

Klasifikasi Jenis Usaha Mikro Kecil Menengah (UMKM) Berdasarkan Skala Usaha Menggunakan Algoritma C4.5

Muhammad fadilah¹, Lili Tanti²

^aFakultas Ilmu Komputer, Universitas Potensi Utama Medan, Indonesia
Email : mhammadfadillah20@gmail.com¹, Email : lilitanti82@gmail.com²

ABSTRAK

Usaha Mikro, Kecil, dan Menengah (UMKM) memiliki peran penting dalam mendukung pertumbuhan ekonomi baik di tingkat lokal maupun nasional. Namun, di Desa Sambirejo Timur, Kecamatan Percut Sei Tuan, proses klasifikasi UMKM berdasarkan skala usaha masih dilakukan secara manual tanpa sistem yang terstruktur, sehingga menyulitkan pengambilan keputusan dan pemberian bantuan secara tepat. Penelitian ini bertujuan untuk membangun model klasifikasi skala usaha UMKM menggunakan algoritma decision tree C4.5. Data yang digunakan sebanyak 1.000 data UMKM dengan atribut utama berupa omzet tahunan, jumlah karyawan, jenis usaha, dan modal awal. Data tersebut melalui tahapan pre-processing seperti penanganan missing value, outlier, normalisasi, transformasi data, serta seleksi fitur menggunakan PCA. Dataset kemudian dibagi dengan rasio 80:20 untuk data latih dan data uji. Model dibangun menggunakan algoritma C4.5 dan diimplementasikan menggunakan Python. Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model mampu mengklasifikasikan jenis UMKM dengan akurasi tinggi sebesar 99,50%. Meskipun demikian, ketidakseimbangan jumlah data antar kelas menyebabkan penurunan performa pada kelas minoritas. Sistem yang dihasilkan dapat membantu pemerintah desa dalam pengelolaan data UMKM secara lebih efektif dan akurat. Penelitian ini memberikan kontribusi terhadap digitalisasi pengelolaan UMKM dan mendorong pengambilan keputusan berbasis data.

Kata Kunci : UMKM, klasifikasi, skala usaha, algoritma C4.5, data mining, decision tree.

Classification of Micro, Small, and Medium Enterprises (MSMEs) Based on Business Scale Using the C4.5 Algorithm

ABSTRACT

Micro, Small, and Medium Enterprises (MSMEs) play a crucial role in supporting economic growth at both local and national levels. However, in Sambirejo Timur Village, Percut Sei Tuan Subdistrict, the classification of MSMEs based on business scale is still performed manually and lacks a structured system, making it difficult to make accurate decisions and allocate appropriate support. This study aims to develop a business scale classification model for MSMEs using the C4.5 decision tree algorithm. The dataset consists of 1,000 MSME records with key attributes including annual turnover, number of employees, business type, and initial capital. The data underwent preprocessing stages such as handling missing values, outlier detection, normalization, data transformation, and feature selection using PCA. The dataset was split using an 80:20 ratio for training and testing. The model was built using the C4.5 algorithm and implemented in Python. Evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results show that the model can classify MSME types with a high accuracy of 99.50%. However, class imbalance caused performance degradation in minority classes. The developed system helps the village administration manage MSME data more effectively and accurately. This research contributes to the digitalization of MSME management and promotes data-driven decision-making.

Keywords : MSME, classification, business scale, C4.5 algorithm, data mining, decision tree.

Disubmit: 11 September 2025

Info Artikel :
Direview: 10 Oktober 2025

Diterima : 3 November 2025

Copyright © 2025 – senadimu. All rights reserved.

DOI: <https://www.doi.org/10.22303>

1. PENDAHULUAN

Desa Sambirejo Timur merupakan salah satu wilayah administratif yang terletak di Kecamatan Percut Sei Tuan, Kabupaten Deli Serdang, Provinsi Sumatera Utara. Secara geografis, desa ini berbatasan dengan Desa Tembung di sebelah barat, Desa Bandar Klippa di utara, Desa Amplas di selatan, dan Desa Sena di sebelah timur. Sungai Denai yang melintasi wilayah ini menjadi batas alami yang memisahkan kawasan dataran tinggi Deli Serdang dan Karo di selatan dengan wilayah pesisir timur Sumatera di utara, serta menjadi pemisah administratif antara Kota Medan dan Kabupaten Deli Serdang. Sebagai lembaga pemerintahan di tingkat desa, Kantor Desa Sambirejo Timur memiliki tanggung jawab dalam berbagai aspek pelayanan publik, termasuk pembinaan dan pengelolaan data pelaku Usaha Mikro, Kecil, dan Menengah (UMKM). UMKM merupakan sektor ekonomi produktif yang dijalankan oleh individu atau badan usaha kecil dan mencerminkan aktivitas ekonomi masyarakat kelas menengah ke bawah. Berdasarkan Undang-Undang Nomor 20 Tahun 2008 Pasal 3 dan 5, UMKM memiliki peran strategis dalam membangun perekonomian nasional melalui prinsip demokrasi ekonomi yang adil dan berkelanjutan (Purwanto & Wulandari, 2023). Namun, permasalahan yang dihadapi saat ini adalah belum adanya sistem klasifikasi UMKM yang berbasis pada skala usaha secara terstruktur. Selama ini, proses klasifikasi tidak mempertimbangkan atribut penting seperti jumlah aset, omzet tahunan, atau jumlah tenaga kerja, sehingga menyulitkan proses perencanaan kebijakan, alokasi bantuan, dan akses pembiayaan yang tepat. Oleh karena itu, diperlukan pendekatan berbasis teknologi untuk mendukung proses klasifikasi UMKM secara lebih objektif dan sistematis.

