BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi telah memasuki lingkup dunia kesehatan, mulai dari sistem layanan hingga diagnosa penyakit berbasis sistem pakar. Namun, deteksi dini mengenai jenis penyakit yang ada masih terbatas sehingga semakin banyak orang atau organisasi yang menggunakan algoritma untuk menganalisis koleksi data yang besar dalam mendiagnosa atau memprediksi sesuatu, dan teknologi semacam ini juga banyak digunakan, seperti bidang medis biasanya memprediksi beberapa penyakit parah pada tahap awal, misalnya, kanker serviks. Kanker serviks adalah salah satu penyakit yang semakin mengancam kesehatan wanita (Deng *et al*., 2019) kanker serviks menjadi penyebab kematian yang signifikan di negara-negara berpenghasilan rendah, lebih dari setengah juta kasus per tahun, dan menewaskan lebih dari seperempat juta pada periode yang sama (Fernandes *et al*., 2017).

Dari kajian literatur menampilkan bahwa kualitas hasil diagnosa penyakit masih ditemukan banyak kelemahan seperti keterbatasan data penyakit hingga akurasi hasil diagnosa, hal ini disebab kan oleh berbagai faktor seperti kurangnya data pengetahuan yang dimiliki oleh sistem, ketidaklengkapan data (*incomplete data*), dan metode optimasi yang kurang tepat. Seperti pada penelitian sebelumnya, Choudhury *et al* membandingkan optimasi 4 metode yaitu *Gaussian naive bayes* (GNB), *Decision tree* (DT), *Support Vector Mechines* (SVM) dan *Logistic Regregession* (LR) dimana pada penelitian tersebut *Decision tree* (DT) jauh lebih unggul dibandingkan metode optimasi lainnya. Dengan dataset sejenis Wu, W., dan Zhou, H. Juga melakukan proses optimasi perbandingan metode *support vector mechine-recursive feature elimination* (SVM-RFE) dan *Support Vector Mechine Principal Component Analysis* (SVM-PCA), dengan hasil SVM-

PCA memiliki kemampuan yang lebih baik meskipun dengan jumlah fitur yang sama.

Permasalahan lain yang muncul adalah banyaknya atribut yang ada berdampak pada lamanya waktu pengambilan keputusan, dalam beberapa tahun terakhir, banyak metode deteksi diusulkan dan diterapkan di lapangan untuk memberikan diagnosis tepat waktu, termasuk pendekatan berbasis data (Wu, W., & Zhou, H., 2017) . pada penelitian ini metode *Optimum Index factor* (OIF) digunakan untuk mereduksi 32 atribut dengan mengukur nilai statistik dari pemilihan kombinasi optimal yang ada. Namun pendekatan berbasis data juga harus di dukung dengan data yang lengkap, agar tidak adanya nilai data yang hilang. Masalah optimasi untuk data yang tidak lengkap merupakan tantangan bagi pengembangan penelitian data karena teknik yang paling banyak digunakan mengasumsikan data lengkap tetapi tidak menyesuaikan dengan masalah *Incomplete* data (Yildirim., 2018). Pada penelitian tersebut Yildirim melakukan perbandingan analisis subsitusi nilai rata-rata dengan algoritma *ensemble* untuk mengatasi *incomplete* data serta evaluasi hasil berdasarkan ukuran akurasi dan waktu eksekusi. Pada penelitian ini menyajikan proses *incomplete* data dengan model *regression Imputation*, yang diharapkan mencapai nilai akurasi dan presisi yang relatif lebih baik.

Sebagian besar penelitian juga berusaha untuk mendapatkan metode yang tepat untuk meningkatkan nilai akurasi, salah satunya mengusulkan metode optimasi untuk beberapa dataset yang tersedia. Optimasi data medis dengan keakuratan hasil -melalui pendekatan optimasi *Genetic Algorithm* telah dilakukan pada penelitian Gorzałczany dan Rudzinski (2017), pada penelitian sejenis Wissanu Thungrut and Naruemon (2019) juga melakukan hal serupa yang mengoptimasikan dataset diabetes dengan metode *Genetic Algorithm*, hasil yang di dapat membuktikan bahwa metode optimasi *Genetic Algorithm* lebih unggul untuk pemecahan masalah sejenis, maka dari itu pada penelitian ini, penulis mengusulkan metode *Genetic Algorithm* pada proses optimasi.

Berdasarkan uraian yang telah dijelaskan diatas, penelitian kali ini dilakukan incomplete data dengan metode *regression imputation*,reduksi atribut melalui pendekatan *Optimum Index Factor* (OIF) serta optimasi menggunakan metode *Genetic Algorithm* dengan judul "**Optimasi *Genetic Algorithm* melalui penanganan Incomplete Data dan Reduksi (Studi Kasus : Dataset Faktor Risiko Kanker Serviks)**".

1.2.Masalah Penelitian

Berdasarkan latar belakang yang telah dijelaskan diatas, masalah yang ingin dicari solusinya dalam penelitian ini dibagi lagi ke dalam dua bagian, yaitu identifikasi masalah dan rumusan masalah.

