

BAB I

PENDAHULUAN

Pada bab I menjabarkan isi terkait latar belakang permasalahan, rumusan masalah, tujuan dan manfaat, batasan masalah, metodologi penelitian, dan sistematika penulisan dalam penelitian ini.

1.1 Latar Belakang

Bakteri *Mycobacterium tuberculosis*, umumnya dikenal sebagai bakteri tahan asam, adalah penyebab *tuberkulosis* (TB), penyakit menular kronis. Ada berbagai spesies *Mycobacterium*, termasuk *M. tuberculosis*, *M. africanum*, *M. bovis*, dan *M. leprae* (BTA). *MOTT* (*Mycobacterium Other Than Tuberculosis*) adalah sekelompok bakteri *Mycobacterium* yang bukan *Mycobacterium tuberculosis* dan yang terkadang dapat mengganggu diagnosis dan pengobatan TB. Biasanya, bakteri penyebab TB menyerang paru-paru. Terkadang penyakit ini juga berdampak pada otak, ginjal, tulang, kelenjar getah bening, dan organ lainnya. Sebagai penyakit yang ditularkan melalui udara yang dapat menyebar melalui udara, penyakit TB dapat menyebar melalui batuk, bersin, dan berbicara. [1].

Salah satu negara yang memberikan kontribusi signifikan terhadap kasus TBC global adalah Indonesia. Di Indonesia saat ini, TBC masih menjadi masalah. Di Indonesia, terdapat 420.994 kasus baru TBC pada 2017. (data per 17 Mei 2018). Pria mengalami 1,4 kali lebih banyak kasus baru *tuberkulosis* pada tahun 2017 daripada wanita. Prevalensi *tuberkulosis* pada pria tiga kali lebih tinggi daripada pada wanita, menurut Studi Prevalensi TB. Hal yang sama berlaku di negara lain. Ini mungkin karena pria lebih mungkin terkena faktor risiko TB seperti merokok dan ketidakpatuhan terhadap rejimen pengobatan. Menurut laporan tersebut, 68,5% responden pria merokok dan 3,7% perempuan yang merokok [2].

Prevalensi *tuberkulosis* (TB) di Kabupaten Bangkalan relatif tinggi. Hingga 998 penduduk setempat mungkin telah mengalami infeksi TB pada tahun 2021. Hal ini disampaikan Rizkiyah Nunik, Kepala Bidang P3 Dinas Kesehatan (Dinkes) Bangkalan Bidang Pengendalian dan Pemberantasan Penyakit. Ia mengklaim pada tahun 2020, terdapat 1.024 orang yang mengidap TBC. Orang dengan infeksi TB

turun 2,5 persen dari 2021 hingga hari ini. Bahkan, mengingat TB adalah penyakit menular, persentase ini dapat dikurangi hingga 100%. [3].

Banyak individu memanfaatkan komputer sebagai alat karena perkembangan pesat teknologi yang terus berkembang, terutama di ranah komputer. Belum lagi komputer telah dimasukkan ke dalam bidang kedokteran sebagai alat untuk melakukan tugas atau mendiagnosis kondisi tertentu yang memfasilitasi pekerjaan manusia [4]. Saat ini, database pasien yang mencari perawatan di Rumah Sakit Syarifah Ambami Rato Ebu terutama digunakan sebagai arsip dan untuk melihat riwayat medis pasien. Penambahan data harus digunakan untuk menganalisis data rekam medis yang sudah ada untuk mengungkapkan informasi baru atau mengidentifikasi pola yang dapat digunakan untuk analisis, evaluasi, atau untuk menggambarkan keadaan pasien tertentu.

Penambahan data umumnya merupakan proses menggunakan teknik kecerdasan buatan, statistik, matematika, dan pembelajaran mesin untuk memvalidasi dan mengidentifikasi informasi yang sangat relevan dan pengetahuan terkait dari berbagai basis data yang luas, mengingat pesatnya perkembangan teknologi yang semakin kompleks. [5]. Salah satu metode data mining yaitu *clustering*. *Clustering* adalah suatu metode membagi data menjadi beberapa kelompok agar data dengan karakteristik yang sama dikelompokkan kedalam *cluster* yang sama [6]. Pada dasarnya *clustering* adalah strategi untuk mengidentifikasi dan mengkategorikan data dengan sifat-sifat yang *similarities* dengan data lainnya. *clustering* merupakan teknik data mining yang diterapkan tanpa pelatihan atau pembelajaran sebelumnya dan tidak membutuhkan target *output* tertentu [6].

