

BAB I

PENDAHULUAN

1.1 Latar Belakang

Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan tulang punggung perekonomian yang berperan krusial dalam penciptaan lapangan kerja, peningkatan daya saing daerah, dan pemerataan ekonomi (Ramadhan et al., 2025). Secara nasional, jumlah unit UMKM di Indonesia mencapai sekitar 64,2 juta, atau 99% dari seluruh pelaku usaha, dengan kontribusi terhadap produk domestik bruto (PDB) sekitar 61%. UMKM juga menyerap ~97% tenaga kerja nasional (sekitar 119,6 juta orang) (Berdaya et al., 2025), sehingga keberadaannya berdampak langsung pada pengurangan pengangguran dan kemiskinan. Di Provinsi Jawa Timur, Kabupaten Sampang termasuk daerah dengan kegiatan UMKM yang signifikan. Sampang memiliki jumlah UMKM yang besar dan tren Nilai Tambah Bruto (NTB) yang meningkat setiap tahun, menandakan potensi pertumbuhan ekonomi daerah yang didorong sektor UMKM (Fitriyah et al., 2024). Fakta-fakta tersebut menunjukkan betapa strategisnya pengembangan UMKM, termasuk di Kecamatan Sampang, sebagai objek penelitian ini.

Meskipun berkontribusi besar, UMKM menghadapi berbagai permasalahan dalam siklus operasionalnya. Penelitian profil UMKM Sampang menemukan sejumlah kendala utama yang kerap dialami pelaku UMKM, antara lain: terbatasnya jangkauan pemasaran, sulitnya akses atau mahalnya bahan baku, tingginya tingkat persaingan, minimnya modal usaha, kurangnya sumber daya manusia terampil, kendala distribusi/logistik, dan rendahnya kerja sama kemitraan (Fitriyah et al., 2024). Sebagai contoh, dalam hal pemasaran dan teknologi, banyak UMKM belum optimal memanfaatkan platform digital. Data survei Katadata tahun 2020 menunjukkan memang sekitar 60,2% pelaku UMKM sudah menggunakan media sosial untuk pemasaran produk dan 57,8% untuk berinteraksi dengan pelanggan (Putu Bagus Wiranata et al., 2024). Namun, ini juga mengindikasikan masih ada puluhan persen UMKM yang belum terdigitalisasi, sehingga mereka tertinggal dalam hal pemasaran

online dan akses pasar yang lebih luas. Selain itu, heterogenitas karakteristik UMKM (dari jenis usaha, skala modal, hingga kapasitas produksi) membuat pendekatan pembinaan yang seragam menjadi kurang efektif diterapkan (Ramdhan & Erlangga, 2025). Permasalahan-permasalahan ini dapat menghambat potensi pertumbuhan UMKM secara keseluruhan apabila tidak diatasi dengan strategi yang tepat.

Dari berbagai masalah tersebut, fokus utama penelitian ini adalah belum adanya segmentasi dan klasifikasi potensi UMKM berbasis data yang sistematis di Kecamatan Sampang. Selama ini, kebijakan dan program pendampingan cenderung bersifat umum dan belum didukung *model* kuantitatif yang mampu membedakan UMKM berpotensi tinggi, sedang, dan rendah secara jelas (Fitriyah et al., 2024). Akibatnya, pemerintah daerah kesulitan menentukan UMKM mana yang seharusnya diprioritaskan untuk akses pembiayaan, pelatihan digital, atau bantuan teknis; misalnya antara UMKM beromzet tinggi yang sudah go-digital dan UMKM kecil yang masih sangat tradisional. Di tingkat internasional, studi tentang segmentasi UMKM menunjukkan bahwa pemetaan segmen yang tepat sangat penting agar program dukungan inovasi dan pertumbuhan dapat lebih terarah (Dua et al., 2025). Beberapa pendekatan sudah dikembangkan untuk memetakan dan mengelompokkan pelaku usaha. Di sisi *unsupervised*, metode klasterisasi seperti *K-Means*, *K-Prototypes*, dan variasinya telah banyak digunakan untuk segmentasi pelanggan maupun UMKM karena mampu mengelompokkan data berdasarkan kemiripan atribut numerik dan kategorikal (Engelberth Worabai et al., 2023). Studi terbaru di Indonesia, misalnya, berhasil menggunakan *K-Prototypes* untuk mengelompokkan UMKM berdasarkan kombinasi variabel modal, pendapatan, dan sektor, sehingga menghasilkan beberapa klaster dengan karakteristik yang berbeda (Ramdhan & Erlangga, 2025). Di sisi *supervised*, berbagai *model* klasifikasi seperti *Logistic Regression*, *Random Forest*, dan *XGBoost* telah digunakan untuk memprediksi kategori kinerja, risiko kredit, maupun tingkat digitalisasi UMKM, dan menunjukkan kinerja yang menjanjikan pada data tabular (Purnamasari & Asa Verano, 2025). Penelitian lain bahkan menggabungkan *clustering* dan *supervised learning (clustering+classification)* untuk segmentasi UMKM digital dan analisis perilaku usaha (Marcelina, 2025).