1.2.1. Identifikasi Masalah

Adapun yang menjadi identifikasi masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana melengkapi data dari beberapa jumlah data yang hilang
2. Bagaimana mereduksi atribut tanpa mempengaruhi hasil
3. Bagaimana mengoptimasi data faktor risiko kanker serviks dengan tingkat akurasi hasil yang relatif lebih optimal

1.2.2. Rumusan Masalah

Berdasarkan identifikasi masalah yang ada, maka rumusan masalah pada penelitian ini adalah bagaimana melengkapi data yang *incomplete* dan mereduksi beberapa atribut tanpa mempengaruhi hasil serta pengoptimasian data yang diharapkan mencapai nilai akurasi presisi yang relatif lebih baik sehingga meningkatkan efisien dan efektivitas dalam pengolahan data kanker serviks.

1.3. Tujuan dan Manfaat Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Meningkatkan efisiensi data dalam pengolahan incomplete data

2. Meningkatkan efisiensi waktu dalam reduksi atribut tanpa mempengaruhi tingkat keakuratan hasil.
3. Meningkatkan efektivitas dalam pengoptimasian data sehingga mendapatkan hasil berupa prediksi yang lebih baik dalam pengolahan data kanker serviks.

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Membantu penyedia layanan kesehatan untuk menghasilkan alat bantu diagnosis dalam menganalisis prediksi berdasarkan optimasi kanker serviks yang cerdas.
2. Hasil penelitian ini dapat digunakan sebagai referensi untuk pengembangan penelitian lebih lanjut dalam bidang prediksi risiko kanker serviks dalam dunia medis.

1.4. Batasan Masalah

Adapun batasan dari penelitian ini adalah sebagai berikut:

1. Dataset yang digunakan adalah data Kanker Serviks yang diambil dari *repositori University of California di Irvine (UCI)*[1] dikumpulkan di 'Hospital Universitario de Caracas' di Caracas, Venezuela. Dataset terdiri dari informasi demografis, kebiasaan, dan catatan medis historis dari 858 sampel pasien 32 atribut serta 4 target. Beberapa pasien memutuskan untuk tidak menjawab beberapa pertanyaan karena masalah privasi (nilai yang hilang).
2. Metode regression imputation yang akan digunakan adalah model pengembangan regression dalam jurnal singh & Valdes (2009) berjudul *Optimal Method of Imputation in Survey Sampling*
3. Metode yang di gunakan untuk mereduksi atribut adalah *Optimum Index Factor* (OIF)
4. Arsitektur *machine learning* untuk pengoptimasian yang digunakan adalah *Genetic Algorithm*
5. *Tools* yang digunakan dalam penelitian ini adalah PYTHON 3.7

[1] https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 *Diacces 17/05/19 14:27*

1.5. Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi Literatur

   Pada tahap ini dilakukan pendalaman mengenai bagaimana konsep *Incomplete Data,* reduksi atribut, proses optimasi dan bagaimana cara menyelesaikan masalah tersebut serta mempelajari beberapa penelitian sebelumnya.

2. Tahap Analisis

   Pada tahap ini dilakukan proses untuk mengidentifikasi data yang dibutuhkan, masalah dan tantangan yang harus diselesaikan dan menjelaskan solusi yang diusulkan untuk menyelesaikan masalah dan tantangan yang ada.

3. Perancangan Sistem

   Perancangan sistem dimulai dengan membuat diagram dari setiap proses yang akan dilakukan dan menetukan *library* apa saja yang akan digunakan dalam membangun sistem.

4. Implementasi

   Pada tahap ini dilakukan implementasi algoritma sesuai dengan rancangan sistem yang telah dibuat sebelumnya menggunakan bahasa pemrograman yang telah ditentukan.

5. Pengujian

   a. Melakukan pengujian untuk nilai akurasi presisi dan recall yang lebih baik dengan menggunakan beberapa data testing.
   b. Melakukan pengujian data testing dengan metode yang telah di tentukan sehingga mendapatkan hasil akurasi terbaik.

6. Evaluasi

   Pada tahap ini, dilakukan evaluasi terhadap hasil pengujian yang sudah dilakukan untuk mengambil kesimpulan dan saran.

1.6. Sistematika Penulisan

Sistematika penulisan laporan penelitian ini terdiri dari 5 bab, dimana secara garis besar masing-masing bab membahas hal – hal berikut ini.

Bab 1 Pendahuluan, berisi penjelasan umum, masalah dan solusi yang akan dilakukan penelitian. Bab 2 berisi studi literatur dan tinjauan singkat terkait masalah dan metode yang berhubungan dengan penelitian yang akan dilakukan. Bab 3 Metodologi Penelitian, berisi identifikasi masalah, langkah-langkah dari metode yang diusulkan, data yang digunakan, alat-alat penelitian dan metode analisis. Bab 4 Hasil dan Pengujian, berisi hasil dari sistem yang dibangun dan analisis berdasarkan pengujian yang dilakukan. Bab 5 Kesimpulan dan Saran, berisi kesimpulan yang diperoleh dari hasil dan pengujian penelitian yang sudah dilakukan dan saran untuk hasil yang lebih baik dalam penelitian yang sejenis.