

BAB I

PENDAHULUAN

Bab ini membahas tentang latar belakang masalah untuk dilakukan penelitian, rumusan masalah yang ada dalam penelitian, batasan masalah dalam penelitian, tujuan dan manfaat dari penelitian ini, dan sistematika penulisan laporan penelitian.

1.1 Latar Belakang

Jumlah penderita diabetes secara global, sekitar 537 juta orang dewasa menderita diabetes, dan 6,7 juta kematian secara langsung dikaitkan dengan diabetes setiap tahun di seluruh dunia. Di wilayah Asia Tenggara berdasarkan survei yang dilakukan oleh *International Diabetes Federation* (IDF), sekitar 90 juta orang dewasa mengidap diabetes, dengan perkiraan jumlah ini akan meningkat hingga 152 juta pada tahun 2045, penderita diabetes di Indonesia sebanyak 19,5 juta orang dewasa dan menempati posisi ke 5 dengan penderita diabetes tertinggi pada tahun 2021.

Diabetes merupakan salah satu penyakit kronis yang semakin meningkat prevalensinya di seluruh dunia. Menurut data dari *World Health Organization* (WHO), diabetes melitus menjadi salah satu penyebab utama morbiditas dan mortalitas global. Untuk mengelola dan mencegah penyakit ini, penting untuk memiliki sistem deteksi yang akurat dan efisien. Pengolahan data medis yang melibatkan berbagai fitur atau variabel klinis dapat memberikan wawasan berharga dalam diagnosis dan penanganan diabetes. (*World Health Organization*, 2023).

Dalam konteks data besar, salah satu tantangan utama adalah menentukan fitur mana yang paling relevan untuk model prediksi. Fitur yang tidak relevan atau redundan dapat mengurangi performa model dan meningkatkan kompleksitas komputasi (Rong dkk., 2019). Oleh karena itu, teknik seleksi fitur atau *feature selection* sangat penting untuk meningkatkan kualitas model prediksi dengan mengurangi dimensi data dan menyaring fitur yang tidak penting. Seleksi fitur adalah proses memilih subset fitur yang paling relevan dari sekumpulan fitur yang ada. Penelitian yang dilakukan oleh (Zhang dan Yang, 2020) berhasil meningkatkan akurasi klasifikasi menggunakan Seleksi fitur dengan perbandingan model pertama

menggunakan semua fitur menghasilkan akurasi sebesar 88,21% dan model kedua menggunakan 10 fitur terbaik menghasilkan akurasi sebesar 89,03%.

Regresi logistik adalah salah satu algoritma klasifikasi yang sering digunakan untuk memprediksi probabilitas dari suatu kejadian biner, metode ini bekerja dengan memodelkan hubungan antara satu atau lebih variabel bebas dengan variabel target yang bersifat kategorikal. Dalam konteks analisis data medis, regresi logistik menawarkan keunggulan dalam interpretasi hasil dan kemampuannya menangani berbagai jenis data, termasuk data kategorikal dan kontinu. Hal ini menjadikannya metode yang sesuai untuk mengidentifikasi faktor-faktor yang mempengaruhi risiko diabetes berdasarkan data pasien (Hosmer dkk., 2013). Banyak penelitian yang membuktikan bahwa Regresi Logistik memiliki performa prediksi yang baik dalam klasifikasi dan regresi dalam bidang kesehatan, biologi, pemasaran, dan berbagai disiplin ilmu lainnya (Zhu dkk., 2019) berhasil melakukan klasifikasi pada penyakit diabetes menggunakan algoritma Regresi Logistik dengan hasil akurasi tertinggi 77% dibanding dengan algoritma lainnya.