Salah satu pendekatan yang relevan dalam menyelesaikan permasalahan klasifikasi adalah data mining. Data mining merupakan proses eksplorasi dan penggalian pola atau pengetahuan baru dari kumpulan data dalam jumlah besar yang dapat digunakan untuk mendukung proses pengambilan keputusan (Sulistio et al., 2023). Dalam konteks klasifikasi UMKM berdasarkan skala usaha, salah satu algoritma data mining yang potensial digunakan adalah decision tree C4.5, yang mampu menangani data atribut dengan struktur pohon keputusan yang mudah dipahami. Efektivitas algoritma C4.5 dalam menyelesaikan permasalahan klasifikasi telah dibuktikan oleh berbagai penelitian terdahulu. Penelitian oleh Amalia et al. (2024) menunjukkan bahwa algoritma C4.5 mampu mengidentifikasi faktor-faktor yang memengaruhi kelayakan pemberian Kredit Usaha Rakyat (KUR) pada UMKM di PT Pegadaian dengan tingkat akurasi mencapai 95,56%. Model tersebut dapat digunakan sebagai sistem pendukung keputusan untuk mengurangi risiko gagal bayar.

Penelitian oleh Suseno et al. (2023) membandingkan algoritma C4.5 dengan K-Nearest Neighbor (K-NN) dalam konteks klasifikasi pemberian bantuan UMKM. Hasilnya menunjukkan bahwa meskipun algoritma K-NN menghasilkan akurasi yang lebih tinggi (94,94%) dibandingkan C4.5 (86,08%), keduanya tetap mampu memberikan hasil klasifikasi yang baik, tergantung pada karakteristik data yang digunakan. Dalam konteks preferensi konsumen, penelitian oleh Fanani et al. (2024) menerapkan algoritma C4.5 dalam kegiatan sosialisasi produk UMKM di Desa Karangdowo, Pekalongan. Hasilnya menunjukkan bahwa pendekatan berbasis data mining ini membantu pelaku UMKM memahami preferensi pasar serta meningkatkan daya saing produk.

Penelitian lain oleh Wajir (2021) menggunakan algoritma C4.5 untuk menganalisis kepuasan konsumen terhadap UMKM MiniMall. Indikator seperti ketersediaan produk, kualitas, dan ketepatan pengiriman menjadi faktor utama dalam pembentukan pohon keputusan. Hasilnya

menunjukkan bahwa C4.5 mampu memberikan wawasan strategis yang bermanfaat bagi pelaku usaha dalam meningkatkan layanan.

Sementara itu, Pratiwi et al. (2022) dalam penelitiannya mengenai klasifikasi kelayakan penerima Bantuan Sosial Tunai (BST) di Desa Keramas, Gianyar, menunjukkan bahwa algoritma C4.5 menghasilkan akurasi sebesar 97,83%, lebih tinggi dibandingkan K-NN (92,29%) dan Naïve Bayes (91,81%).

Berdasarkan berbagai penelitian tersebut, dapat disimpulkan bahwa algoritma C4.5 memiliki kapabilitas yang tinggi dalam mengolah dan mengklasifikasikan data berbasis atribut. Oleh karena itu, penelitian ini bertujuan untuk menerapkan algoritma C4.5 dalam membangun sistem klasifikasi jenis UMKM berdasarkan skala usaha di Desa Sambirejo Timur. Diharapkan sistem ini dapat mendukung proses pengambilan keputusan yang cepat, akurat, dan efisien serta membantu pemerintah desa dalam menyusun strategi pemberdayaan UMKM yang tepat sasaran.

2. METODE

2.1 Sumber dan Cara Pengumpulan Data

Dalam penelitian ini, peneliti menggunakan dua jenis data, yaitu data primer dan data sekunder, guna mendukung proses klasifikasi jenis usaha mikro, kecil, dan menengah (UMKM) berdasarkan skala usaha menggunakan algoritma C4.5.

a. Data Primer

Data primer dikumpulkan secara langsung dari subjek penelitian melalui observasi lapangan dan teknik wawancara kepada pelaku UMKM yang berada di Desa Sambirejo Timur, Kecamatan Percut Sei Tuan, Kabupaten Deli Serdang. Pengumpulan data ini bertujuan untuk memperoleh informasi aktual yang sesuai dengan kondisi usaha di lapangan.

- Wawancara. Dilakukan secara langsung kepada pelaku UMKM maupun pihak yang relevan, seperti aparatur desa atau pihak terkait lainnya yang memiliki informasi mengenai klasifikasi dan karakteristik usaha di wilayah tersebut. Prosedur tanya jawab ini membantu peneliti memperoleh informasi mengenai omset, jumlah karyawan, jenis usaha, dan modal awal—yang merupakan fitur penting dalam proses klasifikasi.
- Dokumentasi. Dilakukan dengan mengumpulkan data tertulis atau visual, seperti formulir survei UMKM, catatan administrasi desa, laporan kegiatan UMKM, serta foto-foto kondisi usaha. Dokumen ini digunakan untuk memverifikasi kebenaran data yang diperoleh melalui wawancara dan observasi langsung.

b. Data Sekunder

Data sekunder diperoleh dari berbagai sumber yang relevan untuk memperkuat hasil analisis, seperti jurnal ilmiah, artikel, laporan pemerintah (misalnya dari Kementerian Koperasi dan UKM), serta dokumen kebijakan klasifikasi UMKM berdasarkan omset, jumlah tenaga kerja, jenis usaha dan modal awal. Data ini berguna dalam proses validasi dan penyusunan model klasifikasi.

2.2 Tahapan Penelitian



Gambar 1: Tahapan Proses Penelitian.

a. Identifikasi Masalah

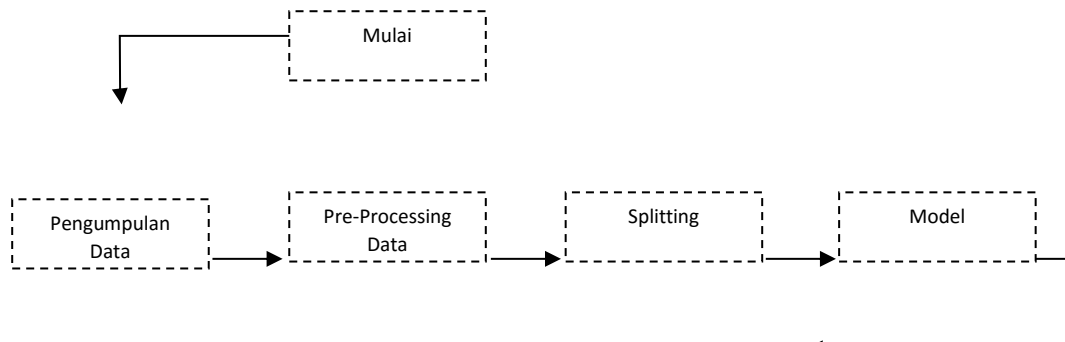
Kantor Desa Sambirejo, Kecamatan Percut Sei Tuan belum memiliki sistem yang mampu mengklasifikasikan skala usaha UMKM secara otomatis berdasarkan data seperti omset, jumlah karyawan, dan modal awal, sehingga dibutuhkan aplikasi berbasis algoritma C4.5 untuk membantu proses klasifikasi secara lebih akurat dan efisien.

