

# **BAB I**

## **PENDAHULUAN**

### **1.1. Latar Belakang**

Tuberkulosis ialah suatu penyakit yang bisa menular, dan bersifat mematikan dalam banyak kasus dan menurut *World Health Organization (WHO)* termasuk penyakit berbahaya karena merupakan salah satu dari sepuluh besar pemicu kematian di dunia. Indonesia menjadi salah satu negara tertinggi yang terinfeksi penyakit Tuberkulosis [1]. Penyakit ini berasal dari berbagai strain mikobakteria, yaitu *Mycobacterium Tuberkulosis*. Tuberkulosis umumnya menyerang paru-paru, akan tetapi juga dapat berdampak pada bagian tubuh lainnya dan sebagian besar penderita TB ialah seseorang dengan rentang usia 15-55 tahun [2]. Sehingga kurang lebih 80% bakteri menyerang pada organ paru-paru [1]. Penyakit Tuberkulosis di Indonesia menjadi sangat serius karena penularannya sangat mudah. Penyakit yang dulunya dikenal dengan TBC kini menjadi TB Paru dapat menular melalui udara yang biasanya ketika penderita akan mengalami bersin atau batuk maka bakteri tersebut akan keluar sehingga terhirup oleh orang sehat lainnya.

Pada bidang kesehatan sebagian besar banyak menggunakan *data mining* untuk mendiagnosis penyakit, misalnya yaitu mendiagnosis kanker payudara, penyakit Tuberkulosis, jantung dan lain-lain [3]. Akan tetapi dalam mendiagnosis membutuhkan data lengkap karena hal tersebut mempengaruhi hasil diagnosis yang akurat. Masalah pada sebuah proses data yang terdapat di *data mining* di sebut *missing data*.

Data hilang atau *missing data* atau *missing value* adalah nilai kosong atau nilai tidak lengkap dalam data. Dalam banyak riset sering mengalami kejadian *missing data*. Fenomena ini diakibatkan oleh berbagai macam masalah, antara lain ketidaktepatan peneliti, metode pengumpulan data yang tidak memadai, timbulnya kesalahan program saat pengumpulan data, dan sebagainya [4].

Dalam kasus ini metode untuk mendapatkan diagnosis yang akurat diantaranya menggunakan *k-Nearest Neighbor (k-NN)* dan *naïve bayes*, akan tetapi

dalam metode naïve bayes sendiri tidak dapat dipakai pada data yang cukup besar jadi metode yang di gunakan dalam penelitian ini ialah metode *k-Nearest Neighbor* (*k-NN*). Metode *k-Nearest Neighbor* (*k-NN*) ialah metode yang sebagian besar dipakai untuk mengatasi missing data atau missing value. Metode *k-NN* memberikan hasil imputasi yang sangat baik bahkan ketika data yang digunakan memiliki missing data yang cukup besar [5], namun, metode ini memiliki kekurangan karena mengharuskan peneliti memasukkan nilai *k* untuk menemukan data yang hilang. Dalam penentuan nilai *k* yang kurang tepat dapat menurunkan kinerja *k-NN*.

Masalah penentuan nilai *k* yang optimal dapat diselesaikan dengan bantuan pendekatan metaheuristik, yang dapat menemukan solusi terbaik untuk suatu masalah dengan memperbaiki kandidat solusi secara iteratif. Pendekatan metaheuristik yang dapat digunakan diantaranya *Particle Swarm Optimization* (PSO), *Ant Colony Optimization* (ACO), dan *Genetic Algorithm* (GA) [6]. Namun dalam penelitian ini menggunakan pendekatan *Particle Swarm Optimization* (PSO) untuk mencari nilai *k* yang optimal. *Particle swarm optimization* adalah suatu metode pencarian metaheuristik yang di kemukakan oleh seorang Russell Eberhart dan James Kennedy pada tahun 1995 berlandasan dari perilaku komunikasi *sekelompok* binatang dan iteraksi sosial, diantaranya sekawanan burung dan sekelompok ikan [7]. metode ini memiliki banyak kelebihan yaitu parameter yang sedikit, untuk penerapannya mudah, cepatnya konvergensi, dan sederhana sehingga banyak digunakan untuk optimasi fungsi, optimasi metode konvensional, dan klasifikasi pola [8]. Dari banyaknya kelebihan yang dimiliki PSO metode ini tidak dapat mencari ruang solusi lebih luas sehingga permasalahan ini menjadi kelemahannya [9].

Penelitian yang di lakukan oleh PSO-(*k-NN*) akan dilanjutkan dengan pengklasifikasian menggunakan *Long Short Term Memory* (LSTM). *Long Short Term Memory* (LSTM) adalah salah satu jenis RNN [10]. LSTM mempunyai tiga gerbang yaitu gerbang input, gerbang forget dan gerbang output [11]. LSTM menyimpan informasi tentang pola dalam data. Beberapa gerbang di setiap neuron LSTM mengatur sendiri memori tiap neuron, sehingga LSTM dapat mengetahui data mana yang harus disimpan dan mana yang harus dibuang. Data teks, video,

dan deret waktu (*time series*) biasanya diproses menggunakan LSTM [12]. LSTM tidak memiliki kekurangan karena dapat mengatur memori untuk setiap input menggunakan sel memori dan gerbang unit [13]. Sehingga, saat memproses data sekuensial yang panjang, LSTM sering digunakan sebagai solusi untuk mengatasi vanishing gradient di RNN.

