VOL. 20. NO. 2 FEBRUARI 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v20i2.7416

STUDI KOMPARATIF ALGORITMA C4.5 DAN RANDOM FOREST PADA DIGITALISASI UMKM KABUPATEN TEGAL

Gita Iftah Royani[1]; Nur Syifa Amelia[2]; Marlina[3]; Fani Nurona Cahya[4*]

Sistem Informasi[1,2,3,4]
Universitas Bina Sarana Informatika, Jakarta, Indonesia[1,2,3,4]
https://www.bsi.ac.id/[1,2,3,4]
royanigitaiftah@gmail.com[1*], nursyifaamelia265@gmail.com[2],
marlina.mln@bsi.ac.id[3], fani.foc@bsi.ac.id[4]

(*) Corresponding Author

Abstract— *Digital transformation has become an essential necessity for Micro, Small, and Medium Enterprises (UMKM) to enhance their competitiveness in the era of Industry 4.0. However, in Tegal Regency, the level of digitalization adoption among MSMEs remains varied and tends to be low, thus requiring further investigation. This study aims to compare the performance of the C4.5 and Random Forest algorithms in classifying the level of digitalization of MSMEs in Tegal Regency. This research employs the CRISP-DM methodology, which includes business understanding, data understanding, data preparation, modeling, evaluation, and implementation. Primary data were collected through questionnaires distributed to 100 MSME respondents and processed using RapidMiner. The results indicate that the Random Forest algorithm demonstrates superior performance, achieving an average F1-score of 89.63%, accuracy of 91.43%, while the C4.5 algorithm records an average F1-score of 86.24%, accuracy of 90%. The highest F1-score for both algorithms is observed in the low digitalization category at 95%, which is consistent with the data distribution showing that the majority of MSMEs (59%) fall within this category. This study systematically integrates the CRISP-DM approach from business understanding to model implementation, resulting in a structured data analysis workflow that can be replicated by local governments or future researchers. Another novelty of this study lies in the finding that although Random Forest exhibits better classification performance than C4.5, the majority of MSMEs remain at a low level of digitalization. These results provide practical contributions as a basis for formulating more targeted and sustainable MSME digitalization policies at the local.*

Keywords*: C4.5, CRISP-DM, Digitalization, UMKM, Random Forest.*

Abstrak— Transformasi digital telah menjadi kebutuhan penting bagi Usaha Mikro, Kecil, dan Menengah (UMKM) untuk meningkatkan daya saing di era Industri 4.0. Namun, di Kabupaten Tegal, tingkat adopsi digitalisasi di kalangan UMKM masih bervariasi dan cenderung rendah, sehingga memerlukan investigasi lebih lanjut. Penelitian ini bertujuan untuk membandingkan kinerja algoritma C4.5 dan Random Forest dalam mengklasifikasikan tingkat digitalisasi UMKM di Kabupaten Tegal. Penelitian ini menggunakan metodologi CRISP-DM, yang meliputi pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Data primer dikumpulkan melalui kuesioner yang dibagikan kepada 100 responden UMKM dan diolah menggunakan RapidMiner. Hasil menunjukkan bahwa algoritma Random Forest menunjukkan kinerja yang lebih unggul, mencapai rata-rata F1-score sebesar 89,63%, akurasi 91,43%, sedangkan algoritma C4.5 mencatat rata-rata F1-score sebesar 86,24%, akurasi 90%. Skor F1 tertinggi untuk kedua algoritma diamati pada kategori digitalisasi rendah sebesar 95%, yang konsisten dengan distribusi data yang menunjukkan bahwa mayoritas UMKM (59%) termasuk dalam kategori ini. Studi ini secara sistematis mengintegrasikan pendekatan CRISP-DM dari pemahaman bisnis hingga implementasi model, menghasilkan alur kerja analisis data terstruktur yang dapat direplikasi oleh pemerintah daerah atau peneliti di masa mendatang. Kebaruan lain dari studi ini terletak pada temuan bahwa meskipun Random Forest menunjukkan kinerja klasifikasi yang lebih baik daripada C4.5, mayoritas UMKM tetap berada pada tingkat digitalisasi yang rendah. Hasil ini memberikan kontribusi

praktis sebagai dasar untuk merumuskan kebijakan digitalisasi UMKM yang lebih tepat sasaran dan berkelanjutan di tingkat lokal.