b. Analisis Kebutuhan

Pengumpulan data dilakukan melalui wawancara dan studi dokumentasi untuk mendapatkan informasi terkait karakteristik UMKM, seperti omset, jumlah karyawan, jenis usaha, dan modal awal. Hasilnya menunjukkan bahwa Kantor Desa Sambirejo belum memiliki sistem klasifikasi skala usaha UMKM yang otomatis, sehingga proses klasifikasi masih manual dan kurang akurat. Oleh karena itu, diperlukan aplikasi berbasis algoritma C4.5 untuk membantu klasifikasi skala usaha UMKM secara lebih akurat, efisien, dan mendukung pengambilan keputusan yang tepat.

c. Perancangan Sistem

Perancangan sistem dilakukan pada tahap ini dengan menggunakan beberapa langkah tahapan yang dilakukan pada penelitian ini sebagai berikut,





Gambar 2: Perancangan Sistem.

d. Implementasi Sistem

Pada tahap ini, sistem klasifikasi skala usaha UMKM diimplementasikan menggunakan bahasa pemrograman Python. Aplikasi dikembangkan dalam bentuk non-web (berbasis desktop atau terminal), dengan mengintegrasikan algoritma C4.5 untuk mengolah data dan menentukan klasifikasi usaha ke dalam kategori mikro, kecil, atau menengah. Fitur utama dalam sistem mencakup fungsi CRUD (Create, Read, Update, Delete) untuk pengelolaan data UMKM, seperti data omset, jumlah tenaga kerja, jenis usaha, dan modal awal. Selain itu, sistem dilengkapi dengan kemampuan untuk melakukan pelatihan model, prediksi data baru, serta evaluasi performa klasifikasi guna memastikan keakuratan hasil.

Dalam proses implementasinya, sistem dirancang dengan mempertimbangkan spesifikasi perangkat keras dan perangkat lunak yang diperlukan agar dapat berfungsi secara optimal dan mendukung operasional di lingkungan Kantor Desa. Klasifikasi

e. Klasifikasi

Tahap Tahap Klasifikasi ini dimaksudkan untuk melakukan perhitungan secara manual ataupun memanfaatkan algoritma C4.5 di Bahasa Pemrograman Python. Adapun pada gambar 2 merupakan tahapan implementasi algoritma yang digunakan guna mengklasifikasi skala usaha UMKM secara otomatis. Berikut tahapan-tahapan yang dilakukan pada proses penelitian ini diantaranya sebagai berikut,

- Tahapan Pengumpulan Data
Tahapan ini merupakan langkah awal dalam proses sistem klasifikasi. Data dikumpulkan melalui observasi, wawancara, dan studi dokumentasi dari pelaku UMKM di Desa Sambirejo Timur. Data yang dikumpulkan mencakup fitur-fitur penting seperti omset tahunan, jumlah karyawan, jenis usaha, dan modal awal, yang akan digunakan sebagai input dalam proses klasifikasi.
- Tahapan Pre-Processing Data
Setelah data terkumpul, dilakukan proses pembersihan dan persiapan data (pre-processing). Ini mencakup diantaranya menghapus data kosong atau duplikat, menormalisasi nilai numerik (jika diperlukan), dan mengubah data kategorikal menjadi numerik agar dapat diproses oleh algoritma C4.5. Tahapan ini bertujuan untuk memastikan kualitas dan konsistensi data sebelum dimodelkan.
- Tahapan Splitting Data
Data yang telah diproses kemudian dibagi menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data). Teknik pemisahan data ini dilakukan agar sistem dapat dilatih dan kemudian diuji menggunakan data yang berbeda, guna mengukur performa algoritma dalam melakukan klasifikasi.
- Tahapan Model Data
Pada tahap ini, data latih digunakan untuk membangun model klasifikasi menggunakan algoritma C4.5. Model ini akan membentuk pohon keputusan (decision tree) yang merepresentasikan aturan klasifikasi berdasarkan atribut-atribut yang dimiliki oleh data UMKM.
- Tahapan Evaluasi Data

Model yang telah dibentuk kemudian dievaluasi menggunakan data uji. Evaluasi dilakukan dengan mengukur akurasi dan performa model melalui metrik seperti confusion matrix, precision, recall, dan F1-score. Tahapan ini penting untuk mengetahui seberapa baik model dapat mengklasifikasikan UMKM ke dalam skala mikro, kecil, atau menengah secara tepat. Menurut Ade Saputra dan Firdaus (2022), dalam pembuatan pohon keputusan menggunakan algoritma C4.5, terdapat beberapa tahapan penting yang harus dilakukan. Pertama, menyiapkan data pelatihan yang berasal dari data historis yang telah terjadi sebelumnya dan telah dikelompokkan ke dalam masing-masing kelas. Langkah selanjutnya adalah menentukan akar pohon keputusan. Pemilihan atribut untuk akar dilakukan dengan menghitung nilai gain dari setiap atribut yang tersedia. Atribut dengan nilai gain tertinggi akan dipilih sebagai akar pohon. Sebelum menghitung nilai gain, perlu terlebih dahulu menghitung nilai entropi, yang digunakan untuk mengukur tingkat ketidakpastian atau impuritas dalam kumpulan data. Berikut rumus untuk menghitung nilai entropi terlampir,

$$\text{Entropy}(S) = \sum_{i=1}^n - p_i \cdot \log_2 p_i \quad (1)$$

Keterangan,

- S = Himpunan
- n = Jumlah kelas yang ada dalam dataset. (Jika kelas terdapat mikro, kecil dan menengah, maka $n = 3$)
- p_i = Proporsi (probabilitas) dari setiap kelas- i terhadap total data dalam himpunan S.
- $\log_2 (p_i)$ = Logaritma basis 2 dari nilai proporsi.

