

BAB I

PENDAHULUAN

1.1. Latar Belakang

Usaha Mikro, Kecil, dan Menengah (UMKM) memiliki peran vital dalam perekonomian daerah. Berdasarkan data Kementerian Koperasi dan UKM, lebih dari 64 juta unit UMKM berkontribusi terhadap lebih dari 60% Produk Domestik Bruto (PDB) nasional serta menyerap lebih dari 90% tenaga kerja (Kementerian Koperasi Republik Indonesia, 2024). Posisi strategis ini menjadikan UMKM sebagai sektor yang perlu dipetakan secara akurat guna mendukung perumusan kebijakan dan program pembinaan yang tepat sasaran. *Clustering* atau analisis klaster merupakan teknik data mining yang lazim digunakan untuk mengelompokkan entitas ke dalam segmen-segmen homogen berdasarkan karakteristik tertentu. Dengan segmentasi klaster, pemerintah daerah dapat mengidentifikasi pola atau kelompok UMKM yang memiliki karakteristik dan kebutuhan serupa, sehingga program pemberdayaan atau bantuan dapat disesuaikan per klaster secara lebih efektif (Marcelina et al., 2023; Syifa & Fahmi, 2021). Studi terbaru di Indonesia telah menerapkan *clustering* pada data UMKM mengenai pengelompokan produk UMKM berdasarkan data penjualan online menggunakan algoritme K-mean menghasilkan tiga klaster kinerja penjualan yang berbeda (tinggi, sedang, rendah), yang membantu UMKM merumuskan strategi pemasaran yang lebih tepat (Marcelina et al., 2023). Hasil klaster semacam itu memberikan gambaran yang berguna bagi pemangku kebijakan untuk meningkatkan daya saing UMKM sesuai segmennya (Marcelina et al., 2023; Syifa & Fahmi, 2021).

Namun, tantangan muncul karena karakteristik data UMKM sering bersifat campuran atau kombinasi atribut numerik dan kategorik. Contohnya, profil UMKM dapat mencakup data numerik seperti omzet, jumlah tenaga kerja, aset, serta data kategorik seperti sektor usaha, skala usaha (mikro/kecil/menengah), atau status legal. Metode *clustering* tradisional umumnya dirancang untuk tipe data tertentu saja. Jika tidak ditangani dengan tepat, hal ini dapat mengurangi kualitas pengelompokan dan kepercayaan terhadap hasil segmentasi. Selain itu, hasil klasterisasi yang tidak konsisten atau tidak stabil akan menyulitkan pengambilan

kebijakan publik misalnya jika pengelompokan UMKM berubah-ubah drastis dengan metode atau inisialisasi berbeda, maka sulit bagi pemerintah daerah untuk menetapkan program yang konsisten bagi segmen UMKM tertentu. Oleh karena itu, penelitian ini menyoroti pendekatan *Ensemble clustering* berbasis *Majority Voting* sebagai solusi untuk meningkatkan kualitas, konsistensi, dan stabilitas segmentasi UMKM di Kabupaten Sampang dibanding pendekatan individual.

Tiga algoritme *clustering* non-hierarkis yang sering digunakan untuk segmentasi adalah K-mean, K-mode, dan K-medoid. Masing-masing memiliki karakteristik dan keterbatasan tersendiri. K-mean populer untuk data numerik dan menghitung centroid kluster sebagai rata-rata dari data dalam kluster (Cendana & Kuo, 2024). Ia menggunakan jarak Euclidean sebagai metrik kesamaan, sehingga hanya cocok untuk variabel numerik kontinu. K-mean efektif dan sederhana, namun sensitif terhadap *outlier*; keberadaan nilai ekstrem dapat sangat mempengaruhi centroid dan menggeser hasil klusterisasi (Tan et al., 2022). Selain itu, K-mean mengasumsikan kluster berbentuk bulat dengan ukuran seragam, sehingga kurang sesuai untuk distribusi data yang non-sferis atau sangat bervariasi (Niu et al., 2020). Ia juga peka terhadap inisialisasi pusat kluster awal, di mana pemilihan pusat awal yang berbeda sering kali menghasilkan hasil kluster yang berbeda pula (Niu et al., 2020).

