

BAB I PENDAHULUAN

1.1 Latar Belakang

Penyakit adalah istilah medis yang digambarkan sebagai gangguan dalam fungsi tubuh yang menyebabkan berkurangnya kapasitas kerja pada organ tubuh. Penyakit dapat terjadi ketika keseimbangan dalam tubuh tidak dapat dipertahankan. Keadaan sakit terjadi pada saat seseorang tidak lagi berada dalam kondisi sehat yang normal [1]. Tingkat keparahan suatu penyakit dapat menyebabkan kematian pada penderitanya. Penyakit dapat menyerang orang tubuh manapun, termasuk dengan organ pernapasan.

Salah satu penyakit pernapasan yang berbahaya adalah penyakit *Tuberculosis* atau tuberkulosis paru. *Tuberculosis* atau dikenal dengan TB Paru merupakan penyakit yang mematikan setelah HIV/AIDS [2]. Penyakit tuberkulosis paru adalah penyakit menular langsung yang disebabkan oleh bakteri *Mycobacterium tuberculosis* yang sebagian besar menyerang paru-paru. Selain itu bakteri TB juga dapat menyerang organ tubuh lain yang biasa disebut dengan TB ekstra paru. TB paru adalah bentuk paling umum dari sekitar 80% dari semua penderita. TB yang menyerang jaringan paru-paru merupakan bentuk penyakit yang mudah TB menular. [3]. Oleh karena itu, berdasarkan lokasi atau organ tubuh yang sakit, *Tuberculosis* atau TB dapat diklasifikasikan ke dalam dua *class*, yaitu TB paru dan TB ekstra paru.

Klasifikasi merupakan pengelompokan data yang mana data tersebut mempunyai kelas label ataupun sebuah target. Dengan berkembangnya ilmu pengetahuan, dalam melakukan pengklasifikasian suatu data atau objek, dapat dibantu dengan menggunakan sistem yang menerapkan ilmu olah data. Telah ada beberapa metode yang digunakan untuk melakukan kegiatan klasifikasi, seperti metode *Long Short Term Memory* (LSTM). *Long Short Term Memory* (LSTM) adalah sebuah jenis khusus dari algoritma *Recurrent Neural Network* (RNN) yang mampu mempelajari ketergantungan jangka panjang. Berdasarkan penelitian yang telah dilakukan oleh Arfan dan Lussiana, 2020, LSTM memiliki nilai *loss* yang lebih baik jika dibandingkan dengan SVR (*Support Vector Regression*). Sehingga penggunaan LSTM mampu melakukan prediksi dengan hasil yang lebih akurat [4].

Dalam beberapa kasus yang berhubungan dengan klasifikasi terdapat beberapa permasalahan, salah satunya adalah masalah *missing value* pada data yang digunakan. *Missing value* merupakan suatu permasalahan dimana terdapat ketidak lengkapan nilai dalam suatu data. *Missing value* dapat terjadi karena beberapa faktor. Beberapa faktor tersebut, diantaranya adalah pengumpulan data dari sebuah pengamatan tidak berlangsung baik dan permasalahan responden yang tidak memberikan jawaban dari beberapa pertanyaan tertentu pada saat peninjauan. Adanya *missing value* dapat mempengaruhi hasil keakuratan dari proses oleh data dalam suatu sistem. Penelitian yang dilakukan oleh Achmad Fikri Sallaby dan Azlan, 2021 menjelaskan bahwa klasifikasi dapat dilakukan jika data sudah lengkap. Berdasarkan penelitian tersebut, pengklasifikasian yang melalui proses imputasi terhadap data kosong terlebih dahulu, memiliki nilai akurasi yang lebih tinggi jika dibandingkan dengan proses pengklasifikasian tanpa melalui imputasi [5]. Imputasi merupakan upaya mengatasi *missing value* dengan melakukan perhitungan untuk menentukan nilai pengganti data. Hal ini yang mendorong perlunya mencari metode khusus untuk penanganan imputasi terhadap permasalahan yang berkaitan dengan penanganan *missing value*.

