

BAB 3

METODE PENELITIAN

Penelitian ini bertujuan untuk membandingkan algoritma *K-Means* dan *Agglomerative clustering* dalam pengelompokan kategori UMKM berdasarkan nama desa, dan sektor ekonomi kreatif di Provinsi Daerah Istimewa Yogyakarta. Metode *clustering* ini digunakan untuk mengidentifikasi pola dan kelompok dalam data UMKM, guna memberikan pemahaman yang lebih mendalam mengenai distribusi dan karakteristik kategori UMKM di wilayah tersebut. Proses penerapan dilakukan secara bertahap, diawali dengan pengumpulan dan pra-pemrosesan data untuk memastikan bahwa data dalam kondisi siap untuk dianalisis, dilanjutkan dengan penerapan algoritma *K-Means* dan *Agglomerative clustering*.

Tahapan awal penelitian adalah Pra-pemrosesan data, yang mencakup pembersihan dan transformasi data agar dapat dianalisis secara efektif. Setelah itu, dilakukan Analisis *clustering* dengan menerapkan algoritma *K-Means* untuk membagi data menjadi beberapa kelompok berdasarkan jarak *Euclidean*, serta *Agglomerative clustering* yang membentuk kelompok data secara hierarkis. Kedua algoritma ini diaplikasikan untuk menemukan pola-pola relevan di antara UMKM berdasarkan lokasi desa dan sektor ekonomi kreatif. Hasil pengelompokan kemudian dianalisis dan dibandingkan guna mengevaluasi keakuratan serta efektivitas masing-masing algoritma dalam membentuk kelompok yang konsisten dan bermakna.

3.1 BAHAN DAN ALAT PENELITIAN

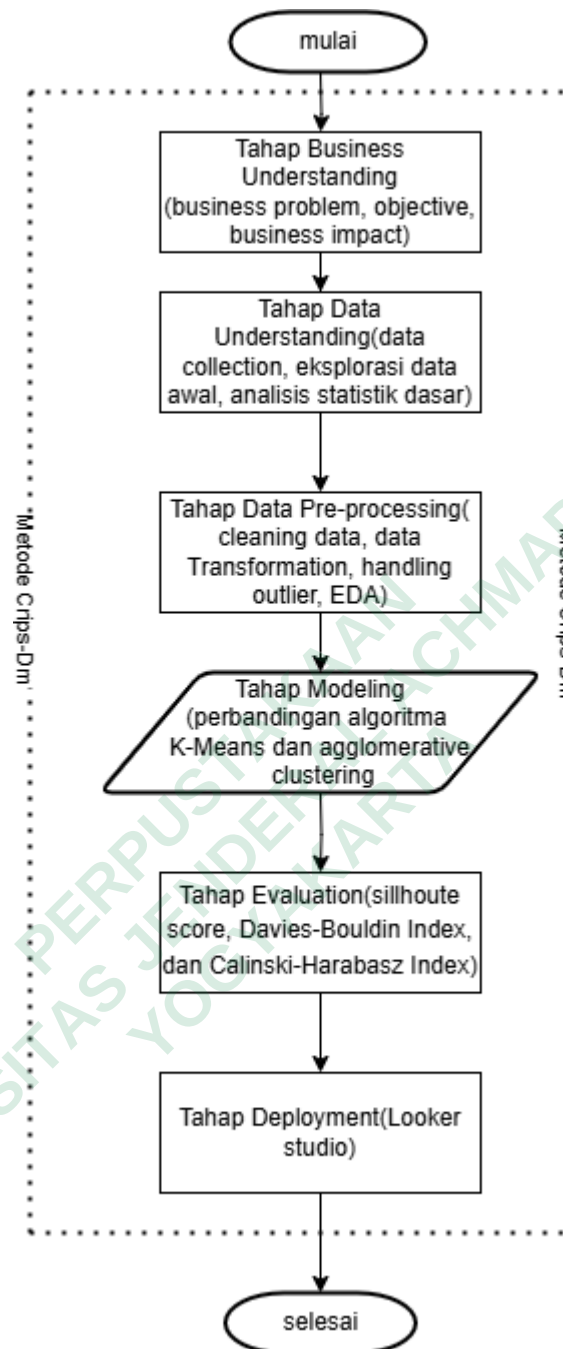
Bahan penelitian berupa data UMKM yang diperoleh dari Dinas Pertanian dan Tata Ruang DIY dengan total 25.065 data. Data tersebut mencakup berbagai informasi terkait usaha mikro, kecil, dan menengah yang tersebar di wilayah DIY. Variabel yang digunakan dalam penelitian ini meliputi nama desa, yang merepresentasikan lokasi keberadaan UMKM serta ekonomi kreatif, yang mencerminkan sektor usaha yang dijalankan oleh masing-masing UMKM.