Berdasarkan permasalahan tersebut penelitian ini bertujuan untuk mengetahui bagaimana performa yang dihasilkan metode PSO dalam penentuan nilai  $k$  optimal pada metode KNN dalam melakukan imputasi missing data pada diagnosis Tuberkulosis.

## **1.2. Perumusan Masalah**

Perumusan masalah pada penelitian ini sebagai berikut :

### **1.2.1. Permasalahan**

Keakuratan missing data pada diagnosis Tuberkulosis data sangat di butuhkan. Permasalahan dalam missing data pada diagnosis TB sering terjadi karena data yang kurang akurat akibat data yang hilang. Dimana untuk menghasilkan diagnosis data yang akurat perlu adanya imputasi data yang tepat. Oleh karena itu, imputasi missing data pada diagnosis Tuberkulosis sangat tepat dalam kinerja ini dengan menggunakan *k-Nearest Neighbor (k-NN)* yang dioptimasi dengan *Particle Swarm Optimization* untuk menghasilkan nilai  $k$  optimal karena kelemahan *k-Nearest Neighbor (k-NN)* yaitu penentuan nilai  $k$  yang kurang tepat dapat menurunkan kinerja *k-NN*. Setelah menghasilkan data lengkap untuk menguji keakuratan data dilakukan klasifikasi dengan menggunakan *Long Short Term Memory*.

### **1.2.2. Metode Usulan**

Solusi yang digunakan untuk mendapatkan sistem dengan kinerja terbaik adalah menggunakan *k-Nearest Neighbor (k-NN)* dengan pendekatan *Particle Swarm Optimization (PSO)* untuk mencari nilai  $k$  yang optimal.

### 1.2.3. Pertanyaan Penelitian

1. Berapakah nilai  $k$  optimal pada metode *k-Nearest Neighbor (k-NN)* yang dioptimasi menggunakan metode *Particle Swarm Optimization (PSO)* untuk melakukan imputasi missing data ?
2. Bagaimana membandingkan kinerja tingkat akurasi menggunakan klasifikasi *PSO-(k-NN) LSTM* dengan *(k-NN-LSTM)* ?

### 1.3. Tujuan & Manfaat

Tujuan dan manfaat dari penelitian ini yaitu:

#### 1.3.1. Tujuan Penelitian

1. Mengimplementasikan imputasi missing data dengan metode *PSO* dalam penentuan nilai  $k$  optimal pada metode *k-NN*.
2. Mengukur tingkat akurasi menggunakan klasifikasi *PSO-(k-NN) LSTM* dengan *(k-NN)-LSTM*.

#### 1.3.2. Manfaat

Hasil penelitian dapat dijadikan referensi bagi peneliti lain yang membahas tentang Imputasi Missing Data Menggunakan Metode *k-Nearest Neighbor* Dengan Optimasi *Particle Swarm Optimization (PSO)* Pada Penyakit Tuberkulosis.

### 1.4. Batasan-batasan

1. Data berupa penyakit Tuberkulosis yang didapatkan dari RSUD Syarifah Ambami Rato Ebu mulai dari 2017-2020 dalam waktu catur wulan dengan jumlah 985 data.
2. Metode yang di gunakan untuk melakukan missing value pada data Tuberkulosis yaitu Algoritma *k-Nearest Neighbor (k-NN)*.
3. Metode yang digunakan untuk melakukan optimasi terhadap parameter  $k$  pada Algoritma *k-Nearest Neighbor (k-NN)* yaitu menggunakan Algoritma *particle swarm optimization (PSO)*.

4. Parameter yang digunakan untuk pengklasifikasian yaitu 6 parameter, yaitu jenis kelamin, foto toraks, status HIV (*human immunodeficiency virus*), riwayat diabetes, dan lokasi anatomi.
5. Pengklasifikasian dikelompokkan menjadi 2 kelas, yaitu paru-paru dan ekstra paru.
6. Pengklasifikasian pada data Tuberkulosis menggunakan Algoritma *Long Short Term Memory (LSTM)*.

### **1.5. Sistematika Proposal**

Usulan rencana penyusunan sistematika penulisan laporan tugas akhir. Penyusunan sistematis laporan tugas akhir ini meliputi:

#### **BAB I PENDAHULUAN**

Pada bab ini membahas tentang latar belakang dari penelitian, perumusan, solusi dari masalah tersebut, tujuan, manfaat, batasan – batasan, dan sistematika proposal.

#### **BAB II KAJIAN PUSTAKA**

Pada bab ini menjelaskan teori – teori yang mendukung penelitian, serta menjadi dasar dalam pemecahan masalah.

#### **BAB III METODOLOGI PENELITIAN**

Pada bab ini menjelaskan apa saja langkah-langkah penelitian yang akan dilakukan.

#### **BAB IV HASIL DAN PEMBAHASAN**

Bab ini berisi hasil implementasi sistem dengan metode yang digunakan. Hasil implementasi terdiri dari lingkungan uji coba dan hasil uji coba sistem.

#### **BAB V PENUTUP**

Bab ini berisi kesimpulan dari penelitian yang disesuaikan dengan pertanyaan penelitian dan saran terkait pengembangan aplikasi selanjutnya.