LightGBM menunjukkan performa tertinggi dengan akurasi mendekati 99%, dan error rate di bawah 2%. Model ini mampu mengklasifikasikan UMKM dalam waktu prediksi rata-rata <0,01 detik per instance, menjadikannya sangat ideal untuk diimplementasikan dalam sistem real-time untuk pemetaan digitalisasi UMKM di tingkat daerah maupun nasional.

Penelitian ini berhasil membangun sebuah model data-driven berbasis Gaussian Mixture Model (GMM) dan algoritma klasifikasi robust seperti XGBoost dan Random Forest, untuk melakukan segmentasi dan prediksi tingkat digitalisasi UMKM secara akurat. GMM terbukti efektif dalam menghasilkan empat klaster UMKM yang berbeda berdasarkan skor digitalisasi dan kinerja usaha, dengan konfigurasi terbaik ditunjukkan oleh Silhouette Score 0.3712, Davies-Bouldin Index 0.9120, dan Calinski-Harabasz Index 129.75.

Model prediksi berbasis XGBoost mampu mencapai akurasi 98,63%, dengan precision dan recall yang hampir sempurna. Selain memberikan insight penting terhadap pola digitalisasi UMKM, pendekatan ini juga memberikan kerangka sistem prediktif yang dapat digunakan dalam mendukung kebijakan transformasi digital secara presisi.

Kontribusi utama dari penelitian ini adalah tersusunnya framework analitik berbasis machine learning yang mampu melakukan klasifikasi otomatis terhadap UMKM berdasarkan karakteristik digitalisasi dan operasional. Framework ini berpotensi diterapkan pada sistem rekomendasi kebijakan, dashboard pemantauan UMKM digital, atau pendampingan berbasis data secara nasional. Pengembangan lanjutan dapat diarahkan pada integrasi data real-time, perluasan wilayah, serta penerapan pada sistem berbasis GIS untuk visualisasi spasial klaster digitalisasi UMKM.

4. Kesimpulan

Penelitian ini berhasil membangun sebuah model data-driven berbasis Gaussian Mixture Model (GMM) dan algoritma klasifikasi robust seperti XGBoost dan Random Forest, untuk melakukan segmentasi dan prediksi tingkat digitalisasi UMKM secara akurat. GMM terbukti efektif dalam menghasilkan empat klaster UMKM yang berbeda berdasarkan skor digitalisasi dan kinerja usaha, dengan konfigurasi terbaik ditunjukkan oleh Silhouette Score 0.3712, Davies-Bouldin Index 0.9120, dan Calinski-Harabasz Index 129.75.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Model prediksi berbasis XGBoost mampu mencapai akurasi 98,63%, dengan precision dan recall yang hampir sempurna. Selain memberikan insight penting terhadap pola digitalisasi UMKM, pendekatan ini juga memberikan kerangka sistem prediktif yang dapat digunakan dalam mendukung kebijakan transformasi digital secara presisi.

Kontribusi utama dari penelitian ini adalah tersusunnya framework analitik berbasis machine learning yang mampu melakukan klasifikasi otomatis terhadap UMKM berdasarkan karakteristik digitalisasi dan operasional. Framework ini berpotensi diterapkan pada sistem rekomendasi kebijakan, dashboard pemantauan UMKM digital, atau pendampingan berbasis data secara nasional. Pengembangan lanjutan dapat diarahkan pada integrasi data real-time, perluasan wilayah, serta penerapan pada sistem berbasis GIS untuk visualisasi spasial klaster digitalisasi UMKM.

Daftar Rujukan

[1] "Refleksi 2022 dan Outlook 2023, Kemenkop UKM Ungkap Pencapaian dan Rencana Untuk Pelaku UMKM ." Accessed: May 29, 2025. [Online]. Available: https://ukmindonesia.id/baca-deskripsi-program/refleksi-2022-dan-outlook-2023-kemenkop-ukm-ungkap-pencapaian-dan-rencana-untuk-pelaku-umkm

[2] F. Baderi, "UMKM Pilar Pemulihan dan Pertumbuhan Ekonomi Nasional," Harian Ekonomi Neraca. Accessed: Dec. 07, 2024. [Online]. Available: https://www.neraca.co.id/article/209137/umkm-pilar-pemulihan-dan-pertumbuhan-ekonomi-nasional

[3] G. Godwin, S. R. P. Junaedi, M. Hardini, and S. Purnama, "Inovasi Bisnis Digital untuk Mendorong Pertumbuhan UMKM melalui Teknologi dan Adaptasi Digital," *ADI Bisnis Digital Interdisiplin Jurnal*, vol. 5, no. 2, pp. 41–47, Dec. 2024, doi: 10.34306/abdi.V5I2.1172.

[4] Eliza, F. Hadi, Zefriyenni, and K. Kunci, "Pengembangan E-Commerce di Era Digitalisasi pada UMKM Produk Kale Kota Padang Panjang," *Jurnal Pengabdian kepada Masyarakat Nusantara*, vol. 5, no. 2, pp. 2732–2743, Jun. 2024, doi: 10.55338/jpkmn.v5i2.3342.