Kata kunci: *C4.5, CRISP-DM, Digitalisasi, UMKM, Random Forest.*

PENDAHULUAN

Perkembangan teknologi di era industri 4.0 mendorong seluruh sektor, termasuk usaha mikro, kecil, dan menengah (UMKM), untuk bertransformasi secara digital agar mampu bertahan dan bersaing. Di Indonesia, UMKM berperan signifikan dengan kontribusi lebih dari 61% terhadap PDB dan menyerap lebih dari 97% tenaga kerja (Kementerian Koperasi dan UKM, 2022). Meski demikian, tingkat pemanfaatan teknologi digital oleh UMKM masih bervariasi, dipengaruhi oleh keterbatasan pengetahuan, akses, serta budaya bisnis konvensional (Birgithri et al., 2024). strategi dan pendekatan yang tepat diperlukan agar UMKM dapat memahami serta menerapkan teknologi digital sesuai dengan kebutuhan usahanya. Pelatihan berkelanjutan dan pendampingan digital dapat menjadi salah satu solusi untuk mengatasi hambatan tersebut. Salah satu pendekatan yang dapat digunakan untuk mengkaji pemanfaatan digitalisasi adalah dengan teknologi *machine learning.*

Machine lerning, yang merupakan cabang dari teknologi kecerdasan buatan, memiliki kemampuan untuk mengolah data berukuran besar dan mendeteksi pola tertentu yang berguna untuk mendukung proses pengambilan keputusan. Penelitian sebelumnya juga jarang mengintegrasikan metodologi CRISP-DM secara lengkap sebagai kerangka kerja analisis, sehingga proses pengolahan data sering kali tidak terdokumentasi secara sistematis. Di sisi lain, masih terdapat keterbatasan penelitian yang menggunakan data primer langsung dari pelaku UMKM untuk mencerminkan kondisi riil di lapangan. Oleh karena itu, penelitian ini mengisi celah tersebut dengan menghadirkan analisis komparatif antara algoritma *C4.5 dan Random Forest* berbasis data primer UMKM di Kabupaten Tegal serta menerapkan CRISP-DM secara menyeluruh. Penelitian ini diharapkan dapat melengkapi kekurangan studi sebelumnya sekaligus menjadi rujukan empiris bagi pengembangan kebijakan dan penelitian lanjutan di bidang digitalisasi UMKM. Algoritma ini dipilih karena keduanya merupakan metode pohon keputusan dengan performa klasifikasi yang baik, di mana Random Forest menunjukkan kinerja yang lebih unggul dibandingkan C4.5 pada beberapa metrik evaluasi (Ismanto & Novalia, 2021). Selain itu, (Bhardwaj & Chaurasia, 2022) menekankan pentingnya pemilihan algoritma yang tepat untuk meningkatkan efisiensi dalam pengambilan keputusan berbasis data.

Tujuan dari penelitian ini adalah untuk melakukan analisis komparatif terhadap algoritma C4.5 dan Random Forest dalam mengklasifikasikan tingkat pemanfaatan digitalisasi pelaku UMKM di Kabupaten Tegal. Perbandingan dilakukan dengan menilai kinerja kedua algoritma berdasarkan metrik *accurcay* dan kemampuan generalisasi. Melalui penelitian ini diharapkan dapat diketahui algoritma yang paling efektif dan optimal dalam klasifikasi tingkat pemanfaatan digitalisasi, sehingga dapat menjadi dasar solusi nyata untuk mendorong peningkatan digitalisasi UMKM. Hasilnya diharapkan mendukung pertumbuhan ekonomi yang inklusif dan berkelanjutan serta memperluas penerapan analisis data melalui teknologi *machine learning*. Kebaruan penelitian ini terletak pada penerapan dan evaluasi algoritma Random Forest untuk memetakan tingkat digitalisasi UMKM pada konteks wilayah Kabupaten Tegal, yang hingga saat ini masih relatif terbatas dikaji secara spesifik dan berbasis data. Berbeda dengan penelitian sebelumnya yang umumnya hanya bersifat deskriptif atau menggunakan metode statistik konvensional, penelitian ini mengintegrasikan pendekatan machine learning untuk mengidentifikasi pola, faktor dominan, serta tingkat kesiapan digital UMKM secara lebih akurat dan adaptif terhadap kompleksitas data.

Selain itu, penelitian ini tidak hanya berfokus pada aspek teknologi, tetapi juga mengombinasikan faktor literasi digital, sumber daya manusia, pelatihan, pendampingan, serta infrastruktur dan akses internet ke dalam satu model prediksi yang komprehensif. Dengan demikian, hasil penelitian mampu memberikan gambaran yang lebih holistik mengenai hambatan dan potensi transformasi digital UMKM di daerah.

Kebaruan lainnya terletak pada **implikasi praktis hasil model**, yang dapat dimanfaatkan sebagai dasar perumusan kebijakan dan perancangan program pemberdayaan UMKM berbasis bukti (*evidence-based policy*). Model yang dihasilkan tidak hanya bersifat akademis, tetapi juga aplikatif, sehingga dapat membantu pemerintah daerah dan pemangku kepentingan dalam menentukan prioritas intervensi digitalisasi UMKM yang lebih tepat sasaran, berkelanjutan, dan kontekstual sesuai karakteristik wilayah.