Dalam penambahan data, ada dua jenis teknik pengelompokan, yaitu pengelompokan hierarkis dan pengelompokan non-hierarkis. Mengingat laju kemajuan teknologi yang semakin cepat, Data *K-Means* adalah salah satu teknik pengelompokan non-hierarkis yang paling banyak digunakan. Karena mudah, mudah diterapkan, dan mampu mengelompokkan volume data yang sangat besar, *K-Means* adalah teknik pengelompokan yang paling populer [6]. Jarak rata-rata antara setiap *cluster* data dapat dikurangi menggunakan *K-Means*. Data dapat diurutkan menjadi satu atau beberapa *cluster* menggunakan *K-Means*. Data akan

dikelompokkan menurut fitur serupa terlebih dahulu, kemudian menurut atribut yang berbeda kedua, dan seterusnya. [7].

Kelemahan mendasar dari *K-Means* adalah bahwa hasil *cluster* hanya menghasilkan optimal lokal dan sangat bergantung pada inisialisasi *centroid* yang dihasilkan secara acak pada awal proses. [7]. Optimum lokal adalah nilai optimal yang didapatkan merupakan nilai yang berada dalam rentang nilai tertentu, sedangkan global optimum adalah nilai optimal yang didapatkan merupakan nilai dari keseluruhan rentang dataset. Oleh karena itu, diperlukan mekanisme untuk menentukan inisialisasi *centroid* yang tepat untuk mendapatkan global terbaik. Algoritma genetika dapat digunakan sebagai cara untuk menyiasati kelemahan *K-Means* [8]. *Centroid* awal yang sangat baik untuk *K-Means* dapat diproduksi menggunakan pendekatan ini. Algoritma evolusioner dipilih karena memiliki karakter pencari solusi global, sehingga sulit untuk terperosok dalam optimasi lokal. Pusat *cluster* yang diturunkan oleh optimasi lokal memiliki kualitas yang mirip dengan iterasi sebelumnya, di mana jika data pengamatan adalah outlier, hasil pengelompokan akan jauh lebih buruk. Akibatnya, optimasi lokal sering dihindari. [8].

Dalam penelitian sebelumnya, pendekatan GA dan *K-Means* digabungkan. Ditemukan bahwa jumlah keseluruhan iterasi turun, menghemat waktu saat membuat kluster untuk mencari informasi yang diperlukan [9].

Penelitian lain menggunakan metodologi yang sama untuk mengelompokkan prestasi akademik dosen. Kombinasi prosedur GA dan *K-Means* menghasilkan nilai SC yang lebih besar daripada pengelompokan menggunakan pendekatan *K-Means* saja, menurut nilai evaluasi *Silhouette Coefficient* (SC). Nilai evaluasi, SC *K-Means* = 0,3803 dan SC GAK-*Means* = 0,4352 [10].

Penelitian serupa yang mengelompokkan data tsunami sebagai sarana mitigasi bencana. Berdasarkan perbandingan nilai Koefisien Siluet, hasil pengujian pada *clustering* menggunakan algoritma GA dan *K-Means* memiliki kualitas *cluster* yang lebih tinggi yaitu 0,9959 dibandingkan dengan pendekatan *K-Means* tanpa algoritma genetika yang memiliki nilai SC sebesar 0,8831 [11].

Karena dataset yang digunakan dalam penelitian ini memiliki struktur data yang tidak biasa dan nilai untuk beberapa atribut memiliki jangkauan yang luas, maka

penelitian lain yang menggunakan pendekatan normalisasi Z-Score juga digunakan. Penentuan parameter nilai K dalam penelitian ini kemudian menggunakan Particle Swarm Optimization sehingga dapat meningkatkan akurasi klasifikasi K-Nearest Neighbor dan dengan demikian meningkatkan kinerja K-Nearest Neighbor. Rumus Jarak Euclidean digunakan dalam pendekatan K-Nearest Neighbor [12].