Kemudian setelah didapatkan nilai entropi untuk setiap atribut, hitung nilai gain dengan menggunakan rumus,

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^n \left(\frac{|S_i|}{|S|} \cdot \text{Entropy}(S_i) \right) \quad (2)$$

Keterangan,

- $\text{Gain}(S, A)$ = Nilai informasi yang diperoleh dari atribut A untuk memisahkan himpunan data S
- $\text{Entropy}(S)$ = Nilai entropi dari himpunan data S secara keseluruhan
- n = Jumlah kelas yang ada dalam dataset
- S_i = Subset data dari S untuk nilai atribut A ke- i
- $|S_i| / |S|$ = Proporsi data pada cabang ke- i dibandingkan seluruh data
- $\text{Entropy}(S_i)$ = Entropi dari subset S_i yaitu tingkat ketidakpastian pada cabang tersebut.

– Tahapan Deployment

Setelah model dikembangkan dan dievaluasi dengan baik, sistem kemudian diimplementasikan atau dideploy agar dapat digunakan oleh pihak Kantor Desa. Sistem ini berbasis bahasa pemrograman Python, berjalan pada lingkungan desktop, dan memungkinkan pengguna untuk memasukkan data baru, melihat hasil klasifikasi, serta melakukan pengelolaan data UMKM.

3. HASIL DAN PEMBAHASAN

Berikut hasil dan pembahasan dari setiap langkah-langkah pada penelitian ini,

3.1. Tahapan Pengumpulan Data

Tahap awal dalam proses pengembangan sistem klasifikasi skala usaha UMKM dimulai dengan pengumpulan data yang dilakukan secara langsung di Desa Sambirejo Timur, Kecamatan Percut Sei Tuan, Kabupaten Deli Serdang sebanyak 1000 records. Metode yang digunakan dalam pengumpulan data meliputi observasi lapangan, wawancara dengan pelaku UMKM dan aparat desa, serta studi dokumentasi terhadap data administratif yang tersedia. Data yang berhasil dikumpulkan terdiri atas beberapa fitur utama yang relevan dengan penentuan skala usaha, yaitu: Nama Usaha, Alamat Usaha, Nama Pemilik, Alamat Pemilik, Jumlah Karyawan, Modal, Omset, dan Kategori. Fitur-fitur ini dipilih berdasarkan kriteria umum pengelompokan skala UMKM sesuai dengan peraturan pemerintah dan akan digunakan sebagai atribut input dalam proses klasifikasi menggunakan algoritma C4.5. Berikut Gambar 3

merupakan dataset yang telah berhasil dikumpulkan dalam format CSV sebagai data sekunder yang akan digunakan,

Unnamed: 0		Nama Usaha	Alamat Usaha	Nama Pemilik	Alamat Pemilik	Jumlah Karyawan	Modal	Omset	Kategori
0	NaN	Warung Mie Sep Kartil	Jl. Sederhana Dsn IX Serosa	Darianti	Jl.Sederhana Gg.musayawarah	1.0	6000000	8000000	UMKM DESA SAMBIREJO TIMUR
1	No	Gedang serba guna	Jl. Sederhana Dsn IX Serosa	Suriatmen	Jl.swadaya Dsn IX Serosa	6.0	700000	150000000	UMKM DESA SAMBIREJO TIMUR
2	1	Bengkel Motor Yaris	Jl. Sederhana Dsn IX Serosa	Riski Hariansyah	Jl.swadaya Dsn IX Serosa	7.0	130000000	115000000	UMKM DESA SAMBIREJO TIMUR
3	2	bengkel motor gokil	Jl. Sederhana Dsn IX Serosa	Reza	Jl.swadaya Dsn IX Serosa	6.0	9000000	9000000	UMKM DESA SAMBIREJO TIMUR
4	3	cia Ponsel	Jl. Sederhana Dsn IX Serosa	Gela Agustina	Jl.Bandar Setia Gg.buntu	3.0	2600000	42000000	UMKM DESA SAMBIREJO TIMUR

Gambar 3: Sampel Dataset Awal berupa file CSV.

3.2. Tahapan Pre-processing Data

Proses pre-processing merupakan bagian penting dari tahap pembersihan data yang bersifat krusial sebelum data digunakan dalam pemodelan klasifikasi menggunakan algoritma C4.5. Tahap ini bertujuan untuk memastikan kualitas dan integritas data agar model yang dibangun mampu menghasilkan output yang akurat dan andal. Beberapa tahapan dalam proses pre-processing atau pembersihan data yang dilakukan dalam penelitian ini antara lain sebagai berikut,

a. Penanganan Missing Data

Missing data merupakan kondisi di mana satu atau lebih nilai dalam dataset tidak tersedia. Dalam konteks penelitian ini, missing data ditemukan pada beberapa atribut penting seperti Nama

Usaha, Jumlah Karyawan, Modal (Rp), dan Omzet (Rp). Ketidakhadiran data ini dapat disebabkan oleh kesalahan input, informasi yang tidak lengkap dari responden, atau gangguan teknis selama proses pengumpulan data. Untuk menangani kondisi ini, digunakan beberapa metode, antara lain penghapusan baris data yang kosong jika jumlahnya signifikan dan tidak dapat diisi secara logis, imputasi menggunakan nilai rata-rata (mean) atau nilai tengah (median) jika data yang hilang relatif sedikit, serta interpolasi untuk memperkirakan nilai yang hilang apabila data memiliki pola berurutan.

```
... Unnamed: 0      20
    Nama Usaha      31
    Alamat Usaha     31
    Nama Pemilik     30
    Alamat Pemilik   30
    Jumlah Karyawan  30
    Modal            30
    Omzet            30
    Kategori         30
    dtype: int64
```

a. Cek Missing Data (Before)

```
... Nama Usaha      0
    Alamat Usaha     0
    Nama Pemilik     0
    Alamat Pemilik   0
    Jumlah Karyawan  0
    Modal            0
    Omzet            0
    Kategori         0
    dtype: int64
```

b. Hapus Missing Data (After)