Algoritma *k-Nearest Neighbor* (*k-NN*) merupakan salah satu metode imputasi yang dapat digunakan dalam mengatasi permasalahan *missing value*. Pada penelitian yang dilakukan oleh Taufiq Rizaldi, dkk, 2018 membahas bahwa penanganan *missing value* dengan menggunakan metode *k-NN* memiliki nilai MSE (*Mean Square Error*) yang lebih kecil, jika dibandingkan dengan menggunakan metode *Naïve Bayes*. Dimana, apabila semakin kecil nilai MSE, maka kinerja dari metode tersebut dianggap semakin baik [6]. Langkah penting dalam Algoritma *k-Nearest Neighbor* (*k-NN*) adalah menentukan nilai *k* dan jarak antar satu *instance* dengan seluruh *instance* pada *dataset*. Dimana sejumlah *k* tetangGaterdekat dari *dataset* dijadikan sebagai nilai estimator. Penentuan nilai *k* harus tepat dan benar. Hal tersebut karena, penentuan nilai *k* yang tidak tepat dapat menurunkan hasil akurasi

Pada beberapa penelitian yang telah dilakukan sebelumnya, seperti pada penelitian yang dilakukan oleh Susanti, dkk, 2018 [7] dan pada penelitian oleh Achmad Fikri Sallaby dan Azlan, 2021 [5]. Pada penelitian tersebut, penentuan

nilai k dalam penggunaan algoritma *K-Nearest Neighbor* (k -NN) untuk penanganan *missing value* masih dilakukan secara sembarang. Hal ini berpotensi menyebabkan kurang maksimalnya nilai akurasi dari pengolahan data yang dilakukan.

Untuk mengatasi masalah tersebut, maka diperlukan upaya dalam menentukan nilai k yang tepat dengan melakukan optimasi terhadap nilai k pada algoritma *K-Nearest Neighbor* (k -NN). Berdasarkan penelitian yang telah dilakukan oleh Abidatul Izzah dan Nurhayatin, 2013, pencarian nilai k terbaik pada k -NN menggunakan GA menghasilkan kinerja yang lebih baik jika dibandingkan dengan metode k -NN yang menggunakan imputasi *mean* dan median [8]. *Genetic Algorithm* (GA) digunakan sebagai upaya dalam memecahkan permasalahan optimasi tersebut. *Genetic Algorithm* (GA) merupakan salah satu dari algoritma komputasi evolusi. Seperti pada penelitian yang dilakukan oleh Ispandi dan Romi Satria Wahon, 2015. Penelitian tersebut melakukan optimasi parameter SVM menggunakan GA dalam kasus prediksi pemasaran langsung [9]. Penelitian lain yang melakukan proses optimasi juga dilakukan oleh Warid Yunus, 2018. Penelitian tersebut melakukan optimasi parameter k pada k -NN menggunakan PSO dalam kasus prediksi penyakit ginjal kronik [10]. Berdasarkan kedua metode optimasi tersebut (GA dan PSO), optimasi menggunakan GA memiliki akurasi yang lebih unggul 1,64 % jika dibandingkan dengan optimasi menggunakan PSO. Dengan menggunakan algoritma ini, *Genetic Algorithm* (GA) dapat mencari solusi terbaik dari kemungkinan – kemungkinan solusi yang tersedia. Pada penelitian yang dilakukan ini, solusi terbaik dari *Genetic Algorithm* (GA) digunakan sebagai nilai k yang dalam menyelesaikan imputasi *k-Nearest Neighbor* (k -NN). Dalam penelitian ini penulis merancang sebuah sistem pencarian nilai k terbaik milik Algoritma *k-Nearest Neighbor* (k -NN) menggunakan *Genetic Algorithm* (GA) untuk penanganan *missing value* pada data *tuberculosis*.

1.2 Perumusan Masalah

Berdasarkan latar belakang di atas, maka terbentuklah beberapa rumusan masalah yang dijelaskan pada sub bab berikut.

1.2.1 Permasalahan

Hilangnya nilai dari suatu data dapat mempengaruhi hasil akurasi dalam proses pengklasifikasian melalui metode LSTM. Dalam penanganan *missing value*, *k-Nearest Neighbor* (*k*-NN) merupakan metode yang paling populer. Namun, metode ini memiliki kelemahan pada pemilihan nilai *k* yang dapat mempengaruhi kinerja klasifikasi. *Genetic Algorithm* (GA) digunakan sebagai metode optimasi dalam menentukan nilai *k* terbaik pada *k*-NN. Namun, pemilihan parameter GA juga mempengaruhi solusi yang dihasilkan.

1.2.2 Metode Usulan

Metode usulan adalah dengan menggunakan *Genetic Algorithm* (GA) untuk mencari nilai *k* terbaik pada metode *k-Nearest Neighbor* (*k*-NN) dalam menangani data yang hilang. Namun, sebelumnya dilakukan beberapa percobaan untuk menguji berbagai parameter pada GA agar menghasilkan solusi (nilai *k*) yang terbaik. Dengan menggunakan nilai *k* yang terbaik, proses imputasi *k*-NN dapat dilakukan untuk mengisi data yang hilang dan memastikan akurasi klasifikasi LSTM yang lebih baik.