Masing-masing solusi di atas memiliki keunggulan sekaligus keterbatasan. Pada metode *clustering*, *K-Means* unggul dari sisi kesederhanaan dan efisiensi komputasi untuk data berukuran sedang, tetapi mengasumsikan bentuk klaster yang cenderung sferis sehingga kurang andal ketika distribusi data UMKM tidak beraturan atau banyak *outlier*; dalam beberapa studi hal ini tercermin pada nilai *Silhouette coefficient* yang relatif rendah. *Hierarchical clustering* mampu mengungkap struktur hierarkis antar klaster dan cukup efisien pada skala kecil, namun kurang skalabel untuk data besar, sedangkan *DBSCAN* dapat membentuk klaster dengan bentuk arbitrer dan mengabaikan *noise* sehingga kualitas klaster bisa lebih baik, tetapi sangat sensitif terhadap pemilihan parameter dan cenderung lebih berat secara komputasi (Engelberth Worabai et al., 2023). *K-Prototypes* menawarkan kemampuan menangani kombinasi fitur numerik–kategorikal yang umum pada profil UMKM dan dapat menemukan segmen yang bermakna, tetapi masih mewarisi kelemahan *K-Means*, seperti kebutuhan menentukan jumlah klaster optimal dan ketergantungan pada inisialisasi *centroid* (Hasanah & Jannah, 2024). Pada teknik *supervised*, regresi logistik mudah diinterpretasi namun kurang mampu menangkap hubungan nonlinier antar faktor, sedangkan SVM dan *Random Forest* umumnya memberikan akurasi lebih tinggi tetapi dapat lambat atau menghasilkan *model* yang kompleks dan kurang transparan; *neural network* sendiri membutuhkan data besar, waktu pelatihan panjang, dan interpretasi yang tidak sederhana (Arora et al., 2023). XGBoost dikenal tangguh untuk klasifikasi maupun regresi karena menggabungkan akurasi tinggi dengan efisiensi komputasi melalui pemrosesan paralel dan *tree pruning*, namun berpotensi *overfitting* bila hiperparameternya tidak ditata dengan baik, dan sebagai *model* yang kompleks, relatif kurang mudah diinterpretasi tanpa analisis tambahan seperti *feature importance* atau SHAP. Hal-hal ini menjadi perhatian dalam merancang solusi baru agar metode terpilih tidak hanya akurat, tetapi juga cukup stabil dan dapat dijelaskan (Arora et al., 2023; Henriques et al., 2020).

Masing-masing solusi di atas memiliki keunggulan sekaligus keterbatasan. Di sisi *clustering*, *K-Means* unggul dalam kesederhanaan dan efisiensi komputasi untuk data berukuran sedang, namun cenderung tidak stabil saat menghadapi data yang tidak