Gambar 4: Proses Penanganan Missing Data

b. Deteksi dan Penanganan Outlier

Outlier adalah data yang memiliki nilai ekstrem dan berbeda jauh dari mayoritas data lainnya. Dalam dataset UMKM, outlier ditemukan pada kolom Modal (Rp) dan Omzet (Rp), seperti usaha dengan omzet sangat tinggi yang tidak wajar untuk skala mikro atau kecil. Dalam penelitian ini, penanganan outlier dilakukan melalui beberapa tahapan berikut,

- Deteksi Statistik. Outlier diidentifikasi menggunakan pendekatan statistik seperti Z-Score dan Interquartile Range (IQR). Gunanya untuk mendeteksi nilai-nilai yang berada diluar rentang normal distribusi data.
- Verifikasi Manual. Setelah data dicurigai sebagai outlier, dilakukan pemeriksaan manual untuk memastikan apakah nilai tersebut merupakan kesalahan input atau memang mencerminkan kondisi yang valid. Hal ini penting agar tidak menghapus data sah yang dapat merepresentasikan variabilitas UMKM secara real.
- Transformasi Data. Jika outlier valid namun mempengaruhi distribusi data secara signifikan, dilakukan transformasi logaritmik untuk mengurangi dampaknya terhadap analisis dan pemodelan.
- Penghapusan atau Koreksi. Outlier yang terbukti merupakan hasil kesalahan input dengan konteks klasifikasi akan diperbaiki atau dihapus dari dataset untuk menjaga kualitas data. Berikut Gambar 5 hasil outlier data yang telah dilakukan pada penelitian ini,

```
... Jumlah Karyawan  19
    Modal            97
    Omzet            102
    dtype: int64
```

Gambar 5: Hasil Perhitungan Outlier.

c. Publikasi Data

Setelah dilakukan proses penanganan missing data dan deteksi outlier selesai, tahap publikasi data memastikan bahwa dataset yang digunakan sudah bersih, akurat, dan siap untuk dianalisis atau dipublikasikan dalam laporan. Beberapa aspek yang perlu diperhatikan dalam publikasi data diantaranya adalah,

- Konsistensi Format. Seluruh data harus disajikan dalam format yang seragam, khususnya pada kolom numerik seperti nilai dalam satuan Rupiah (Rp). Hal ini penting agar tidak terjadi pencampuran antara angka, teks atau symbol yang dapat mengganggu proses analisis. Berikut Gambar 6 merupakan hasil konsistensi format terlampir,

	Jumlah Karyawan	Modal	Omset
0	10	6000000	8000000
1	60	700000	150000000
2	20	130000000	115000000
3	60	9000000	9000000
4	30	2600000	42000000

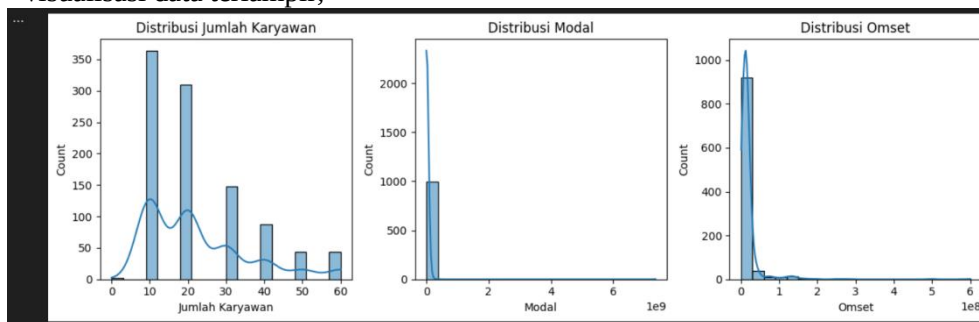
Gambar 6: Hasil Konsistensi Format.

- Anonimisasi Data. Informasi pribadi seperti nama pemilik usaha atau identitas responden perlu disamarkan atau dihapus guna menjaga kerahasiaan data dan mematuhi prinsip etika penelitian. Berikut Gambar 7 merupakan hasil anonimisasi data terlampir,

	Nama Usaha	Alamat Usaha	Nama Pemilik	Alamat Pemilik	Jumlah Karyawan	Modal	Omset	Kategori
0	Warung Mie Sep Kartil	Jl. Sederhana Dsn IX Serosa	Anonim	Disembunyikan	10	6000000	8000000	UMKM DESA SAMBIREJO TIMUR
1	Gedang serba guna	Jl. Sederhana Dsn IX Serosa	Anonim	Disembunyikan	60	700000	150000000	UMKM DESA SAMBIREJO TIMUR
2	Bengkel Motor Yaris	Jl. Sederhana Dsn IX Serosa	Anonim	Disembunyikan	20	130000000	115000000	UMKM DESA SAMBIREJO TIMUR
3	bengkel motor gokil	Jl. Sederhana Dsn IX Serosa	Anonim	Disembunyikan	60	9000000	9000000	UMKM DESA SAMBIREJO TIMUR
4	cia Ponsel	Jl. Sederhana Dsn IX Serosa	Anonim	Disembunyikan	30	2600000	42000000	UMKM DESA SAMBIREJO TIMUR

Gambar 7: Hasil Anonimisasi Data.

- Visualisasi Data. Dataset yang telah dibersihkan disajikan dalam bentuk grafik, tabel, atau ringkasan statistik deskriptif guna memudahkan interpretasi dan memperjelas distribusi serta karakteristik data sebelum dilakukan pemodelan. Berikut Gambar 8 merupakan hasil visualisasi data terlampir,



Gambar 7: Hasil Anonimisasi Data.

d. Normalisasi Data

Normalisasi merupakan proses penting dalam tahap pre-processing data, khususnya untuk menyamakan skala nilai pada fitur numerik yang memiliki rentang berbeda. Dalam penelitian ini, normalisasi dilakukan terhadap atribut numerik seperti modal usaha dan omzet, yang memiliki skala nilai yang cukup besar dan bervariasi. Tujuan dari normalisasi adalah untuk menghindari dominasi fitur tertentu dalam proses klasifikasi, serta mempercepat konvergensi algoritma pembelajaran. Metode yang digunakan adalah Min-Max Scaling, yaitu metode normalisasi yang mengubah nilai asli menjadi nilai dalam rentang 0 hingga 1, tanpa mengubah

pola distribusi relatif antar data. Dengan menerapkan metode ini, setiap nilai dalam kolom numerik akan dikonversi ke skala yang seragam (0–1), sehingga algoritma klasifikasi seperti C4.5 dapat bekerja secara optimal tanpa bias terhadap atribut dengan skala yang lebih besar. Hasil normalisasi data dapat dilihat pada Gambar 8 terlampir,

...	Jumlah Karyawan	Modal	Omset
0	0.166667	0.000755	0.012346
1	1.000000	0.000034	0.249249
2	0.333333	0.017627	0.190858
3	1.000000	0.001163	0.014014
4	0.500000	0.000293	0.069069

Gambar 8: Hasil Normalisasi Data.

e. Transformasi Data.