K-mode adalah perluasan K-mean untuk data kategorik (Huang, 1998). Algoritme ini menggantikan konsep mean dengan mode dalam menentukan pusat kluster dan menggunakan metrik disimilaritas sederhana seperti jarak Hamming. Dengan demikian, K-mode mampu mengelompokkan data nominal atau ordinal tanpa perlu transformasi ke numerik (Cendana & Kuo, 2024). Algoritma ini terbukti konvergen dan memberikan hasil pasti untuk data kategorik (Taufik et al., 2025). Namun, ia tidak dapat secara langsung menangani atribut numerik dan apabila digunakan pada data campuran tanpa penyesuaian, informasi numerik bisa hilang atau terdistorsi. Untuk menangani hal ini, beberapa peneliti menggunakan pendekatan seperti K-Prototypes yang menggabungkan prinsip K-mean dan K-mode (Ikotun et al., 2023).

K-medoid serupa dengan K-mean tetapi memilih medoid sebagai pusat

klaster, yakni objek aktual dalam data. K-medoid lebih tahan terhadap *outlier* karena pusat klaster merupakan titik data aktual, bukan nilai rata-rata. Algoritme ini dapat menggunakan metrik jarak yang beragam, termasuk Gower distance untuk data campuran. K-medoid cocok untuk data yang mengandung variabel numerik dan kategorik sekaligus, namun kompleksitas komputasinya lebih tinggi daripada K-mean dan hasilnya tetap bisa dipengaruhi oleh pemilihan awal medoid (Khedairia & Khadir, 2022; Tan et al., 2022).

Ketiga algoritma tersebut menghadapi keterbatasan saat menangani data UMKM yang bersifat campuran. K-mean tidak dapat memproses atribut kategorik, K-mode mengabaikan informasi numerik, dan K-medoid membutuhkan definisi metrik jarak yang tepat serta waktu komputasi yang lebih tinggi. Salah satu solusi adalah memisahkan terlebih dahulu atribut numerik dan kategorik, lalu mengelompokkan masing-masing dengan algoritme yang sesuai, kemudian menggabungkan hasilnya. Pendekatan ini telah diterapkan oleh (Putri et al., 2022) yang mengelompokkan data numerik dengan K-Medoid, data kategorik dengan QROCK, lalu hasil klaster digabung dan dikelompokkan ulang untuk menghasilkan klaster akhir yang lebih akurat dan stabil.

Ensemble clustering atau pengelompokan *Ensemble* merupakan teknik menggabungkan beberapa hasil klasterisasi menjadi satu hasil konsensus. Pendekatan ini memanfaatkan keragaman dari berbagai algoritme atau skenario *clustering* untuk menangkap struktur data yang berbeda, lalu menggabungkannya untuk menghasilkan klaster yang lebih baik, lebih stabil, dan lebih akurat (Yadav & Sharma, 2024). Salah satu pendekatan *Ensemble* yang efektif dan mudah dipahami adalah *Majority Voting*, di mana hasil klaster akhir ditentukan berdasarkan suara terbanyak dari beberapa algoritme dasar. Metode ini terbukti robust dan stabil, serta mengurangi risiko bias dari satu algoritme saja (Babu & Malathi, 2023).

Majority Voting dalam *Ensemble clustering* melibatkan proses di mana setiap algoritme memberikan label klaster untuk setiap objek data, dan label akhir dipilih berdasarkan label yang paling banyak muncul. Pendekatan ini telah diterapkan oleh (Babu & Malathi, 2023) dalam pengelompokan data besar dengan

mengombinasikan lima algoritme berbeda. Hasilnya menunjukkan bahwa akurasi dan stabilitas hasil *clustering* meningkat dibanding penggunaan algoritme tunggal. Selain itu, *Majority Voting* juga mudah diimplementasikan dan dapat dijelaskan kepada pemangku kepentingan non-teknis, menjadikannya cocok untuk konteks pengambilan kebijakan publik.

Validasi hasil *clustering* diperlukan untuk memastikan kualitas pengelompokan. Dua indeks validitas internal yang umum digunakan adalah *Silhouette Coefficient (SC)* dan *Davies-Bouldin Index (DBI)*. SC mengukur seberapa dekat suatu objek dengan anggota klasternya dibanding klaster lain, dengan nilai berkisar antara -1 hingga +1. Nilai mendekati +1 menunjukkan bahwa objek tersebut cocok dengan klasternya dan jauh dari klaster lain. Sementara *DBI* mengevaluasi rata-rata kesamaan antar klaster yang paling mirip, dengan nilai lebih rendah menunjukkan klaster yang lebih baik (Goldstein et al., 2024). Dengan menggunakan kedua indeks ini, peneliti dapat menentukan apakah klaster yang dihasilkan oleh *Ensemble* lebih baik dibanding klaster dari algoritme individu.