VOL. 20. NO. 2 FEBRUARI 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v20i2.7416

BAHAN DAN METODE

Dataset

Dataset diperoleh melalui penyebaran kuesioner secara langsung kepada 100 responden pelaku UMKM di Kabupaten Tegal. Penentuan jumlah responden tersebut disesuaikan dengan ketersediaan data lapangan**,** keterbatasan waktu penelitian, serta tujuan penelitian yang berfokus pada analisis komparatif kinerja algoritma C4.5 dan Random Forest**,** bukan pada generalisasi populasi secara nasional. Proses pelabelan kelas dilakukan secara *rule-based* berdasarkan hasil kuesioner, sehingga setiap responden memiliki label tingkat digitalisasi yang konsisten dan terukur. Pendekatan ini memungkinkan algoritma machine learning mempelajari pola digitalisasi UMKM secara sistematis.

Indikator Kuesioner

Indikator kuesioner disusun berdasarkan indikator digitalisasi UMKM yang diadaptasi dari penelitian terdahulu dan kebijakan terkait transformasi digital UMKM. Indikator mencakup aspek penggunaan teknologi digital, pemasaran digital, pengelolaan data usaha, literasi digital, serta pemanfaatan platform digital dalam operasional usaha.

Reliabilitas Instrumen

Kuesioner telah melalui uji validitas isi (*content validity*) melalui penyesuaian indikator dengan literatur dan penelitian sebelumnya. Reliabilitas instrumen diuji menggunakan metode statistik (misalnya Cronbach's Alpha) untuk memastikan konsistensi internal item pertanyaan, dengan hasil yang menunjukkan instrumen berada pada kategori reliabel.

Populasi Penelitian

Populasi dalam penelitian ini adalah pelaku UMKM yang berada di Kabupaten Tegal dan masih aktif menjalankan usaha pada saat penelitian dilakukan.

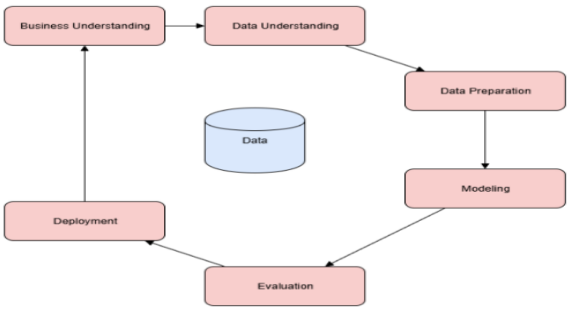
Teknik Sampling

Teknik sampling yang digunakan adalah *purposive sampling,* dengan kriteria responden merupakan pelaku UMKM yang memiliki aktivitas usaha aktif dan bersedia mengisi kuesioner penelitian. Teknik ini dipilih untuk memastikan data yang diperoleh relevan dengan tujuan penelitian dan sesuai dengan kebutuhan analisis komparatif algoritma.

Untuk memastikan kualitas data, kuesioner yang digunakan dalam penelitian ini telah melalui proses validasi isi *(content validity)* dengan mengacu pada indikator digitalisasi UMKM dari penelitian terdahulu. Selain itu, pengisian kuesioner dilakukan langsung oleh pelaku UMKM sehingga data yang diperoleh merepresentasikan kondisi aktual di lapangan. Meskipun ukuran dataset terbatas, data yang digunakan bersifat *primer dan relevan***,** sehingga tetap layak digunakan sebagai dasar analisis komparatif algoritma. Penelitian selanjutnya diharapkan dapat menggunakan dataset dengan jumlah responden yang lebih besar serta cakupan wilayah yang lebih luas untuk meningkatkan generalisasi hasil penelitian.

Metode CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah kerangka kerja standar dengan langkah-langkah yang terstruktur, mulai dari memahami kebutuhan bisnis hingga evaluasi dan penerapan hasil (Hasanah et al., 2021).



Sumber: (Hasil Penelitian, 2025)
Gambar 1. Tahapan CRISP-DM

Data pada penelitian ini telah terhubung secara sistematis dengan bagian lain dalam penelitian. Dataset yang diperoleh melalui kuesioner dijelaskan pada subbab Dataset, kemudian digunakan sebagai dasar pada tahapan *CRISP-DM,* khususnya pada tahap *Data Understanding*, *Data Preparation*, dan *Modeling*. Keterkaitan data juga tercermin pada proses pelabelan kelas tingkat digitalisasi UMKM yang selanjutnya dianalisis menggunakan algoritma C4.5 dan Random Forest. Untuk memperjelas keterhubungan tersebut, penulis telah memperbaiki redaksi dan menambahkan penjelasan transisi antar subbagian agar alur data dari pengumpulan hingga evaluasi model menjadi lebih eksplisit dan mudah dipahami.
Penelitian terdahulu, menerapkan algoritma K-Means *Clustering* untuk untuk mengelompokkan UMKM berdasarkan aspek asset dan omset. Dari hasil analisis, sebesar 53% data ke cluster 1, 40% data ke cluster 2, dan 7% data ke cluster 3.