terdistribusi dengan baik atau banyak *outlier*. Di sisi klasifikasi, meskipun XGBoost dikenal sangat tangguh dan efisien, *model* ini bisa mengalami *overfitting* jika hiperparameternya tidak ditata dengan baik dan kurang mudah diinterpretasi. Untuk mengatasi keterbatasan tersebut, *model K-Means* dan XGBoost dipilih karena kombinasi keduanya dapat mengatasi kekurangan yang ada pada masing-masing metode. *K-Means* dapat memberikan segmentasi yang efisien, sementara XGBoost dapat memanfaatkan hasil segmentasi ini untuk klasifikasi yang lebih akurat dan stabil, menjadikannya *model* yang lebih unggul dalam menangani data yang kompleks dan heterogen seperti yang ada pada UMKM. Untuk mengatasi kekurangan tersebut, penelitian ini mengusulkan *model K-Means* dan XGBoost sebagai solusi. Pada tahap pertama, *K-Means* digunakan untuk melakukan segmentasi UMKM berdasarkan indikator potensi (misalnya omzet, aset, produktivitas, dan skor digital), sehingga diperoleh beberapa kluster dengan karakteristik berbeda. Pada tahap kedua, hasil kluster ini dimanfaatkan oleh XGBoost untuk melakukan klasifikasi potensi, baik dengan menggunakan fitur kluster sebagai target maupun sebagai fitur tambahan dalam *model*. Pendekatan serupa terbukti efektif pada penelitian lain yang menggabungkan *K-Means* dan XGBoost untuk mendeteksi anomali pada data log berskala besar, di mana *K-Means* memisahkan pola data dan XGBoost menghasilkan aturan klasifikasi yang akurat dan interpretabel untuk data baru (Henriques et al., 2020). Selain itu, studi tentang segmentasi pasar dan UMKM menunjukkan bahwa integrasi *unsupervised clustering* dengan *supervised classification* mampu meningkatkan skalabilitas dan adaptivitas segmentasi, terutama saat menghadapi data yang terus berkembang (Dua et al., 2025). XGBoost sendiri telah berulang kali dilaporkan unggul dalam prediksi pada data tabular, termasuk untuk kasus UMKM dan perilaku online, sehingga cocok dipadukan dengan *K-Means* untuk menghasilkan *model* yang kuat namun tetap efisien (Marcelina, 2025).

Melalui penelitian ini, diharapkan diperoleh Dengan menerapkan *model K-Means* Dan XGBoost, penelitian ini menargetkan dua hasil utama yang terukur. Pertama, diharapkan terbentuk segmentasi UMKM Kecamatan Sampang yang valid dengan nilai indeks kualitas kluster (misalnya *Silhouette coefficient*) pada rentang yang wajar untuk

data sosial ekonomi, sehingga setiap klaster memiliki profil potensi yang jelas (Kuswardana et al., 2025). Kedua, *model* XGBoost yang dibangun di atas struktur klaster tersebut diharapkan mampu mengklasifikasikan potensi UMKM (misalnya tinggi, sedang, rendah) dengan tingkat akurasi yang lebih tinggi dibanding pendekatan *baseline*, mengacu pada studi-studi yang menunjukkan keunggulan XGBoost dan *model* integrasi dalam meningkatkan akurasi prediksi dan stabilitas performa (Henriques et al., 2020). Secara praktis, hasil ini akan menghasilkan peta segmen potensi UMKM beserta *model* klasifikasi otomatis yang dapat dimanfaatkan pemerintah daerah untuk mengidentifikasi UMKM prioritas dan merancang program pembinaan yang lebih tepat sasaran.

1.2 Rumusan Masalah

Perumusan masalah pada penelitian ini disusun untuk menjelaskan apa permasalahan utama yang dihadapi, pendekatan apa yang diusulkan untuk menyelesaikan permasalahan tersebut, serta pertanyaan penelitian apa yang akan dijawab melalui analisis dan pengujian metode yang dipilih.

1.2.1 Permasalahan

Dari sudut pandang informatika, pengelolaan data UMKM di tingkat daerah, termasuk Kecamatan Sampang, umumnya masih bersifat administratif dan deskriptif, sebatas pencatatan profil usaha seperti omzet, aset, sektor, dan jumlah tenaga kerja tanpa diolah lebih lanjut menjadi informasi analitis yang dapat mendukung pengambilan keputusan strategis. Padahal, data UMKM yang heterogen dan berjumlah besar sebenarnya berpotensi dimanfaatkan dengan teknik *data mining* dan *machine learning* untuk memperoleh segmentasi dan pemetaan potensi usaha. Hingga saat ini belum terdapat *model* analitik yang secara sistematis melakukan segmentasi dan klasifikasi potensi UMKM berbasis data, sehingga penentuan kelompok UMKM berpotensi tinggi, sedang, atau rendah masih cenderung mengandalkan penilaian manual dan subjektif dari pihak dinas atau pendamping, bukan berdasarkan pola data yang terukur. Kondisi ini menimbulkan beberapa masalah informatik, antara lain tidak adanya *model* segmentasi otomatis untuk mengelompokkan UMKM berdasarkan