Alat yang digunakan pada penelitian ini terdiri dari perangkat keras dan perangkat lunak yang mendukung analisis data dan penerapan algoritma *clustering*. Komputer yang digunakan harus memiliki spesifikasi yang memadai untuk menjalankan sistem operasi dan aplikasi yang diperlukan, serta terhubung dengan koneksi internet yang stabil. Sistem operasi dan aplikasi yang digunakan dalam penelitian ini meliputi:

1. Sistem Operasi: Windows 11, dengan RAM 12 GB yang dikenal karena kestabilannya dan dukungan yang baik untuk pengembangan perangkat lunak.
2. Bahasa Pemrograman: Python 3.10.12, yang menjadi bahasa pemrograman yang banyak digunakan untuk analisis data dan pemrograman algoritma.
3. *Library* untuk Analisis Data: Pandas, NumPy, dan Scikit-learn, yang menyediakan alat untuk manipulasi data, komputasi numerik, dan penerapan algoritma *clustering*.
4. *Software* untuk Visualisasi Data: Matplotlib dan Seaborn, Looker Studio, Tableau, Power BI digunakan untuk menghasilkan visualisasi yang membantu dalam interpretasi hasil analisis *clustering*.
5. *Code Editor*: Google Collab, yang menyediakan lingkungan pengembangan yang nyaman dan efisien untuk menulis serta menjalankan kode.

3.2 JALAN PENELITIAN

Metode penelitian yang digunakan pada penelitian ini adalah metode CRISP-DM dengan pendekatan metode *clustering* yang menggunakan perbandingan algoritma *K-Means* dan *Agglomerative clustering*. Metode ini menggambarkan langkah-langkah yang sistematis dalam proses pengumpulan, pemrosesan, dan analisis data untuk pengelompokan kategori UMKM berdasarkan nama desa, dan sektor ekonomi kreatif di Provinsi Daerah Istimewa Yogyakarta. Tahapan-tahapan dalam metode ini adalah sebagai berikut: 1) *Buseniss Understanding*; 2) *Data Understanding* 3) *pre-processing data*, 4) *Modelling*, 5) *Evaluation*, 6) *Deployment* yang dapat di lihat pada gambar 3.1



Gambar 3. 1 Jalan Penelitian

Pada gambar 3.1 merupakan jalan penelitian menggunakan metode CRISP-DM dengan pendekatan *clustering* yang membandingkan algoritma *K-Means* dan *agglomerative clustering*. Adapun tahapannya adalah sebagai berikut:

1. *Business Understanding*

Pada tahap *Business Understanding*, dilakukan identifikasi kebutuhan bisnis secara mendalam untuk merumuskan strategi pengelompokan kategori UMKM di Provinsi Daerah Istimewa Yogyakarta. Meskipun terdapat 25.065 UMKM yang tersebar di berbagai wilayah, data yang belum terstruktur dan terintegrasi menyulitkan pemangku kebijakan dalam memahami kondisi riil di lapangan. Kurangnya pengelompokan yang jelas membuat proses analisis dan penyusunan strategi pengembangan menjadi tidak optimal serta wisatawan yang kesulitan dalam melihat kategori UMKM. Oleh karena itu, dibutuhkan pendekatan berbasis data untuk mengelompokkan kategori UMKM secara efektif berdasarkan sektor ekonomi kreatif dan distribusi geografis, guna mendukung pengambilan keputusan serta pemberdayaan potensi ekonomi lokal secara lebih tepat sasaran.

2. *Data Understanding*

Data yang telah di kumpulkan perlu dilakukan eksplorasi untuk memahami karakteristik, distribusi, dan kualitasnya. Data yang digunakan mencakup informasi mengenai nama desa, dan sektor ekonomi kreatif dari 25.065 UMKM di Yogyakarta. Analisis awal dilakukan untuk mengidentifikasi pola, inkonsistensi, serta kemungkinan data yang hilang atau tidak valid. Visualisasi data dan statistik deskriptif digunakan untuk mendapatkan wawasan awal mengenai persebaran kategori UMKM berdasarkan wilayah dan jenis usaha. Proses ini bertujuan untuk memastikan bahwa data yang digunakan memiliki kualitas yang baik dan sesuai untuk analisis lebih lanjut dalam proses pengelompokan.