1. *Business Understanding* (Pemahaman Bisnis)
   Dalam penelitian (Yudiana et al., 2023) dijelaskan bahwa tahap ini meliputi pemahaman tentang tujuan, kebutuhan, batasan, dan

VOL. 20. NO. 2 FEBRUARI 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

perspektif bisnis. Semua ini kemudian diubah menjadi definisi masalah dan strategi yang akan digunakan dalam proses *data mining*.

2. *Data Understanding* (Pemahaman Data)
   Tahap *Data Understanding* diawali dengan pengumpulan dan pemeriksaan data, identifikasi kualitas, serta penemuan pola awal untuk hipotesis (Kurniawan & Yasir, 2022).
3. *Data Preparation* (Persiapan Data)
   Tahap *Data Preparation* dilakukan dengan menyiapkan data agar layak dianalisis melalui pemilihan atribut, pembersihan, transformasi format, dan pembagian data (Dhewayani et al., 2022).
4. *Modeling* (Pemodelan)
   Tahap *Modeling* menerapkan teknik *data mining* yang paling sesuai dengan tujuan dan jenis data, kemudian melatih model dengan *data training* (Aria, 2025). Kinerja C4.5 dan Random Forest dibandingkan dengan metrik *accuracy*, *precision*, *recall*, dan *F1-Score*.

a. Confussion Matrix

Tabel 1. *Confussion Matrix*

	Predicted Class			
Actual Class		Yes	No	Total
	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

Sumber: (Hasil Penelitian, 2025)

Confusion matrix terdiri dari TP (data positif yang diklasifikasi benar), TN (data negatif yang diklasifikasi benar), FP (data negatif yang diklasifikasi salah sebagai positif), dan FN (data positif yang diklasifikasi salah sebagai negatif) (Muhammad Nur Ihsan Muhlashin & Stefanie, 2023).

b. Accuracy

Accuracy mengukur ketepatan prediksi model terhadap data aktual, dihitung dari rasio prediksi benar terhadap total prediksi (Argina, 2020).

c. Precision

Precision mengukur *accurcay* model dalam mengklasifikasikan data positif secara benar, yaitu rasio data positif yang tepat terhadap seluruh prediksi positif (Rininda et al., 2024).

d. Recall

Recall mengukur kemampuan model mendeteksi seluruh data positif, dihitung dari rasio prediksi positif yang benar terhadap total data positif aktual (Rininda et al., 2024).

e. F1-Score

F1-Score menggabungkan *precision* dan *recall* dalam satu metrik menggunakan rata-rata harmonik, untuk menyeimbangkan *accurcay* dan kelengkapan deteksi data positif (Rais et al., 2025). Penentuan nilai *Accuracy, Precision, Recall,* dan *F1-Score* dapat dilihat dari persamaan (Argina, 2020) berikut:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precission = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2\ \frac{precision\ x\ recall}{precision+recall} \quad (4)$$

5. *Deployment* (Implementasi)
   Pada tahap akhir ini, hasil dari proses pengolahan dan pengujian data dirangkum dalam sebuah laporan akhir. Kemudian disajikan dalam bentuk visual agar lebih mudah dipahami bagi pihak yang membutuhkan (Salsabila et al., 2021).

Metode Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini yaitu sebagai berikut:

1. Observasi
   Melakukan observasi ke Dinas Koperasi, UKM, dan Perdagangan Kabupaten Tegal untuk memperoleh data awal UMKM sebagai responden kuesioner serta menggali informasi singkat terkait program pelatihan digitalisasi yang pernah atau sedang dilaksanakan.
2. Kuesioner
   Kuesioner penelitian ini dibuat menggunakan *Google Form* dan disebarkan secara *online* seperti *WhatsApp* dan *Instagram*.
3. Studi Pustaka
   Pada tahap ini dilakukan penelusuran penelitian sebelumnya untuk mengumpulkan referensi yang relevan, seperti jurnal, buku, dan sumber lain yang mendukung penelitian ini (Abdullah, 2025).