Penelitian untuk mengimplementasikan teknik seleksi fitur menggunakan *Pearson Correlation* pada dataset diabetes dan mengevaluasi bagaimana pengaruh pemilihan fitur terhadap performa algoritma Regresi Logistik penting karena dataset diabetes, sering kali memiliki banyak variabel atau fitur yang tidak semuanya relevan untuk tujuan analisis. Variabel yang tidak relevan dapat mengaburkan hasil prediksi dan menyebabkan model menjadi kurang efisien atau akurat. Seleksi fitur yang tepat dapat meningkatkan kinerja model prediksi dengan mengurangi kompleksitas data tanpa mengorbankan akurasi.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, maka rumusan masalah dari penelitian ini yaitu seberapa besar kenaikan akurasi dari klasifikasi penyakit diabetes dengan model Regresi Logistik tanpa seleksi fitur dan dengan ditambahkan seleksi fitur (*Pearson Correlation*) pada proses klasifikasi.

1.3 Ruang Lingkup

Ruang lingkup digunakan sebagai acuan mengenai batasan dari pembuatan program pada penelitian ini, sehingga tidak menyimpang maupun melebihi batas pembahasan dari penelitian dan tetap sesuai dengan target yang ingin dicapai. Adapun ruang lingkup dari penulisan penelitian ini sebagai berikut:

1. Menggunakan data sekunder yang diambil berjumlah 390 data dan bersumber dari situs web kaggle.com diambil pada tanggal 7 April 2024.
2. Melakukan klasifikasi pada dataset diabetes dengan membuat program menggunakan metode Regresi Logistik.
3. Mengimplementasikan *Pearson Correlation* sebagai metode seleksi fitur pada Regresi Logistik untuk menentukan fitur-fitur yang paling relevan dengan target prediksi, yaitu kondisi diabetes.
4. Membandingkan hasil akurasi antara penggunaan seleksi fitur dengan tidak menggunakan seleksi fitur.
5. Membuat program menggunakan metode Regresi Logistik dengan bahasa pemrograman *Python 3*.
6. Melakukan pelatihan dan pengujian program yang dibuat menggunakan platform *Google Colaboratory* di lingkungan *Jupyter Notebook*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengetahui pengaruh seleksi fitur (*Pearson Correlation*) dengan model Regresi Logistik pada klasifikasi terhadap penyakit diabetes, serta mengetahui perbandingan akurasi antara penggunaan seleksi fitur dengan tidak menggunakan seleksi fitur.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah memberikan informasi tentang pengaruh seleksi fitur terhadap performa model klasifikasi dan mengurangi kompleksitas model dengan menggunakan fitur-fitur yang paling relevan, sehingga dapat menjadi referensi penelitian lain dengan tujuan untuk meningkatkan performa pemodelan.

1.6 Sistematika Penulisan

Tujuan dari sistematika penulisan adalah untuk memberikan gambaran tentang pembahasan yang ada didalam laporan penelitian secara urut dan jelas. Berikut adalah sistematika penulisan laporan penelitian ini:

BAB I PENDAHULUAN

Bab ini membahas tentang latar belakang masalah untuk dilakukan penelitian, rumusan masalah yang ada dalam penelitian, batasan masalah dalam penelitian, tujuan dan manfaat dari penelitian ini, dan sistematika penulisan laporan penelitian.

BAB II LANDASAN TEORI

Bab ini membahas tentang tinjauan pustaka tentang penelitian sebelumnya yang berkaitan dengan penelitian ini, dasar dari teori-teori yang akan digunakan dalam penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai metode serta langkah kerja yang dilakukan dalam melakukan klasifikasi penyakit diabetes menggunakan Regresi Logistik dengan implementasi seleksi fitur pada Regresi Logistik.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas hasil dari penelitian tentang implementasi seleksi fitur menggunakan *Pearson Correlation* pada algoritma Regresi Logistik dalam melakukan klasifikasi penyakit diabetes.

BAB V PENUTUP

Bab ini membahas tentang kesimpulan dari bab-bab yang telah dibahas sebelumnya dalam penelitian, serta saran-saran untuk pengembangan penelitian di masa mendatang.