BAB I
PENDAHULUAN

1.1. Latar Belakang

Machine learning adalah teknik yang umum digunakan dalam pemodelan dengan cara mengenali pola pada data. Bidang ini merupakan cabang dari ilmu komputer yang memanfaatkan algoritma agar komputer dapat belajar dari data [1]. Pendekatan machine learning bertujuan untuk meniru dan menggantikan manusia dalam mengatasi masalah [2]. Secara umum, pembelajaran mesin diklasifikasikan menjadi tiga kategori utama, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning* [3]. Pada *unsupervised learning*, *clustering* menjadi salah satu teknik yang sering digunakan. *Clustering* bertujuan untuk mengelompokkan data yang memiliki karakteristik serupa ke dalam satu kelompok (*cluster*), sementara data yang berbeda akan ditempatkan dalam kelompok lain [4]. Secara umum, metode *clustering* terbagi menjadi dua jenis, yaitu metode clustering non-hierarki dan metode *clustering* hierarki [5].

K-Means merupakan salah satu metode *clustering* non-hierarki yang membagi data ke dalam dua atau lebih kelompok [6]. Algoritma ini banyak digunakan karena memiliki kecepatan komputasi yang tinggi serta konsepnya yang intuitif, di mana ia berusaha meminimalkan fungsi objektif dalam bentuk paling sederhana [7]. K-Means sering digunakan karena algoritmanya yang sederhana dan mudah dalam implementasi [8-9]. Cara kerja K-Means dimulai dengan memilih *k* titik awal sebagai *centroid* awal dari titik yang ada. Selanjutnya pada setiap iterasi, setiap titik akan dihitung jaraknya ke *centroid* menggunakan Euclidean Distance. Pengelompokan kelas dari setiap titik ditentukan berdasarkan jarak terdekat dari pusat *cluster*. Namun, jarak antara titik data ke *centroid* yang dihitung dengan Euclidean Distance memicu pemilihan *centroid* atau pusat *cluster* yang buruk dan terjebak pada solusi minimum lokal sehingga tidak mendapatkan solusi paling optimal dari data [10-11]. Terlebih lagi, pada penelitian [12] menulis bahwa *initial centroid cluster* pada K-Means ditentukan secara random menyebabkan masalah konvergensi. Oleh karena itu, pemilihan *centroid* untuk algoritma K-Means

menjadi masalah yang penting untuk diteliti guna meningkatkan kinerja algoritma K-Means.

Salah satu cara untuk mengatasi kekurangan algoritma K-Means tersebut dapat menggunakan pendekatan optimasi. Strategi optimasi umumnya digunakan dengan mengubah masalah pengelompokan menjadi fungsi objektif, di mana variabel yang akan yang diamati, diintegrasikan, dan diidentifikasi adalah *centroid*. Strategi optimasi ini dikategorikan menjadi dua jenis; pendekatan optimasi metaheuristik dan pendekatan optimasi deterministik. Metode deterministik sering kali membutuhkan banyak asumsi untuk penerapannya, tidak seperti pendekatan metaheuristik yang umumnya lebih mudah beradaptasi. Oleh karena itu, penelitian ini akan menggunakan pendekatan metaheuristik. Algoritma metaheuristik adalah strategi pencarian canggih yang dirancang untuk menyelesaikan masalah optimasi secara efisien dengan mengeksplorasi ruang solusi melalui metode khusus [13].

Pada penelitian sebelumnya telah banyak yang menggunakan pendekatan metaheuristik untuk optimasi algoritma *clustering*, termasuk K-Means. Terdapat dua penelitian yang relevan terhadap penggunaan metode metaheuristik untuk optimasi K-Means, Penelitian pertama [8] mengimplementasikan algoritma *Bee Colony* untuk mencari titik pusat K-Means. Penelitian tersebut menghasilkan akurasi 83.16%-83.30% lebih baik dari K-Means yaitu 83.09%. Penelitian kedua [14] menggunakan algoritma Firefly dalam optimasi K-Means *clustering*. Hasil yang diperoleh adalah algoritma Firefly mampu mengungguli dua algoritma lainnya (K-Means dan PSO) dengan *F-Score* rata-rata dari lima dataset yang diuji coba adalah 0,786, sedangkan untuk K-Means 0,76 dan PSO 0,773. Optimasi algoritma K-Means juga dilakukan oleh [11] menggunakan Modified Particle Swarm Optimization untuk mencari pusat *cluster* terbaik. Algoritma Modified Particle Swarm Optimization (MPSO) yang dibandingkan dengan K-Means murni dan algoritma klasik berbasis PSO terbukti memberikan kinerja yang lebih baik di semua aspek.

Selain metode-metode di atas, Differential Evolution juga menjadi metode yang dapat diklasifikasikan sebagai pendekatan metaheuristik dan berguna dalam menemukan solusi permasalahan optimasi yang cukup kompleks. Algoritma

tersebut bertujuan mencari nilai parameter untuk mengoptimalkan fungsi seperti meminimalkan atau memaksimalkan [15]. Differential Evolution juga digunakan pada penelitian-penelitian terdahulu untuk mengoptimasi suatu algoritma. Penelitian [16] menggabungkan Artificial Neural Network dan Differential Evolution untuk mengklasifikasi kanker kulit. Peneliti menggunakan 2 dataset, yaitu HAM10000 dan PH2 *dermatoscopic*. Hasilnya, *accuracy* ANN-DE mencapai sekitar 97,4% lebih tinggi dari algoritma *Genetika* (GA)-ANN 92-94%, ANN 86-88%, dan SVM 86-89%. Penelitian yang lainnya [17] menerapkan Differential Evolution terhadap algoritma LightGBM untuk mendapatkan kombinasi struktur parameter dan performa model yang optimal. Kesimpulan dalam penelitian [17] menghasilkan bahwa algoritma DE (mampu menemukan optimasi optimal dengan cepat yaitu hanya sekitar 15 generasi optimasi.