Transformasi data dilakukan terhadap fitur kategorikal agar dapat diproses oleh algoritma klasifikasi C4.5 yang membutuhkan input numerik. Salah satu metode yang digunakan adalah One Hot Encoding, yaitu teknik representasi data kategorikal dalam bentuk vektor biner. Metode ini mempertahankan nilai unik dari masing-masing kategori tanpa memberikan bobot numerik yang bisa menyesatkan proses klasifikasi. Setelah model selesai digunakan, hasil klasifikasi berbentuk vektor biner ini dapat dikembalikan ke label aslinya melalui proses One Hot Decoding, sehingga dapat ditampilkan dalam bentuk yang mudah dipahami, seperti kategori “Mikro”, “Kecil”, atau “Menengah”. Dengan enentuan kelas target (Kategori UMKM) dilakukan berdasarkan kombinasi dua variabel numerik utama, yaitu: modal awal usaha dan omzet tahunan. Kedua atribut ini sebelumnya telah dinormalisasi menggunakan metode Min-Max Normalization untuk menyesuaikan skala ke rentang [0, 1]. Pendekatan klasifikasi dibuat dalam bentuk fungsi logika sebagai berikut,

- Jika nilai normalisasi Modal < 0.33 dan Omset < 0.33 → diklasifikasikan sebagai UMKM Mikro
- Jika nilai Modal < 0.66 dan Omset < 0.66 → diklasifikasikan sebagai UMKM Kecil
- Selebihnya diklasifikasikan sebagai UMKM Menengah

Fungsi logika ini dirancang agar sesuai dengan skala normalisasi. Nilai target yang dihasilkan menjadi label kelas yang digunakan dalam algoritma C4.5.

...	mat. Gg. Baltun	Alamat Gg. Bahayu	...	Alamat J.enderhane Gg.muyawarah	Alamat J.enderhane Gg.serasa 13	Alamat J.aidomullo	Alamat J.sandaga Dst IX serasa	Alamat J.veteran	Alamat Lorong Jaya	Alamat Lorong Winau	Alamat belvet par 2	Alamat belvet par 4	Kategori UMKM
	False	False	...	False	False	False	False	False	False	False	False	False	Mikro
	False	False	...	False	False	False	False	False	False	False	False	False	Mikro
	False	False	...	False	False	False	False	False	False	False	False	False	Mikro
	False	False	...	False	False	False	False	False	False	False	False	False	Mikro
	False	False	...	False	False	False	False	False	False	False	False	False	Mikro

Gambar 9: Hasil Transformasi Data.

f. Feature Selection.

Dalam proses klasifikasi, reduksi dimensi menjadi langkah penting untuk meningkatkan efisiensi dan performa model, terutama ketika data memiliki banyak fitur. Salah satu teknik yang umum digunakan untuk seleksi fitur adalah Principal Component Analysis (PCA). PCA merupakan metode statistik yang digunakan untuk mengubah kumpulan variabel yang saling berkorelasi menjadi sejumlah variabel baru yang tidak berkorelasi satu sama lain, yang disebut sebagai Principal Components (PCs). Tujuan utama dari PCA adalah untuk mempertahankan sebanyak mungkin informasi (varians) dari data asli dengan jumlah fitur yang lebih sedikit, sehingga mendukung proses klasifikasi menjadi lebih efisien tanpa mengorbankan akurasi. Hasil pca feature selection dapat dilihat pada Gambar 10 terlampir,

```
... (      PC1      PC2      PC3
0 -0.210499 -0.007657 -0.001746
1  0.632920  0.190046 -0.012980
2 -0.035721  0.163553  0.006539
3  0.622016 -0.044595 -0.000258
4  0.125105  0.033465 -0.004539,
array([0.92666178, 0.0557146 , 0.01762362]))
```

Gambar 10: Hasil PCA Feature Selection.

3.3. Tahapan Splitting Data

Dalam penelitian ini, dataset dibagi menjadi dua subset, yaitu data latih (training set) dan data uji (testing set) dengan rasio 80:20. Artinya, sebanyak 80% dari total data digunakan untuk melatih model klasifikasi, sedangkan 20% sisanya digunakan untuk menguji performa model yang telah dibangun. Pembagian ini dilakukan untuk memastikan bahwa model algoritma C4.5 dapat melakukan generalisasi dengan baik terhadap data baru yang belum pernah dilihat sebelumnya. Data latih berfungsi untuk membangun struktur pohon keputusan, sedangkan data uji digunakan untuk mengevaluasi akurasi dan efektivitas model dalam mengklasifikasikan jenis UMKM berdasarkan fitur-fitur seperti modal, omzet, jumlah karyawan, dan jenis usaha. Secara umum, rasio 80:20 merupakan standar umum dalam pembelajaran mesin yang dianggap cukup untuk memberikan pelatihan yang optimal tanpa mengorbankan validitas pengujian. Hasil splitting data dapat dilihat pada Gambar 11 terlampir,

```
... ((798, 3),
(200, 3),
Kategori UMKM
Mikro      794
Kecil      2
Menengah   2
Name: count, dtype: int64,
Kategori UMKM
Mikro      199
Menengah   1
Name: count, dtype: int64)
```

Gambar 11: Hasil Splitting Data.