Algoritma C4.5

Algoritma C4.5 digunakan untuk klasifikasi dalam *machine learning* dan *data mining* (Fajar et al., 2025). Algoritma ini memilih atribut terbaik dengan teknik *Gain Ratio*, yaitu menghitung *Entropy* dan *Gain* sebagai dasar pemisahan data (Abdullah, 2025). C4.5 juga mampu menangani data kontinu dan atribut tidak lengkap, serta menghasilkan pohon keputusan yang mudah dipahami (Melfia et

VOL. 20. NO. 2 FEBRUARI 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v20i2.7416

al., 2023).

$$Entropy\ (S) = -\sum_{i=1}^{n} \quad p_i\ .Iog_2\ (p_i) \quad (5)$$

Keterangan:
S = Himpunan kasus
n = Jumlah kelas atau partisi dalam *S*
p_i = Proporsi kasus pada kelas ke-*i* terhadap total kasus dalam *S*

$$Gain\quad (S,\quad a)\quad =\quad Entropy(S)\ -\ \sum_{i=1}^{n}\ \frac{|Si|}{|S|}.\ Entropy(S_i) \quad (6)$$

Keterangan:
S = Himpunan Kasus
a = Atribut yang dievaluasi
n = Jumlah nilai partisi dari atribut *a*
$|S_i|$ = Jumlah kasus dalam partisi ke-*i*
$|S|$ = Jumlah kasus dalam *S*

Algoritma Random Forest

Random Forest merupakan algoritma yang fleksibel dan praktis, menghasilkan klasifikasi akurat melalui kombinasi prediksi dari banyak pohon keputusan acak menggunakan mekanisme *voting* mayoritas (Khasanah et al., 2021). Mekanisme seleksi fiturnya membantu memilih atribut terbaik dan efektif untuk data berukuran besar (Supriyadi et al., 2020). Penentuan atribut menggunakan perhitungan *Entropy* dan *Information Gain* seperti persamaan berikut:

$$Entropy\ (Y) = -\sum_{i}\quad p(c|Y)\ log_2\ p\ (c|Y) \quad (7)$$

Keterangan:
Y = Himpunan Kasus
p (c|Y) = Proporsi nilai Y terhadap kelas c.

$$Information\ Gain\ (Y, a) = Entropy\ (Y) - \sum_{v}\quad \varepsilon values(a)\frac{Yv}{Ya}\ Entropy\ (Y_v) \quad (8)$$

Keterangan:
Values(a) = Nilai yang mungkin dalam himpunan kasus *a.*
Y_v = Subkelas dari *Y* dengan kelas *v* yang berhubungan dengan kelas *a.*
Y_a = Semua nilai yang sesuai dengan *a.*

HASIL DAN PEMBAHASAN

Business Understanding

Tahap awal bertujuan memahami tujuan bisnis, yaitu mengklasifikasikan tingkat digitalisasi (tinggi, sedang, rendah) UMKM di Kabupaten Tegal.

Data Understanding

Dataset penelitian ini diperoleh dari kuesioner yang dijawab oleh 100 pelaku UMKM di Kabupaten Tegal. Data digunakan untuk menganalisis dan membandingkan kinerja algoritma C4.5 dan Random Forest dalam mengklasifikasikan tingkat digitalisasi, dengan evaluasi berbasis *accuracy*, *precision*, *recall*, dan *F1-Score.*

Dataset memuat 11 atribut hasil pengolahan dari 20 variabel input, meliputi informasi umum UMKM (nama, jenis usaha, lama berdiri, jumlah karyawan) dan indikator penggunaan teknologi digital, seperti kepemilikan media sosial, *online shop*, aplikasi keuangan, pelatihan digital, serta dampak digitalisasi. Satu atribut digunakan sebagai label target, yaitu tingkat digitalisasi UMKM.

Data Preparation

Sebelum analisis, data dipersiapkan dengan memastikan kualitasnya, yaitu menghapus duplikasi, melengkapi data kosong, dan menyaring responden yang belum memanfaatkan teknologi digital. Selanjutnya, dipilih atribut yang relevan, seperti jenis usaha, penggunaan teknologi, frekuensi pembaruan, analisis digital, dampak penjualan, dan tingkat digitalisasi sebagai fokus utama.

Tabel 2. *Data Cleaning*

No	Jenis_usaha	Penguna an_Tekn ologi	Kepemil ikan_ols hop	Penggu naan_m edia	Websi te_usa ha	Freku ensi_upd ate	Pencatat an_keua ngan	Pembay aran_dig ital	Analis is_digi tal	Dampa k_penj_ digi	Tingkat _Digitali sasi
1	Kulin er	Ya	Ya	Ya	Tidak	Tidak Pernah	Tidak	Ya	Tidak	Sangat Tinggi	Sedang
2	Jasa	Ya	Ya	Ya	Tidak	Jarang	Ya	Tidak	Tidak	Tinggi	Sedang
3	Kulin er	Ya	Ya	Ya	Ya	Setiap Hari	Tidak	Ya	Tidak	Sangat rendah	Sedang
4	Jasa	Tidak	Tidak	Tidak	Tidak	Tidak Pernah	Tidak	Ya	Tidak	Sangat Rendah	Rendah

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v20i2.7416
VOL. 20. NO. 2 FEBRUARI 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