Berdasarkan uraian sebelumnya dan dengan mempertimbangkan kelebihan dan kekurangan metode *K-Means*, penulis memutuskan untuk melakukan penelitian menggunakan *K-Means* dengan optimasi Algoritma Genetika dalam mengklasterkan data penyakit *Tuberkulosis* di Kabupaten Bangkalan. Data *Tuberkulosis* akan diproses menggunakan Algoritma Genetika untuk mencari *centroid* terbaik, kemudian hasilnya akan digunakan sebagai *centroid* awal pada *K-Means clustering*. Sehingga diharapkan dapat membagi kelompok penderita *Tuberkulosis* kedalam *cluster-cluster* yang memiliki kemiripan karakteristik berdasarkan dari kedekatan nilai-nilai atributnya. Untuk mengukur kualitas dan kekuatan *cluster* yaitu seberapa baik suatu objek ditempatkan dalam satu *cluster* dengan menggunakan uji evaluasi *Silhouette Coefficient*.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Algoritma *K-Means* banyak digunakan karena prosesnya sederhana, mudah diimplementasikan, dan dapat mengklasterkan data yang besar. Namun kelemahan utama *K-Means* adalah hasil *cluster* yang bergantung pada inisialisasi *centroid* yang dibangkitkan secara *random* pada awal proses, sehingga hasil *clustering* sering terjebak pada solusi yang bersifat optimum lokal. Dengan demikian pemilihan *centroid* awal dapat menentukan hasil partisi dari dataset, sehingga dibutuhkan suatu mekanisme untuk mencapai optimum global dengan menentukan inisialisasi *centroid* yang tepat.

1.2.2 Usulan Solusi

Berdasarkan permasalahan yang telah dijabarkan, maka peneliti mengusulkan Algoritma Genetika untuk menghasilkan *centroid* awal yang digunakan pada *K-Means Clustering* dalam pengelompokan data *Tuberkulosis*.

1.2.3 Pertanyaan Penelitian

Bagaimana pengaruh Algoritma Genetika dalam penentuan *centroid* awal *K-Means Clustering* terhadap hasil evaluasi *Silhouette Coefficient* pada pengelompokan dataset *Tuberkulosis*?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Mengetahui pengaruh Algoritma Genetika dalam penentuan *centroid* awal *K-Means* terhadap hasil evaluasi *Silhouette Coefficient* pada pengelompokan data *Tuberkulosis*.

1.3.2 Manfaat

Manfaat yang diperoleh dari penelitian ini adalah peneliti dapat mengetahui pengaruh Algoritma Genetika sebagai metode untuk mengoptimasi *centroid* pada *K-Means* dalam membangun sebuah sistem pemodelan *clustering*.

1.4 Batasan Masalah

Adapun batasan masalah pada penelitian ini, antara lain:

1. Data yang digunakan berupa data penyakit Tuberculosis yang didapatkan dari RSUD Syarifah Ambami Rato Ebu mulai dari 2017-2020 dalam waktu catur wulan dengan jumlah 985 data.
2. Parameter yang digunakan yaitu 5 parameter, yaitu umur, jenis kelamin, foto toraks, status HIV (*human immunodeficiency virus*), dan riwayat diabetes.
3. Algoritma yang digunakan untuk melakukan *clustering* yaitu Algoritma *K-Means*.
4. Algoritma yang digunakan sebagai optimasi titik pusat *cluster (centroid)* awal yaitu Algoritma Genetika.
5. Evaluasi *cluster* menggunakan *Silhouette Coefficient*.
6. Menggunakan bahasa pemrograman *Python*.

1.5 Sistematika Penulisan

Sistematika proposal secara garis besar terdiri dari:

1. BAB I PENDAHULUAN

Pada Bab I berisi tentang latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah, dan sistematika penulisan.

2. BAB II KAJIAN PUSTAKA

Pada Bab II berisi tentang penelitian - penelitian terdahulu yang juga mengimplementasikan Algoritma *K-Means Clustering* dan optimasi *centroid* awal Algoritma Genetika dalam penelitiannya dan terdapat pembahasan teori-teori yang dapat membantu untuk menyelesaikan penelitian ini.

3. BAB III METODE USULAN

Pada Bab III berisi tentang kebutuhan sistem, deskripsi sistem, data penelitian, *preprocessing* data, arsitektur sistem, dan skenario uji coba

4. BAB IV HASIL DAN PEMBAHASAN

Pada bab IV berisi tentang hasil dan pembahasan, bagian ini menjelaskan mengenai implementasi pembuatan sistem dan juga uji coba sistem.

5. BAB V PENUTUP

Pada bab V berisi tentang kesimpulan dan saran untuk pengembangan penelitian selanjutnya.

6. REFERENSI

Berisi semua pustaka atau referensi yang digunakan untuk penyusunan proposal skripsi.