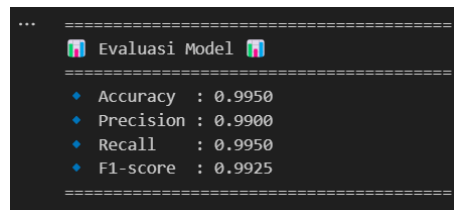
3.4. Tahapan Model Data

Pada tahap ini dilakukan proses pembangunan model klasifikasi menggunakan algoritma Decision Tree C4.5. Dalam pustaka scikit-learn, algoritma ini diimplementasikan menggunakan DecisionTreeClassifier dengan parameter criterion="entropy", yang mengindikasikan bahwa model

menggunakan pengukuran Information Gain sebagai dasar pemilihan atribut pada setiap simpul pohon. Dengan menggunakan entropy dan information gain, model akan membangun pohon keputusan yang optimal untuk memetakan fitur-fitur numerik dan kategorikal menjadi label kelas yang relevan (Mikro, Kecil, Menengah). Tahapan ini penting sebagai inti dari proses klasifikasi dalam penelitian data mining berbasis supervised learning. Selanjutnya model tersebut akan membuat prediksi akurasi pada data testing yang diuji.

3.5. Tahapan Evaluasi Data

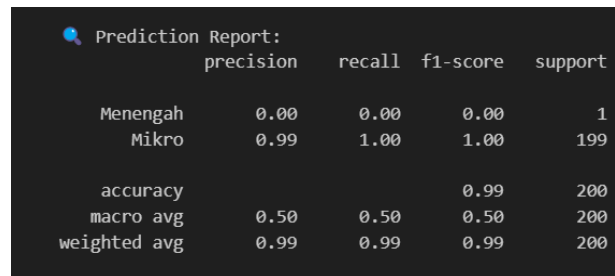
Proses pada evaluasi model ini merupakan hasil akhir dari klasifikasi jenis usaha mikro kecil menengah (UMKM). Hasil evaluasi algoritma C4.5 dapat dilihat pada Gambar 12 serta prediksi report pada Gambar 13 terlampir,



```

=====
Evaluasi Model
=====
• Accuracy : 0.9950
• Precision : 0.9900
• Recall   : 0.9950
• F1-score : 0.9925
=====
  
```

Gambar 12: Hasil Evaluasi Data.



```

Prediction Report:
precision  recall  f1-score  support

Menengah   0.00    0.00    0.00         1
Mikro      0.99    1.00    1.00        199

accuracy   0.99         200
macro avg  0.50    0.50    0.50         200
weighted avg 0.99    0.99    0.99         200
  
```

Gambar 13: Hasil Prediksi Report.

Berdasarkan hasil evaluasi model klasifikasi UMKM menggunakan algoritma Decision Tree (C4.5), diperoleh tingkat akurasi sebesar 99,50%, precision sebesar 99,00%, recall sebesar 99,50%, dan F1-score sebesar 99,25%. Nilai-nilai ini menunjukkan bahwa model secara keseluruhan memiliki kinerja yang sangat tinggi dalam melakukan klasifikasi data UMKM. Namun, jika ditinjau lebih dalam melalui laporan klasifikasi (classification report), ditemukan bahwa model sangat baik dalam mengklasifikasikan kelas "Mikro" (dengan precision dan recall sempurna), tetapi tidak mampu mengklasifikasikan kelas "Menengah" dengan benar (precision dan recall = 0). Hal ini disebabkan oleh ketidakseimbangan jumlah data antar kelas (class imbalance), di mana kelas "Mikro" memiliki 199 data, sedangkan kelas "Menengah" hanya 1 data. Kondisi ini menunjukkan bahwa meskipun model memiliki akurasi tinggi secara umum, model cenderung bias terhadap kelas mayoritas, dan memiliki kinerja buruk terhadap kelas minoritas. Oleh karena itu, dalam pengembangan model lebih lanjut, perlu dilakukan strategi penyeimbangan data seperti oversampling (misalnya SMOTE) atau undersampling, agar model dapat lebih adil dan akurat dalam mengklasifikasikan semua kategori skala usaha UMKM.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi jenis Usaha Mikro, Kecil, dan Menengah (UMKM) berdasarkan skala usaha menggunakan algoritma C4.5. Melalui tahapan pre-processing data seperti penanganan missing value, deteksi outlier, normalisasi, transformasi data dengan one-hot encoding, serta seleksi fitur menggunakan PCA, diperoleh dataset yang bersih dan siap untuk digunakan dalam pemodelan. Pembagian data latih dan data uji dengan rasio 80:20 menghasilkan model klasifikasi yang menunjukkan performa sangat tinggi, dengan akurasi mencapai 99,50%, precision 99,00%, recall 99,50%, dan F1-score 99,25%. Namun, model masih mengalami kendala dalam mengklasifikasikan kelas minoritas akibat ketidakseimbangan data (class imbalance), khususnya pada kelas "Menengah". Implementasi model dalam bentuk aplikasi berbasis Python mampu membantu proses klasifikasi UMKM secara otomatis, cepat, dan akurat. Sistem ini diharapkan dapat mendukung Kantor Desa Sambirejo Timur dalam pengelolaan data UMKM dan pengambilan keputusan yang lebih efektif, transparan, dan berbasis data. Secara keseluruhan, penelitian ini membuktikan bahwa algoritma C4.5 efektif digunakan dalam digitalisasi sistem klasifikasi UMKM serta memberikan kontribusi nyata dalam penerapan data mining pada sektor pemerintahan desa.