Konsistensi dan stabilitas segmentasi sangat penting dalam konteks kebijakan publik daerah. Hasil *clustering* yang tidak stabil dapat membingungkan pengambil kebijakan dan menyulitkan dalam pelaksanaan serta evaluasi program. *Ensemble clustering* menawarkan solusi untuk mengatasi masalah ini. Dengan menggabungkan hasil dari beberapa algoritma, *Ensemble clustering* menghasilkan segmentasi yang lebih konsisten dan tahan terhadap variasi data maupun parameter (Liu et al., 2022). Ini sangat penting untuk pengambilan keputusan jangka panjang, seperti dalam perencanaan bantuan dan pelatihan UMKM oleh pemerintah daerah.

Dengan mempertimbangkan berbagai tantangan pada data UMKM, keterbatasan algoritma individu, dan kebutuhan akan hasil segmentasi yang berkualitas dan stabil, maka pendekatan *Ensemble clustering* berbasis *Majority Voting* dipilih dalam penelitian ini. Dengan dukungan validasi indeks (*Silhouette* dan *DBI*) yang menunjukkan kualitas klaster baik, diharapkan hasil segmentasi UMKM di Kabupaten Sampang menggunakan *Ensemble clustering* dapat diandalkan sebagai *evidence-based input* dalam perencanaan program pengembangan UMKM daerah.

1.2. Perumusan Masalah

Pada bagian perumusan masalah, diuraikan mengenai permasalahan, metode usulan serta pertanyaan penelitian yang diuraikan sebagai berikut:

1.2.1. Permasalahan (*Problem Statement*)

Permasalahan utama dalam penelitian ini adalah belum tersedianya segmentasi UMKM Kabupaten Sampang yang stabil, adil, dan representatif akibat keterbatasan jika hanya menggunakan satu metode klasterisasi. Metode seperti K-mean hanya cocok untuk data numerik, K-mode untuk data kategorik, dan K-medoid untuk data campuran dengan keunggulan terhadap *outlier*, namun masing-masing menghasilkan struktur klaster yang berbeda dan tidak selalu sejalan. Hal ini menyebabkan hasil segmentasi menjadi tidak konsisten dan sulit dijadikan dasar pengambilan keputusan, terutama untuk penentuan sasaran bantuan, pelatihan, dan program digitalisasi UMKM. Oleh karena itu, dibutuhkan pendekatan yang mampu menggabungkan kelebihan ketiga metode tersebut agar dapat menghasilkan klaster yang lebih stabil, dapat dijelaskan dengan baik, serta relevan untuk kebijakan pembangunan UMKM secara berkelanjutan.

1.2.2. Metode Ulsulan

Metode yang diusulkan dalam penelitian ini adalah *Ensemble clustering* berbasis *Majority Voting*, yang menggabungkan tiga algoritma partisional yaitu K-mean untuk data numerik, K-mode untuk data kategorik, dan K-medoid dengan jarak Gower untuk data campuran. Setiap algoritma dijalankan pada subset data yang sesuai, lalu hasil klasterisasi digabung melalui voting mayoritas agar diperoleh label akhir yang lebih stabil dan tidak bergantung pada satu metode saja. Pemilihan jumlah klaster dilakukan dengan membandingkan nilai evaluasi internal seperti *Silhouette* dan *Davies–Bouldin Index*, serta mempertimbangkan kemudahan interpretasi lapangan. Pendekatan ini bertujuan menghasilkan segmentasi UMKM yang lebih akurat, adil, dan konsisten sehingga dapat langsung dimanfaatkan untuk merancang kebijakan bantuan, pelatihan, dan digitalisasi yang tepat sasaran.

1.2.3. Pertanyaan Penelitian

Berdasarkan uraian latar belakang, dapat diambil pernyataan penelitian diantaranya sebagai berikut:

1. Apakah pendekatan *Ensemble clustering* berbasis *Majority Voting* mampu menghasilkan segmentasi UMKM Kabupaten Sampang dengan kualitas klaster yang lebih baik dibanding metode tunggal?
2. Bagaimana karakteristik dan profil masing-masing klaster UMKM yang dihasilkan oleh pendekatan *Ensemble clustering* berbasis *Majority Voting*?

1.3. Tujuan & Manfaat

Pada bagian tujuan dan manfaat, diuraikan sasaran penelitian serta kontribusi yang diharapkan sebagai berikut:

1.3.1 Tujuan Penelitian

1. Mengetahui apakah pendekatan *Ensemble clustering* berbasis *Majority Voting* dapat menghasilkan segmentasi UMKM Kabupaten Sampang dengan kualitas klaster yang lebih baik dibandingkan metode tunggal.
2. Menganalisis karakteristik dan profil masing-masing klaster UMKM yang dihasilkan oleh pendekatan *Ensemble clustering* berbasis *Majority Voting*.

