BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Diabetes Melitus adalah salah satu penyakit yang disebabkan oleh hiper-glikemia atau kadar glukosa yang banyak dalam darah ditambah dengan adanya kelainan pada proses metabolisme akibat kurangnya insulin yang ditandai dengan kebiasaan sering minum, sering kencing, dan sering makan. (Prabowo 2015) memprediksi pada tahun 2030, hingga 21,3 juta orang Indonesia diperkirakan menderita diabetes. Diabetes di Indonesia masih menjadi masalah kesehatan masyarakat yang serius ditandai dengan pertumbuhan jumlah penduduk, penuaan, gaya hidup tidak sehat, kebiasaan makan yang tidak sehat, diet yang tidak sehat dan obesitas (Nasution 2021). Hal tersebut tidak disadari oleh orang-orang yang tergolong sibuk seperti pekerja kantoran atau staff pengajar yang rutin di depan layar komputer dan di dalam kelas memberikan perkuliahan, tentu kondisi tersebut memiliki tingkat resiko yang sangat tinggi terhadap gangguan metabolisme dan akhirnya menderita penyakit diabetes (Irwansyah 2021). Menurut (Nasution 2021) bahwa sebagian besar pekerja ringan dengan aktivitas fisik rendah memiliki kadar gula darah tinggi dan berisiko mengidap diabetes.

Saat ini, perkembangan *database* dalam bidang kesehatan sangat bertumbuh pesat. *Database* kesehatan banyak menyimpan data-data terkait dunia kesehatan seperti data diagnosa penyakit, data rekam medis pasien, data kandungan obat dan lainnya dimana data-data tersebut sangat penting untuk meningkatkan kualitas alat-alat medis (Vidyanto 2019). Bidang ilmu yang potensial mengolah data menjadi pengetahuan baru sering disebut dengan istilah *data mining*. Dalam bidang kesehatan, *data mining* dapat digunakan untuk memprediksi sejumlah penyakit yang salah satunya adalah Diabetes Melitus. *Data mining* merupakan proses untuk memperoleh informasi pada suatu database besar dan didalam prosesnya menggu- nakan penalaran statistik, matematika dan kecerdasan buatan (Prasetyowati 2017). Saat ini, algoritma *data mining* sering digunakan dalam banyak kasus tergantung pada ukuran dan kualitas data. Pada prosesnya *data mining* akan mengekstrak informasi dengan cara menganalisis adanya pola-pola

keterkaitan tertentu dari sejumlah besar data yang tersimpan dalam repositori dan data berkualitas rendah dapat menyebabkan performa yang rendah. Di antara algoritma *data mining* lainnya, pohon keputusan memiliki berbagai keunggulan diantaranya sederhana, mudah diterapkan, membutuhkan sedikit pengetahuan, mampu menangani data bersifat kontinu, diskrit dan kategorikal, serta mampu menangani data yang sangat besar (Han 2012).

Beberapa algoritma yang digunakan membuat pohon keputusan antara lain ID3 (Quinlan,1986), ID5 (Utgoff,1989), C4.5 (Quinlan,1993) dan CART (Breiman, Friedman, Olshen dan Stone,1984). Algoritma C4.5 dan *Classification and Regression Tree* (CART) adalah bentuk terbarukan dari algoritma *Iterative Dichotomiser 3* (ID3) dimana CART dikembangkan oleh ahli statistik, sedangkan C4.5 dikembangkan oleh ilmuwan komputer di bidang *machine learning*. Pemilihan algoritma C4.5 pada penelitian ini didasarkan karena algoritma memuat proses kriteria *split* dari ID3 yang dimodifikasi yang disebut *Gain Ratio* (Santhosh 2013). Kriteria *split* terbagi dari serangkaian *node* pilihan, dihubungkan melalui cabang, bergerak menurun kebawah dari simpul akar sampai berakhir di simpul daun. Menurut (Patrick 2020), Algoritma C4.5 merupakan pengembangan dari Algoritma ID3 yang bisa menangani nilai hilang, mengatasi data dengan tipe kontinu dan melakukan pemangkasan pohon (*pruning tree*) menghasilkan komposisi praktis dan sederhana yang berpengaruh pada pembentukan pohon keputusan yang lebih akurat.

Penilitian telah dilakukan (Prasannavenkatesan 2015) menggunakan berbagai algoritma dalam mendiagnosa Diabetes Melitus diantaranya, *Decision Tree* (akurasi 66.89%), *Naive Bayes* (akurasi 77.24%), *Extra Trees* (akurasi 72.41%), *k-Nearest Neighbor* (akurasi 71.72%), *Radial Basis Function* (akurasi 69.27%), *Multi Layer Perceptron* (akurasi 80.68%). Penelitian lain oleh (Arnita 2019) dengan menggunakan metode Neural Network dan algoritma Learning Vector Quantization menghasilkan akurasi sebesar 90%. Hasil penelitian-penelitian tersebut memperlihatkan bahwa algoritma pohon keputusan memberi tingkat akurasi lebih rendah dibanding algoritma lainnya dengan perbedaan yang tidak signifikan. Dalam pendekatan pohon keputusan (*decision tree*),

pemangkasan pohon merupakan proses menghilangkan atau memotong beberapa cabang (*node*) yang tidak diperlukan. *Node* yang tidak diperlukan dapat memicu adanya data noise dan fitur yang kurang relevan (Bahzad 2021). Noise adalah data yang berisi nilai-nilai yang salah atau anomali, yang umumnya disebut outlier. Kemungkinan lain penyebab dari noise yang harus diperhatikan adalah pengukuran data, perekaman dan transmisi data, yang mengakibatkan anomali dan ketidakakuratan (Vercellis 2009). Data yang mengandung noise tinggi dapat dilihat pada sebaran data *varians* yang tidak merata atau *heterogen*. Dalam *data mining* noise yang terdapat pada data yang memiliki banyak kelas atau multi kelas dapat mengurangi akurasi pada klasifikasinya (Al Riza 2015).