No	Jenis_usaha	Penguna an_Tekn ologi	Kepemil ikan_ols hop	Penggu naan_m edia	Websi te_usa ha	Freku nsi_upd ate	Pencatat an_keua ngan	Pembay aran_dig ital	Analis is_digi tal	Dampa k_penj_ digi	Tingkat _Digitali sasi
5	Kulin er	Ya	Tidak	Ya	Tidak	Bebera pa kali Seming gu	Tidak	Tidak	Tidak	Sangat Rendah	Rendah
...	...	...	...	...	...	...	...	...	...	...	...
1 0 0	Kulin er	Ya	Tidak	Ya	Tidak	Bebera pa kali seming gu	Ya	Ya	Ya	Sangat Tinggi	Sedang

Sumber: (Hasil Penelitian, 2025)

Selanjutnya, data diubah ke format yang sesuai, misalnya skala 1–5 dikonversi menjadi kualitatif: sangat rendah, rendah, sedang, tinggi, dan sangat tinggi.

Tabel 3. Distribusi Kategori Tingkat Digitalisasi

Kategori Tingkat Digitalissi	Jumlah Responden	Persentase
Tinggi	12	12%
Sedang	29	29%
Rendah	59	59%

Sumber: (Hasil Penelitian, 2025)

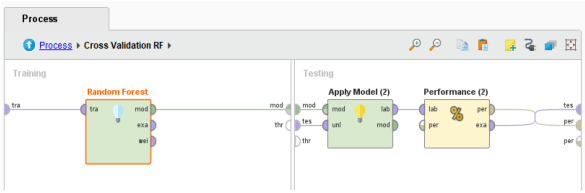
Modeling

Pada tahap ini, model klasifikasi dibangun dengan algoritma C4.5 dan Random Forest menggunakan RapidMiner Studio 9.9. Data kuesioner yang telah diproses diimpor dengan operator *Retrieve*, lalu dibagi menjadi data latih dan uji dengan rasio 70:30 menggunakan *Split Data*.

Operator *Multiply* menduplikasi data agar kedua algoritma diuji pada data yang sama. Evaluasi dilakukan dengan *Cross Validation* untuk menilai *accurcay*, *precision*, *recall*, dan *F1-Score*. Hasil evaluasi inilah yang menjadi dasar membandingkan performa kedua algoritma dalam klasifikasi data UMKM.

Algoritma C4.5

Berikut merupakan proses *Cross Validation* pada algoritma C4.5 di RapidMiner.



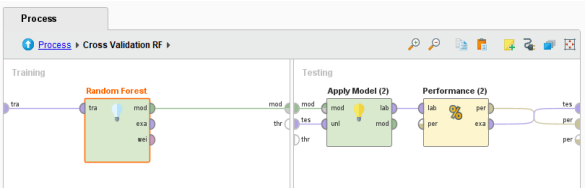
Sumber: (Hasil Penelitian, 2025)
Gambar 2. Proses *Cross Validation*

Algoritma C4.5 menunjukkan bahwa kemampuan analisis digital adalah faktor utama penentu tingkat digitalisasi UMKM di Kabupaten Tegal, didukung oleh kepemilikan toko online dan penggunaan media sosial.

Sumber: (Hasil Penelitian, 2025)
Gambar 3. Pohon Keputusan Algoritma C4.5

Algoritma Random Forest

Sama seperti *Cross Validation* pada C4.5, model dilatih dengan data latih dan diuji pada data uji. Evaluasi *accurcay* dan performa dilakukan dengan operator *Performance*.



Sumber: (Hasil Penelitian, 2025)
Gambar 4. Proses *Cross Validation* RF

Gambar 4 berikut, algoritma Random Forest mengonfirmasi bahwa kemampuan analisis digital tetap menjadi faktor utama dalam menentukan tingkat digitalisasi UMKM di Kabupaten Tegal. Hasil mayoritas pohon dalam ensemble menunjukkan pola yang sama dengan C4.5, yaitu dukungan dari kepemilikan toko online dan penggunaan media sosial turut meningkatkan kategori digitalisasi. Hal ini memperkuat kesimpulan bahwa semakin lengkap penerapan teknologi, semakin tinggi pula tingkat digitalisasi UMKM, sedangkan keterbatasan faktor-faktor tersebut menyebabkan digitalisasi cenderung rendah.