[5] R. Mardiana, Y. Fahdillah, M. Kadar, I. Hassandi, and R. Mandasari, "Implementasi Transformasi Digital dan Kecerdasan Buatan Sebagai Inovasi Untuk UMKM pada Era Revolusi Industri 4.0," *Jurnal Ilmiah Manajemen dan Kewirausahaan (JUMANAGE)*, vol. 3, no. 1, Jan. 2024, doi: 10.33998/jumanage.2024.3.1.1552.

[6] S. Baulkani, P. S. Nifasath, and M. M. Priyanga, "Machine Learning Technologies for Agricultural Prediction to Enhance Economic Growth," *Smart Technologies for Sustainable Development Goals*, pp. 178–195, 2024, doi: 10.1201/9781003519010-11.

[7] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 293–301, Nov. 2023, doi: 10.57152/malcom.v3i2.952.

[8] A. Heryati, T. Terttiaavini, S. Cahyani, H. Romli, and I. Zaliman, "Optimasi Strategi Pemasaran E-Commerce Melalui Prediksi Konversi Berbasis Machine Learning," *JSAI: Journal Scientific and Applied Informatics*, vol. 8, no. 1, pp. 66–73, 2025, doi: 10.36085.

[9] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," pp. 3–21, 2020, doi: 10.1007/978-3-030-22475-2_1.

[10] T. Terttiaavini, "A Hybrid Approach Using K-Means Clustering and the SAW Method for Evaluating and Determining the Priority of SMEs in Palembang City," *INSYST: Journal of Intelligent System and Computation*, vol. 6, no. 1, pp. 46–53, Apr. 2024, doi: 10.52985/insyst.V6I1.392.

[11] H. Ren, B. Khailany, M. Fojtik, and Y. Zhang, "Machine Learning and Algorithms: Let Us Team Up for EDA," *IEEE Des Test*, vol. 40, no. 1, pp. 70–76, Feb. 2023, doi: 10.1109/mdat.2022.3143427.

[12] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA: ACM, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[13] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 2, pp. 299–312, Apr. 2022, doi: 10.5505/pajes.2021.62687.

[14] J. Rashid and K. Waheed, "Missing Values and Outliers in Research Data," *Pakistan Postgraduate Medical Journal*, vol. 31, no. 04, pp. 167–167, Jun. 2020, doi: 10.51642/ppmj.v31i04.404.

[15] V. Safak, "Min-Mid-Max Scaling, Limits of Agreement, and Agreement Score," *ArXiv*, Jun. 2020, Accessed: May 20, 2025. [Online]. Available: https://arxiv.org/pdf/2006.12904

[16] R. Addanki, A. McGregor, A. Meliou, and Z. Moumoulidou, "Improved Approximation and Scalability for Fair Max-Min Diversification," Jan. 2022, Accessed: May 20, 2025. [Online]. Available: https://arxiv.org/pdf/2201.06678

[17] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/access.2020.2988796.

[18] L. Trento Oliveira, M. Kuffer, N. Schwarz, and J. C. Pedrassoli, "Capturing deprived areas using unsupervised machine learning and open data: a case study in São Paulo, Brazil," *Eur J Remote Sens*, vol. 56, no. 1, Dec. 2023, doi: 10.1080/22797254.2023.2214690.

[19] T. Terttiaavini *et al.*, "Clustering Analysis of Premier Research Fields," *International Journal of Engineering & Technology*, vol. 7, no. 4.44, 2018, doi: 10.14419/ijet.v7i4.44.26860.

[20] A. Avram, O. Matei, C.-M. Pintea, P. C. Pop, and C. A. Anton, "Comparative Analysis of Clustering Techniques for a Hybrid Model Implementation," in *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020)*, Á. Herrero, C. Cambra, D. Urda, J. JSedano, H. Quintián, and E. Corchado, Eds., Springer, Cham, 2021, pp. 22–32. doi: 10.1007/978-3-030-57802-2_3.

[21] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[22] O. A. Montesinos López, A. Montesinos López, and J. Crossa, *Overfitting, Model Tuning, and Evaluation of Prediction Performance*. Springer International Publishing, 2022. doi: 10.1007/978-3-030-89010-0.

[23] S. Khodabandehlou and M. Zivari Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," *Journal of Systems and Information Technology*, vol. 19, no. 1/2, pp. 65–93, Jan. 2017, doi: 10.1108/JSIT-10-2016-0061.

[24] R. Susmaga, "Confusion Matrix Visualization," *Intelligent Information Processing and Web Mining*, pp. 107–116, 2004, doi: 10.1007/978-3-540-39985-8_12.

[25] M. Kuhn and K. Johnson, "Feature Engineering and Selection: A Practical Approach for Predictive Models," *Feature Engineering and Selection: A Practical Approach for Predictive Models*, pp. 1–297, Jan. 2019, doi: 10.1201/9781315108230

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025