Berdasarkan penjelasan dan tinjauan pada penelitian sebelumnya, pada penelitian ini akan diajukan metode Differential Evolution (DE) untuk menentukan *initial centroid* dari algoritma K-Means. Sehingga solusi yang diperoleh dapat bersifat global dan kinerja algoritma K-Means pun akan semakin baik. Jenis data yang akan digunakan pada penelitian ini ada dua yaitu data hasil *generate random* dan data survei potensi desa dari Badan Pusat Statistik tahun 2021. Data tersebut akan digunakan dalam proses pencarian *centroid* awal optimal dengan DE. Perolehan *centroid* awal yang optimal selanjutnya akan digunakan untuk modelling menggunakan K-Means.

Untuk mengevaluasi algoritma yang diajukan, kinerja dari algoritma K-Means+DE akan dibandingkan dengan kinerja K-Means asli berdasarkan pengukuran menggunakan Silhouette Score dan *running time* sehingga pada tahap selanjutnya dapat dilakukan analisis hasil. Selain dibandingkan dengan K-Means asli, K-Means+DE juga akan dibandingkan dengan K-Means kombinasi algoritma optimasi metaheuristik lain, yang dalam hal ini adalah Genetic Algorithm (GA). Algoritma GA dipilih karena merupakan algoritma optimasi berbasis populasi, sama dengan algoritma DE. Setelah semua perbandingan dilakukan, langkah terakhir adalah mengimplementasikan algoritma K-Means+DE terhadap data asli yakni, data survei potensi desa tahun 2021.

Dengan mengoptimasi algoritma K-Means menggunakan Differential Evolution, diharapkan penelitian ini dapat menghasilkan algoritma *clustering* yang memiliki performa yang lebih handal serta dapat melakukan pengelompokan lebih akurat. Hasil penelitian ini juga diharapkan mampu menjadi referensi dan solusi dalam pekerjaan c*lustering* bagi peneliti sehingga model yang akan digunakan mampu menghasilkan keputusan yang tepat dalam pengelompokan data.

1.2. Rumusan Masalah

Algoritma *clustering* K-Means memiliki kelemahan dalam mencari *centroid* yang optimal yang mana *initial centroid* ditentukan secara random serta sering terjebak pada local optima sehingga *clustering* yang dihasilkan cenderung kurang baik. Oleh karena itu untuk mendapatkan suatu solusi yang bersifat global, penentuan titik pusat awal seharusnya tidak dilakukan secara acak. Kekurangan dari algoritma K-Means dalam penentuan *centroid* awal atau *initial centroid* tersebut dapat ditutupi dengan pendekatan optimasi metaheuristik yang dalam hal ini adalah Differential Evolution. Oleh karena itu, penelitian ini dilakukan untuk mengamati performa dari algoritma K-Means yang dioptimasi menggunakan Differential Evolution melalui *initial centroid* optimal.

1.3. Pertanyaan Penelitian

1. Bagaimana kinerja algoritma Differential Evolution mengoptimasi K-Means dalam menentukan *initial centroid* ?
2. Apakah Differential Evolution mampu meningkatkan kinerja K-Means dalam melakukan pengelompokan/*clustering* pada data?
3. Bagaimana perbandingan kinerja K-Means sebelum dan setelah dilakukan optimasi menggunakan algoritma Differential Evolution?
4. Bagaimana perbandingan performa antara algoritma K-Means+DE dan K-Means+GA ?
5. Bagaimana Silhouette Score dan jumlah *cluster* terbaik pada implementasi K-Means+DE terhadap data survei potensi desa tahun 2021?

1.4. Batasan Masalah

1. Terdapat 2 jenis data yang digunakan pada penelitian ini, yaitu data yang didapatkan dari *generate* data random serta data survei potensi desa dari Badan Pusat Statistik tahun 2021.
2. Jumlah *cluster* ditentukan di awal penelitian.
3. Data asli pada penelitian ini terbatas pada penggunaannya untuk mengamati kinerja algoritma K-Means+*DE*.

1.5. Tujuan Penelitian

1. Penelitian ini menggunakan metode Differential Evolution (DE) dalam penentuan *initial centroid* pada metode K-Means.
2. Meningkatkan kinerja algoritma K-Means menggunakan algoritma DE dalam melakukan *clustering*.
3. Membandingkan kinerja K-Means sebelum dan setelah dilakukan optimasi menggunakan algoritma Differential Evolution.
4. Melakukan komparasi antara K-Means+DE dan K-Means+GA.
5. Mengamati implementasi K-Means+DE dalam mengelompokkan data asli berdasarkan evaluasi Silhouette Score dan Jumlah Cluster.

1.6. Manfaat Penelitian

1. Mengetahui fenomena penentuan *initial centroid* menggunakan algoritma Differential Evolution.
2. Mendapatkan hasil *clustering* yang menggunakan *centroid* optimal dari proses algoritma Differential Evolution dan K-Means *clustering*.
3. Dengan penelitian ini, diharapkan dapat memicu penelitian yang lebih luas terkait penentuan suatu titik pusat atau *centroid* optimal pada suatu *cluster* khususnya algoritma K-Means.