Masalah data noise pada data berdimensi tinggi akan diselesaikan dengan salah satu metode statistika dengan menerapkan Analisis Komponen Utama (PCA) yang berfungsi untuk mereduksi dimensi atau fitur data dengan tetap menjaga karakteristik penting dalam data sebelum diproses oleh algoritma pengklasifikasi (Bailey 2012). Reduksi dimensi merupakan metode dimana komponen yang paling penting dipilih sehingga komputasi dapat menghasilkan hasil yang lebih optimal dan dapat menjelaskan sebanyak mungkin variasi dalam data. Analisis komponen utama adalah salah satu metode statistik multivariat yang secara *linear* mengubah bentuk sekumpulan variabel asli menjadi beberapa variabel yang lebih kecil dan tidak berkorelasi dan mampu mewakili informasi dari variabel-variabel asli (Astuti 2018). Semakin banyak data yang ada maka hasil pengujian pada data akan semakin sesuai dengan prediksi hasil klasifikasi. Namun, jika data yang tersedia sangat sedikit maka pengujian data dengan prediksi hasil klasifikasi akan kurang akurat. Tujuan utamanya adalah untuk menghilangkan fitur yang tidak signifikan dengan memilih fitur-fitur yang aplikatif dan tidak berkorelasi tanpa mengubah data yang terlampir dalam informasi yang sebenarnya (Sumaiya 2016).

Penelitian terdahulu tentang reduksi dimensi yaitu oleh Dang dengan judul model klasifikasi penyakit tumor menggunakan Analisis Komponen Utama sebagai reduksi fitur dan algoritma ID3 untuk menyeleksi fitur pada tahun 2016. Hasil seleksi fitur diklasifikasikan menggunakan model *Multi Layer Perceptron*

(MLP). Data yang diuji adalah data DNA dari *Leukemia, Prostate and Diffuse large B-cell lymphoma* (DLBCL). Hasil uji reduksi fitur tersebut, mengkombinasikan ekstraksi dengan seleksi fitur serta MLP sebagai model klasifikasi terbukti meningkatkan keakuratan model dari dataset DNA (Astuti 2018).

Berdasarkan uraian latar belakang yang telah diuraikan diatas maka penulis tertarik untuk mengambil judul Penerapan Analisis Komponen Utama Untuk Meningkatkan Akurasi Klasifikasi pada Algoritma Decision Tree C4.5 dalam Mendiagnosa Penyakit Diabetes Melitus.

1.2 Ruang Lingkup

Ruang lingkup bahasan pada penelitian ini yaitu data yang akan diteliti mengalami peningkatan akurasi pengklasifikasian algoritma C4.5 dengan metode Analisis Komponen Utama agar memberikan hasil prediksi diagnosa penderita Diabetes Melitus yang lebih akurat.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah yang akan dibahas adalah:

1. Bagaimana penggunaan metode Analisis Komponen Utama untuk mengatasi masalah data noise yang terproses pada model klasifikasi algoritma C4.5?
2. Bagaimana perbandingan hasil klasifikasi data Diabetes Melitus menggunakan algoritma C4.5 sebelum dan sesudah direduksi ?

1.4 Batasan Masalah

Dalam penyusunan penelitian ini diperlukan batasan masalah untuk menghindari meluasnya permasalahan yang akan dibahas, yaitu:

1. Masalah yang ingin diatasi yaitu data noise yang terjadi pada algoritma C4.5 dengan asumsi bahwa data telah memenuhi asumsi klasik yang lain.
2. Metode yang digunakan dalam mengatasi masalah data noise adalah

metode Analisis Komponen Utama.

3. Data yang digunakan yaitu data sekunder.
4. Tingkat ketepatan klasifikasi dibatasi oleh nilai akurasi, presisi dan sensitivitas.

1.5 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan metode Analisis Komponen Utama untuk mengatasi masalah data noise yang terjadi pada model klasifikasi algoritma C4.5.
2. Mendapatkan perbandingan akurasi hasil klasifikasi data Diabetes Melitus menggunakan algoritma C4.5 sebelum dan sesudah direduksi.

1.6 Manfaat Penelitian

Adapun manfaat yang diharapkan pada penelitian ini yaitu sebagai berikut:

1. Bagi penulis, penelitian ini berguna untuk menambah ilmu pengetahuan mengenai informasi mendalam terhadap metode seleksi fitur dataset menggunakan Analisis Komponen Utama dapat meningkatkan akurasi model klasifikasi algoritma pohon keputusan C4.5.
2. Bagi pembaca, penelitian ini berguna untuk menambah pengetahuan mengenai hasil perbandingan dari kinerja model klasifikasi algoritma pohon keputusan C4.5 sebelum dan sesudah digunakan Analisis Komponen Utama.
3. Bagi akademik, dapat digunakan sebagai tambahan informasi dan sumber bacaan mengenai metode Analisis Komponen Utama bagi yang hendak melakukan penelitian serupa dan penelitian selanjutnya.