REFERENSI

- [1] Andikos, A. F. (2019). Perancangan Aplikasi Multimedia Interaktif Sebagai Media Pembelajaran Pengenalan Hewan Pada Tk Islam Bakti 113 Koto Salak. (Indonesia Jurnal Sakinah) Jurnal Pendidikan Dan Sosial Islam, 1(1), 34–49. <http://jurnal.konselingindonesia.com/>
- [2] Arianto Pradana, A. P., & Ibnu Hardi, I. H. (2021). Sistem Informasi Alat Kesehatan Berbasis Web. Jurnal Ilmiah Sistem Informasi, 1(1), 14–21. <https://doi.org/10.51903/juisi.v1i1.257>
- [3] Budiarto, E., Rino, R., Hariyanto, S., & Susilawati, D. (2022). Penerapan Data Mining Untuk Rekomendasi Beasiswa Pada SD Maria Mediatrix Menggunakan Algoritma C4.5. Algor, 3(2), 23–34. <https://doi.org/10.31253/algor.v3i2.1019>
- [4] Fanani, M. R., Rizaluddin, M., & Hidayat, H. (2024). Sosialisasi Produksi Penjualan Produk UMKM yang diminati konsumen menggunakan Algoritma C4.5 di Desa Karangdowo Kec. Kedungwuni Kab. Pekalongan. SOROT : Jurnal Pengabdian Kepada Masyarakat, 3(1), 22–25. <https://doi.org/10.32699/sorot.v3i1.6337>
- [5] Juan, K., & Budi, S. (2023). Pengembangan Menu Digital Menggunakan ReactJs Implementasi Hasil Belajar Studi Independen di Frontend Engineering Program Ruangguru CAMP (Career Acceleration Bootcamp). Jurnal Strategi, 5(1), 130–142.
- [6] Junianto, A., Hartiyah, S., & Mutmainah, K. (2023). Pengaruh Pengalaman Usaha, Skala Usaha, Pengetahuan Akuntansi, Dan Perspektif Informasi Terhadap Penerapan Sistem Informasi Akuntansi (Studi Empiris UMKM Di Kecamatan Wonosobo). Jamasy: Jurnal Akuntansi, Manajemen & Perbankan Syariah, 2(April), 11.
- [7] Murni, I., Br pa, A. S., Lubis, B. R., & Ikhwan, A. (2023). Pengamanan Pesan Rahasia dengan Algoritma Vigenere Cipher Menggunakan PHP. Journal on Education, 5(2), 3466–3476. <https://doi.org/10.31004/joe.v5i2.1027>

- [8] Nurliana Nasution, Feldiansyah Bakri Nasution, & Hasan, M. A. (2023). Pkm Pelatihan Pembuatan Web Berbasis Framework Codeigniter Untuk Siswa Smk. *J-COSCIS : Journal of Computer Science Community Service*, 3(1), 10–19. <https://doi.org/10.31849/jcscis.v3i1.11001>
- [9] Pranata, D., & Suharyanto, C. E. K. O. (2020). *Jurnal Comasie*. *Comasie*, 05(03), 97–106.
- [10] Purwanto, E., & Wulandari, R. I. (2023). Digital Marketing Sebagai Strategi Pengembangan Pasar Pada Umkm Di Kelurahan Rungkut Menanggal. *JICS : Journal Of International Community Service*, 2(02), 50–58. <https://doi.org/10.62668/jics.v2i02.725>
- [11] Putri, R. R., & Effendi, S. (2023). Pengaruh Persepsi, Pengetahuan Akuntansi, dan Skala Usaha terhadap Penggunaan Sistem Informasi Akuntansi Pada Usaha Mikro, Kecil, dan Menengah. *SCIENTIA JOURNAL : Jurnal Ilmiah Mahasiswa*, 5(5). <https://doi.org/10.33884/scientiajournal.v5i5.7481>
- [12] Ramadhan, A., Baskara, D. R., Khairulloh, F. A., Komarudin, M. I., Daffa, M. A., Fadhilah Khairuddin, R., Fatimah, S., Sari, W., & Suryaningrat. (2023). Sosialisasi Pengenalan Website Berbasis HTML SMK Muhammadiyah 02 Tangerang Selatan. *Jurnal Pengabdian Kepada Masyarakat*, 1(4), 294–298.
- [13] Rizaludin, M., & Fikriah, F. (2023). Prediksi Perilaku Pelanggan Pada Produk UMKM Batik Dengan Menggunakan Algoritma Decision Tree. *Teknomatika*, 13(02), 8–16.
- [14] Saputra, A. A., Pakpahan, A. G. S., Kurtubi, A., Amiruddin, A., Fridaniarta, B., Wicaksono, E. Y., Saputra, H., Putra, M. Y. A., & Azahra, R. Y. (2023). Pelatihan Dan Pembuatan Website Menggunakan Html Dan Css. *Beujroh : Jurnal Pemberdayaan Dan Pengabdian Pada Masyarakat*, 1(1), 119–125. <https://doi.org/10.61579/beujroh.v1i1.41>
- [15] Saputra, A., & Firdaus. (2022). Algoritma C4.5 Meningkatkan Daya Saing dan Keterampilan Teknologi Digital bagi Pelaku Usaha Mikro Kecil Menengah Kota Padang dengan Konsep Costumer Relationship Management (CRM). *Jurnal Pendidikan Tambusai*, 6, 1497–1505.
- [16] Sari, I. K. (2021). Blended Learning sebagai Alternatif Model Pembelajaran Inovatif di Masa Post-Pandemi di Sekolah Dasar. *Jurnal Basicedu*, 5(4), 2156–2163. <https://doi.org/10.31004/basicedu.v5i4.1137>
- [17] Srirahayu, A., & Pribadie, L. S. (2023). Review Paper Data Mining Klasifikasi Data Mining. *Jurnal Ilmiah Informatika Global*, 14(1). <https://doi.org/10.36982/jiig.v14i1.2981>
- [18] Sulistio, M. R., Suarna, N., & Nurdiawan, O. (2023). Analisa Penerapan Metode Clustering X-Means Dalam Pengelompokan Penjualan Barang. *Jurnal Teknologi Ilmu Komputer*, 1(2), 37–42. <https://doi.org/10.56854/jtik.v1i2.49>
- [19] Suseno, A. T., Amin, M. Al, & Mahardika, F. (2023). Analisis Pemberian Bantuan UMKM Menggunakan Algoritma K-NN dan C4.5. *Journal of Computer System and Informatics (JoSYC)*, 5(1), 248–256. <https://doi.org/10.47065/josyc.v5i1.4312>
- [20] Wahyudi, M., Septiani, N., Tanjung, M. R., Saputra, R. N., Khairul, & Iqbal, M. (2024). Pelatihan Design Untuk Meningkatkan Kreativitas Promosi UMKM. *Jurnal Pengabdian Harapan Bangsa*, 2(1), 176–181. <https://doi.org/10.56854/jphb.v2i1.159>
- [21] Zulfa, I., & Wanda, R. (2023). KLIK: Kajian Ilmiah Informatika dan Komputer Rancangan Sistem Informasi Akademik Berbasis Website Menggunakan PHP dan MySQL. *Media Online*, 3(4), 393–399. <https://djournals.com/klik>