3. *Data Pre-Processing*

Pada tahap data *pre-processing*, verifikasi data dilakukan untuk memastikan konsistensi, kelengkapan, dan keakuratan data sebelum digunakan dalam analisis. Proses ini mencakup pengecekan apakah data sesuai dengan format yang diharapkan dan tidak mengandung nilai yang tidak valid atau tidak konsisten. *Cleaning data* melibatkan penghapusan atau perbaikan data yang hilang, duplikasi, dan kesalahan entri agar lebih reliabel. Sementara itu, *handling outlier* dilakukan

untuk mengatasi nilai yang jauh berbeda dari pola umum data, yang dapat memengaruhi hasil analisis. Metode seperti *interquartile range* (IQR), *z-score*, atau transformasi data digunakan untuk mendeteksi dan menangani *outlier*, baik dengan menghapus, mengganti, atau menyesuaikan nilainya agar tidak mengganggu proses pengelompokan.

4. *Modeling*

Peneliti menggunakan metode *clustering* dengan menggunakan Perbandingan antara algoritma *K-Means* dan *Agglomerative clustering* dalam proses pembentukan kluster kategori UMKM berdasarkan data yang telah melalui tahap pra-pemrosesan. Implementasi kedua algoritma ini menggunakan bahasa pemrograman Python dengan dukungan pustaka numerik dan pembelajaran mesin, seperti *scikit-learn*, yang menyediakan fungsi-fungsi statistik serta algoritma yang diperlukan dalam analisis klusterisasi. Algoritma *K-Means* diterapkan untuk mengelompokkan data ke dalam sejumlah kluster dengan tujuan meminimalkan variansi intra-kluster melalui iterasi penyesuaian pusat kluster (*centroid*). Pemilihan centroid awal dilakukan secara acak (random) dari data yang tersedia, sebagaimana metode standar pada algoritma *K-Means* untuk memulai proses pengelompokan (Nasution & Hasibuan, 2020), sehingga menghasilkan kluster dengan karakteristik yang lebih homogen. Sementara itu, *Agglomerative clustering* sebagai metode klusterisasi berbasis hierarki melakukan proses pengelompokan secara bertahap berdasarkan tingkat kemiripan antar-sampel data, dimulai dari setiap titik sebagai kluster tersendiri hingga terbentuk hierarki kluster yang lebih terstruktur (Shalileh, 2023). Perbandingan antara kedua algoritma ini bertujuan untuk mengevaluasi efektivitas dan akurasi masing-masing metode dalam mengelompokkan kategori UMKM berdasarkan aspek geografis dan sektor ekonomi kreatif, sehingga dapat diperoleh pendekatan klusterisasi yang paling optimal dan relevan dengan konteks penelitian.

5. *Evaluation*

Hasil perbandingan pengelompokan kategori UMKM yang telah diperoleh dievaluasi menggunakan metrik klasterisasi kuantitatif, seperti *Silhouette Score* dan *Davies-Bouldin Index*, yang berfungsi sebagai indikator efektivitas serta tingkat kesesuaian hasil klasterisasi. *Silhouette Score* memberikan penilaian terhadap konsistensi setiap titik data dalam klasternya dengan menghitung jarak rata-rata antara titik tersebut dan anggota klaster lainnya serta jarak rata-rata dengan klaster terdekat lainnya, sehingga dapat menilai separabilitas antar-klaster secara keseluruhan. Sementara itu, *Davies-Bouldin Index* mengukur rasio antara jarak antar-klaster dan penyebaran intra-klaster, dengan nilai yang lebih rendah menunjukkan klasterisasi yang lebih optimal dan terpisah dengan jelas. Melalui evaluasi ini, performa kedua algoritma *K-Means* dan *Agglomerative clustering* dapat dibandingkan secara objektif. Hasil perbandingan algoritma terbaik akan dipakai pada visualisasi pengelompokan kategori UMKM di Provinsi Daerah Istimewa Yogyakarta.

6. *Deployment*

Pada tahap akhir, peneliti melakukan interpretasi terhadap hasil pengelompokan yang telah terbentuk, serta mengevaluasi performa algoritma yang digunakan. Setelah dilakukan perbandingan dan analisis evaluasi, algoritma dengan performa terbaik dipilih untuk divisualisasikan menggunakan Looker studio guna memperjelas distribusi kategori UMKM di setiap wilayah. Visualisasi hasil klasterisasi ini bertujuan untuk memberikan pemahaman yang lebih jelas mengenai pola distribusi pengelompokan yang terbentuk, sehingga para pemangku kebijakan dan juga wisatawan dengan mudah dapat melihat kategori UMKM yang sesuai dengan minat mereka.