kemiripan atribut kinerja dan karakteristik digital, tidak adanya *model* klasifikasi yang mampu memprediksi potensi UMKM baru ke dalam segmen tertentu secara konsisten, serta belum adanya pemanfaatan metode *machine learning* modern, seperti kombinasi *clustering* dan *ensemble learning*, untuk meningkatkan kualitas informasi yang dihasilkan dari data UMKM yang sudah tersedia.

1.2.2 Metode Usulan

Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan penerapan *model K-Means* dan *XGBoost* yang mengombinasikan teknik *unsupervised learning* dan *supervised learning* dalam satu kerangka permodelan. Pada tahap segmentasi, algoritma *K-Means* digunakan untuk mengelompokkan UMKM berdasarkan variabel yang merepresentasikan potensi, seperti omzet per tahun, aset, kapasitas produksi, jumlah tenaga kerja, dan indikator adopsi teknologi/digital, sehingga terbentuk beberapa klaster potensi (misalnya tinggi, menengah, dan rendah) yang dinilai kualitasnya menggunakan metrik seperti *Silhouette coefficient* dan/atau *Davies-Bouldin Index*. Tahap berikutnya menggunakan algoritma *XGBoost* (*Extreme Gradient Boosting*) sebagai *model* klasifikasi untuk memprediksi kelas potensi UMKM berdasarkan fitur yang sama, dengan evaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*. Melalui kombinasi kedua metode tersebut, diharapkan diperoleh *model* integrasi yang mampu melakukan segmentasi dan klasifikasi potensi UMKM secara lebih akurat dibandingkan penggunaan satu metode saja.

1.2.3 Pertanyaan Penelitian

Dengan mempertimbangkan kasus yang telah diidentifikasi dan pendekatan yang telah diajukan, berikut adalah beberapa pertanyaan yang muncul:

1. Seberapa baik algoritma *K-Means* dalam mengelompokkan UMKM di Kecamatan Sampang berdasarkan tingkat potensi usahanya?
2. Seberapa baik model *K-Means* dan *XGBoost* dalam menentukan kategori potensi UMKM di Kecamatan Sampang, serta apakah hasil pengelompokan dari *K-Means* dapat meningkatkan ketepatan prediksi dibandingkan *XGBoost* tanpa tambahan informasi cluster?

1.3 Tujuan dan Manfaat

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

1.3.1 Tujuan

1. Mengukur dan menganalisis kualitas segmentasi potensi UMKM di Kecamatan Sampang yang dihasilkan oleh *K-Means* menggunakan *Silhouette coefficient* sebagai metrik utama.
2. Menguji dan membandingkan kinerja klasifikasi potensi UMKM antara integrasi *model K-Means dan XGBoost* dan *model XGBoost Baseline*, serta menilai pengaruh penggunaan informasi kluster terhadap peningkatan akurasi, *precision*, *recall*, dan *F1-score*.

1.3.2 Manfaat

Setelah tujuan penelitian tercapai, manfaat yang diharapkan adalah sebagai berikut:

- a. Bagi Pemerintah Daerah: Dihasilkan sebuah *model* analisis (prototipe) yang mampu melakukan segmentasi dan klasifikasi potensi UMKM secara otomatis dan objektif berdasarkan data. *Model* ini dapat menjadi dasar pengambilan keputusan berbasis bukti (*evidence-based policy making*) untuk merancang program pemberdayaan, pelatihan, pendampingan, dan bantuan permodalan yang lebih tepat sasaran sesuai dengan segmen potensi UMKM (misalnya potensi tinggi, menengah, dan rendah).
- b. Bagi Pelaku UMKM: Secara tidak langsung, dengan adanya segmentasi dan klasifikasi potensi yang lebih jelas, pelaku UMKM berpeluang mendapatkan program pembinaan yang lebih relevan dengan kondisi usahanya. UMKM pada segmen potensi rendah, misalnya, dapat diprioritaskan untuk pelatihan dasar manajemen dan digitalisasi, sedangkan UMKM pada segmen potensi tinggi dapat difokuskan pada peningkatan kapasitas produksi dan perluasan pasar. Hal ini diharapkan dapat meningkatkan kinerja, daya saing, dan keberlanjutan usaha UMKM di Kecamatan Sampang.
- c. Bagi Akademisi: Penelitian ini dapat menjadi referensi ilmiah mengenai penerapan *model K-Means dan XGBoost* pada studi kasus nyata di bidang