Evaluation

Evaluasi performa model dilakukan dengan metrik *accuray*, *precision*, *recall*, dan *F1-Score* untuk

menilai kemampuan algoritma dalam mengklasifikasikan tingkat digitalisasi UMKM di Kabupaten Tegal. Berikut hasil evaluasi C4.5 dalam bentuk *confusion matrix* dan matrix performa di RapidMiner.
accuracy: 90.00% +/- 11.76% (micro average: 89.86%)

Tabel.4 *Configuration Matrix* Algoritma C4.5

	true Sedang	true Rendah	true Tinggi	class precission
pred. Sedang	18	3	2	78.26%
pred. Rendah	1	38	0	97.44%
pred. Tinggi	1	0	6	85.71%
class recall	90.00%	92.68%	75.00%	

Sumber: (Hasil Penelitian, 2025)

Tabel berikut adalah hasil evaluasi algoritma Random Forest menggunakan RapidMiner. Accuracy: 91.43% +/- 9.99% (micro average: 91.30%)

Tabel 5. Configuration Matrix Algoritma RF

	true Sedang	true Rendah	true Tinggi	class precission
pred. Sedang	17	3	0	85.00%
pred. Rendah	1	38	0	97.44%
pred. Tinggi	2	0	8	80.00%
class recall	85.00%	92.68%	100.00%	

Sumber: (Hasil Penelitian, 2025)

Perbandingan *F1-Score* antara algoritma C4.5 dan Random Forest ditampilkan berdasarkan kategori klasifikasi tingkat digitalisasi, yakni Rendah, Sedang, dan Tinggi. Nilai *F1-Score* ini menunjukkan sejauh mana kedua model mampu menyeimbangkan *precision* dan *recall* dalam proses pengklasifikasian data.

Tabel 6. Nilai *F1-Score*

Classified	*F1-Score C4.5*	*F1-Score Random Forest*
Rendah	95.00%	95.00%
Sedang	83.72%	85.00%
Tinggi	80.00%	88.89%

Sumber: (Hasil Penelitian, 2025)

Berdasarkan Tabel 6, kedua algoritma memiliki keunggulan masing-masing dalam mengklasifikasikan tingkat digitalisasi UMKM ke kategori Rendah, Sedang, dan Tinggi. Pada kategori Rendah, performa keduanya setara, dengan *F1-Score* sebesar 95,00% untuk C4.5 maupun Random Forest. Sementara itu, untuk kategori Sedang, Random Forest unggul tipis dengan 85,00% dibanding C4.5 sebesar 83,72%. Perbedaan mencolok terlihat pada kategori Tinggi, di mana Random Forest mencapai 88,89%, lebih tinggi dari C4.5 yang hanya 80,00%. Secara keseluruhan, Random Forest memberikan hasil klasifikasi lebih baik, terutama pada kategori digitalisasi tinggi, berkat pendekatan *ensemble learning* yang membuat model lebih stabil dan akurat dibanding pohon keputusan tunggal.

Deployment

5. Hasil klasifikasi C4.5 dan Random Forest disajikan dalam laporan dan visualisasi guna mengidentifikasi UMKM yang membutuhkan intervensi digital, penelitian ini menghadirkan analisis komparatif antara algoritma *C4.5 dan Random Forest* berbasis data primer UMKM di Kabupaten Tegal serta menerapkan CRISP-DM secara menyeluruh. Penelitian ini dapat melengkapi kekurangan studi sebelumnya sekaligus menjadi rujukan empiris bagi pengembangan kebijakan dan penelitian lanjutan di bidang digitalisasi UMKM khusunya Kota Tegal.

KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma C4.5 dan Random Forest sama-sama mampu mengklasifikasikan tingkat digitalisasi UMKM di Kabupaten Tegal dengan baik, meskipun Random Forest memberikan hasil yang lebih unggul. Algoritma C4.5 memiliki *accuracy* 90% dengan rata-rata *F1-Score* 86,24%, sedangkan Random Forest mencapai *accuracy* 91,43% dan rata-rata *F1-Score* 89,63%. Random Forest juga lebih stabil dalam mendeteksi tingkat digitalisasi yang tinggi. Penelitian ini juga menunjukkan bahwa mayoritas pelaku UMKM di Kabupaten Tegal masih berada pada tingkat digitalisasi rendah dengan distribusi kategori sebesar 59%, yang terlihat dari nilai *F1-Score* tertinggi di kategori ini pada kedua algoritma, yaitu 95%. Hasil dari pohon keputusan juga menunjukkan bahwa kemampuan analisis digital dan kepemilikan website masih rendah, sehingga tingkat digitalisasi UMKM belum optimal. Kondisi ini mungkin disebabkan oleh terbatasnya literasi digital, minimnya pelatihan dan pendampingan, kurangnya SDM terampil, serta infrastruktur dan akses internet yang belum merata. Akibatnya, kesadaran pelaku usaha untuk memanfaatkan teknologi dalam bisnis juga masih perlu ditingkatkan. Secara keseluruhan, Random Forest direkomendasikan sebagai algoritma yang lebih sesuai untuk pemetaan digitalisasi UMKM di Kabupaten Tegal. Temuan ini diharapkan dapat