1.3.2 Manfaat

Berdasarkan rumusan masalah yang ada, dapat ditentukan manfaat dari penelitian ini diantaranya adalah sebagai berikut:

1. Memberikan bukti empiris bahwa *Ensemble Majority Voting* pada data campuran UMKM mampu meningkatkan validitas internal (SC lebih tinggi, *DBI* lebih rendah) serta stabilitas dibanding algoritme tunggal.
2. Menambah referensi metodologis tentang penanganan data campuran melalui kombinasi K-mean, K-mode, dan K-medoid dalam konteks segmentasi kebijakan publik daerah.
3. Menghasilkan peta segmen UMKM yang lebih tepat sasaran dan mudah diinterpretasi, sehingga memperjelas prioritas pembinaan per klaster.

1.4 Batasan-batasan

Berikut batasan-batasan pada penelitian ini:

1. Data diperoleh dari Dinas Koperasi, Perindustrian, dan Perdagangan Kabupaten Sampang, dengan fokus pada UMKM di Kabupaten Sampang tahun 2025 dengan jumlah sebanyak 6870 data.
2. Ruang lingkup data terbatas pada UMKM Kabupaten Sampang sesuai ketersediaan dan periode pencatatan yang disediakan; variabel yang digunakan mencakup gabungan fitur numerik omzet, aset, tenaga kerja, kapasitas produksi dan kategorik legalitas, *Marketplace*, social media setelah proses pembersihan dasar.
3. Metodologi dibatasi pada unsupervised *clustering* dengan K-mean, K-mode, K-medoid (jarak Gower) sebagai base learners dan *Majority Voting* sebagai mekanisme consensus.
4. Evaluasi berfokus pada validitas internal, tanpa ground truth eksternal dan tanpa klaim kausalitas; analisis waktu komputasi dicatat pada skala data Kabupaten secara deskriptif.

1.5 Sistematika Proposal

Sistematika penulisan proposal ini penulis susun dengan mengikuti aturan penulisan proposal pada program studi S1 Sistem Informasi Universitas Trunodjoyo Madura. Berikut merupakan metodologi penulisan proposal pada penelitian ini:

BAB I PENDAHULUAN

Bab ini menguraikan latar belakang penelitian, perumusan masalah, problem statement dari sudut pandang Informatika, metode usulan (kerangka *Ensemble Majority Voting*: K-mean, K-mode, K-medoid), pertanyaan penelitian, tujuan dan manfaat penelitian (bukan pengembangan perangkat lunak), batasan-batasan penelitian, serta sistematika penulisan proposal.

BAB II TINJAUAN PUSTAKA

Bab ini memaparkan teori dan konsep penunjang yang relevan, meliputi: karakteristik dan konteks UMKM Kabupaten Sampang, konsep data campuran pada unsupervised learning, dasar-dasar *clustering* partisional (K-mean, K-mode, K-medoid), jarak Gower untuk data campuran, gagasan *Ensemble*/consensus

clustering (dengan penekanan pada *Majority Voting* dan isu label alignment), serta metrik evaluasi internal (*Silhouette*, *DBI*). Bab ini juga memuat penelitian terdahulu yang relevan sebagai pijakan state of the art dan identifikasi celah riset.

BAB III METODE USULAN

Bab ini menjelaskan metodologi penelitian yang diusulkan, mencakup: tahapan penelitian (pra-proses data, pemilihan fitur numerik/kategorik, penanganan missing, standardisasi seperlunya), deskripsi dataset UMKM Kabupaten Sampang, perancangan dan pelaksanaan tiga base learners (K-mean, K-mode, K-medoid) dan fusi *Majority Voting*, serta skenario pengujian (evaluasi *Silhouette* & *DBI*) Alur diejelaskan secara operasional agar dapat direplikasi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil penerapan metode *clustering* pada data UMKM Kabupaten Sampang serta analisis terhadap hasil segmentasi yang diperoleh. Pembahasan meliputi hasil pengujian metode K-Means, K-Modes, K-Medoids, *Ensemble Majority Voting*, *Ensemble Weighted Voting*, dan Enhance K-Means, serta evaluasi menggunakan *Silhouette Coefficient* dan *Davies–Bouldin Index (DBI)* untuk menentukan metode dengan performa terbaik.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil penelitian mengenai metode *clustering* yang paling optimal dalam segmentasi UMKM di Kabupaten Sampang serta saran untuk pengembangan penelitian selanjutnya.

REFRENSI

Bab berisi tentang keterangan mengenai sumber rujukan yang digunakan dalam penyusunan Tugas Akhir. Pada penelitian ini referensi disusun secara otomatis dengan menggunakan aplikasi mendeley.