UMKM dan ekonomi daerah. Selain itu, hasil penelitian dapat menjadi dasar bagi pengembangan penelitian lebih lanjut, misalnya dengan membandingkan algoritma *clustering* dan klasifikasi lain, menambahkan variabel analisis baru, atau mengintegrasikan *model* ke dalam sistem pendukung keputusan berbasis web atau aplikasi.

1.4 Batasan Masalah

Agar penelitian ini terarah dan fokus, maka ruang lingkup penelitian dibatasi pada hal-hal berikut:

1. Sumber dan Jenis Data: Data yang digunakan adalah data UMKM di Kecamatan Sampang yang diperoleh dari Dinas Koperasi, Usaha Mikro dan Tenaga Kerja setempat, dengan fokus pada data berjumlah 1056 data tahun 2024.
2. Objek Penelitian: Objek penelitian terbatas pada UMKM yang terdata secara resmi di wilayah administratif Kecamatan Sampang.
3. Metode: Metode yang menjadi fokus utama penelitian adalah *model K-Means* dan *XGBoost*, di mana *K-Means* digunakan untuk segmentasi (klasterisasi) potensi UMKM dan *XGBoost* digunakan untuk klasifikasi potensi berdasarkan hasil segmentasi tersebut.
4. Variabel: Atribut atau variabel yang digunakan dalam proses segmentasi dan klasifikasi terbatas pada data yang tersedia dari dinas terkait, seperti Nama Pemilik, Nama Usaha, Jalan, Desa, Sektor, NIB/SKU, Omzet/Thn, Aset, Sosmed, *Marketplace*, Kepemilikan Lahan, Jml. Tenaga Kerja, Kapasitas Produksi/Thn.
5. Luaran Penelitian: Penelitian ini berfokus pada pembangunan *model* analitis untuk segmentasi dan klasifikasi potensi UMKM serta analisis profil klaster yang dihasilkan guna mendukung pengambilan keputusan. Pengembangan perangkat lunak Sistem Pendukung Keputusan siap pakai dengan antarmuka pengguna lengkap tidak menjadi fokus utama, dan jika dibuat hanya sebatas prototipe analitik.

1.5 Sistematika Penulisan

Sistematika penulisan dari penelitian ini dapat dijelaskan sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian, perumusan masalah (termasuk permasalahan, metode usulan, dan pertanyaan penelitian), tujuan dan manfaat penelitian, batasan-batasan penelitian, serta sistematika penulisan proposal.

BAB II TINJAUAN PUSTAKA

Bab ini memuat landasan teori yang relevan dengan penelitian, antara lain konsep dan definisi UMKM, potensi UMKM, dasar-dasar *data mining* dan *machine learning*, teori mengenai algoritma *K-Means* dan XGBoost, serta penelitian-penelitian terdahulu yang berkaitan dengan segmentasi dan klasifikasi UMKM atau objek serupa.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan metode yang digunakan dalam penelitian, meliputi rancangan penelitian, sumber dan jenis data, variabel penelitian, tahapan pengolahan data (*preprocessing* dan rekayasa fitur), rancangan *model K-Means* untuk segmentasi, rancangan *model XGBoost* untuk klasifikasi, skema *model K-Means* Dan XGBoost, serta metode evaluasi yang digunakan (metrik kluster dan metrik klasifikasi).

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian, analisis data, dan evaluasi menggunakan *K-Means* dan XGBoost, serta pembahasan terkait temuan penelitian.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari penelitian serta saran untuk pengembangan UMKM Kecamatan Sampang dan penelitian lebih lanjut.

- PUSTAKA Daftar referensi yang digunakan dalam penelitian disertakan dalam bagian ini.
- LAMPIRAN Bagian ini mencakup lampiran yang berisi data dan informasi pendukung yang relevan dengan penelitian.