1.2.3 Pertanyaan Penelitian

Berikut pertanyaan – pertanyaan penelitian terkait proposal skripsi berjudul “Pencarian Nilai *k* Terbaik pada *k*-NN Menggunakan *Genetic Algorithm* (GA) dalam Penanganan *Missing Value* pada Klasifikasi *Tuberculosis*” :

1. Berapa nilai terbaik dari parameter, meliputi *crossover rate* (*cr*), *mutation rate* (*mr*), *population size*, dan *maximum generation* pada *Genetic Algorithm*?
2. Berapa nilai *neuron* dan *epoch* terbaik pada *Long Short Term Memory*?
3. Berapa perbandingan nilai evaluasi berupa akurasi dari hasil pengklasifikasian *tuberculosis* menggunakan *k*-NN+GA+LSTM dengan hasil pengklasifikasian *tuberculosis* menggunakan *k*-NN+ LSTM tanpa optimasi?

1.3 Tujuan Dan Manfaat

Berdasarkan rumusan masalah dan usulan solusi pada sub bab sebelumnya, didapatkan tujuan dan manfaat yang dijelaskan pada sub bab berikut.

1.3.1 Tujuan

Adapun tujuan yang dihasilkan dari penelitian ini meliputi :

1. Untuk mengetahui nilai terbaik dari parameter, meliputi *crossover rate* (*cr*), *mutation rate* (*mr*), *population size*, dan *maximum generation* pada *Genetic Algorithm*.
2. Untuk mengetahui nilai *neuron* dan *epoch* terbaik pada klaifikasi LSTM melalui varian imputasi *k*-NN.
3. Untuk mengetahui berapa perbandingan nilai evaluasi berupa akurasi dari hasil pengklasifikasian *tuberculosis* menggunakan *k*-NN+GA+LSTM dengan hasil pengklasifikasian *tuberculosis* menggunakan *k*-NN+ LSTM tanpa optimasi.

1.3.2 Manfaat

Manfaat yang didapatkan dalam usulan penelitian yang dilakukan yaitu, dapat menghasilkan sebuah pengetahuan baru. Dengan melakukan optimasi nilai Algoritma *k-Nearest Neighbor* (*k*-NN) menggunakan *Genetic Algorithm* (GA), maka dapat menghasilkan *output* dengan akurasi yang lebih baik.

1.4 Batasan

Beberapa batasan dalam usulan penelitian yang telah dilakukan meliputi :

1. Penelitian dilakukan di wilayah Kabupaten Bangkalan, Provinsi Jawa Timur.
2. Penelitian menggunakan data *tuberculosis* milik RSUD Syarifah Ambami Rato Ebu pada tahun 2017 sampai dengan 2020 yang berjumlah 985 data.
3. Pengklasifikasian data *tuberculosis* menggunakan 6 buah atribut atau parameter, meliputi umur, jenis kelamin, foto toraks, status HIV, riwayat diabetes, dan hasil TCM (Test Cepat Molekuler).
4. Terdapat 741 data yang memiliki nilai *missing value* pada seluruh atribut.
5. Tidak ada korelasi antar atribut.
6. Pengklasifikasian data *tuberculosis* menghasilkan 2 buah *class*, yaitu *class* paru dan ekstra paru.

1.5 Sistematika Proposal

Beberapa sistematika proposal penelitian ini disusun ke dalam tiga bab, dengan rincian sebagai berikut.

BAB I PENDAHULUAN

Bab I mengulas mengenai latar belakang, perumusan masalah, tujuan dan manfaat, batasan masalah, serta sistematika proposal.

BAB II KAJIAN PUSTAKA

Bab II membahas mengenai beberapa ilmu terkait metode usulan pilihan dan beberapa penelitian terkait.

BAB III METODE PENELITIAN

Bab III mengulas mengenai jenis penelitian, langkah-langkah penelitian, perancangan sistem, *dataset* yang digunakan, metode evaluasi, skenario ujicoba, dan perkiraan jadwal penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab IV mengulas mengenai hasil implementasi sesuai arsitektur sistem dan hasil uji coba sesuai dengan skenario uji coba.

BAB V KESIMPULAN DAN SARAN

Bab V mengulas mengenai kesimpulan dan saran. Kesimpulan didapatkan dari hasil uji coba yang telah dilakukan. Kemudian, saran digunakan untuk pengembangan penelitian terkait selanjutnya.