menjadi acuan dalam perumusan program pemberdayaan digitalisasi UMKM yang lebih terarah dan berkelanjutan. Berdasarkan temuan penelitian ini, terdapat beberapa saran untuk pengembangan ke depan. Penelitian selanjutnya disarankan melibatkan lebih banyak responden dari berbagai jenis UMKM agar data lebih representatif dan hasil analisis lebih akurat. Selain itu, algoritma pembanding sebaiknya diperluas, tidak hanya C4.5 dan Random Forest, tetapi juga algoritma lain seperti SVM, Naïve Bayes, atau XGBoost untuk melihat kelebihan masing-masing. Penggunaan teknik seleksi fitur yang lebih beragam, seperti *Recursive Feature Elimination* (RFE) atau *Principal Component Analysis* (PCA), juga dapat membantu memilih atribut paling relevan dan membuat model lebih efisien. Ke depan, hasil klasifikasi diharapkan dapat diintegrasikan menjadi rekomendasi praktis bagi pemerintah daerah atau program pendampingan digitalisasi UMKM, sehingga manfaatnya dapat dirasakan langsung oleh pelaku usaha.

REFERENSI

Abdullah, A. (2025). *Prediksi Banjir Di Kota Pontianak Menggunakan Metode*. *8*(1), 40–50.

Aria, R. R. (2025). *Implementasi Algoritma K-Means untuk Pengelompokan Data Imunisasi Balita dengan Metode CRISP-DM*. *9*(1), 189–197.

Bhardwaj, R. B., & Chaurasia, S. R. (2022). Use of ANN, C4.5 and Random Forest Algorithm in the Evaluation of Seismic Soil Liquefaction. *Journal of Soft Computing in Civil Engineering*, *6*(2), 92–106. https://doi.org/10.22115/SCCE.2022.314762.1380

Birgithri, A., Syafira, T., & Louise, N. (2024). Analisis Strategi Pemasaran UMKM untuk Meningkatkan Pertumbuhan Bisnis di Era Digital. *Technomedia Journal*, *9*(1), 117–129. https://doi.org/10.33050/tmj.v9i1.2268

Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2022). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. *Jurnal Teknologi Dan Informasi*, *12*(1), 64–77. https://doi.org/10.34010/jati.v12i1.6674

Fajar, A., Jeffersen, S., Fadilla, R., Zaini, A. R., Sucipto, A., & Lubis, B. O. (2025). *PREDIKSI KELAYAKAN SISWA SMA NEGERI JAKARTA SELATAN*. *9*(2), 3446–3455.

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Ismanto, E., & Novalia, M. (2021). Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas. *Techno.Com*, *20*(3), 400–410. https://doi.org/10.33633/tc.v20i3.4576

Khasanah, N., Komarudin, R., Afni, N., Maulana, Y. I., & Salim, A. (2021). Skin Cancer Classification Using Random Forest Algorithm. *Sisfotenika*, *11*(2), 137. https://doi.org/10.30700/jst.v11i2.1122

Kurniawan, D., & Yasir, M. (2022). Optimization Sentimen Analysis using CRISP-DM and Naive Bayes Methods Implemented on Social Media. *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, *6*(2), 74. https://doi.org/10.22373/cj.v6i2.12793

Melfia, M. A., Ramadhini, K. A., Maulana, M. A., Dasiva, A. L., Briantoro, A. A., & Lubis, B. O. (2023). Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes. *Jurnal SAINTEKOM*, *13*(1), 42–54. https://doi.org/10.33020/saintekom.v13i1.352

Muhammad Nur Ihsan Muhlashin, & Stefanie, A. (2023). Klasifikasi Penyakit Mata Berdasarkan Citra Fundus Menggunakan YOLO V8. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(2), 1363–1368. https://doi.org/10.36040/jati.v7i2.6927

Rais, A. N., Putra, J. L., Informatika, P. S., Bina, U., Informatika, S., Pusat, K. J., Studi, P., Informasi, S., Kampus, A., Tegal, K., Pusat, K. J., Informatika, P. S., Informasi, F. T., Mandiri, U. N., Melayu, C., & Timur, K. J. (2025). *OPTIMASI PREDIKSI RISIKO KREDIT DENGAN PREPROCESSING DAN*. *9*(1), 59–65.

Rininda, G., Hartami Santi, I., & Kirom, S. (2024). Penerapan Svm Dalam Analisis Sentimen Pada Edlink Menggunakan Pengujian Confusion Matrix. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(5), 3335–3342. https://doi.org/10.36040/jati.v7i5.7420

Salsabila, F., Fitrianti, I., Umaidah, Y., & HeryanA, N. (2021). *P Enerapan M Etode C Ustomer R Elationship M Anagement P Ada*. *26*, 38–46.

Yudiana, Y., Yulia Agustina, A., & Nur Khofifah, dan. (2023). Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan. *Indonesian Journal of Islamic Economics and Business*, *8*(1), 01–20.