BAB I
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan teknologi, kebutuhan akan informasi secara cepat dan tepat menjadi hal yang sangat penting saat ini dalam mengambil keputusan. Pentingnya melakukan prediksi terhadap sebuah kasus akan memberikan beberapa peluang dalam memperbaiki situasi dan kondisi ataupun kemungkinan-kemungkinan kejadian yang tidak diinginkan terjadi dimasa yang akan mendatang. Hal ini sudah banyak diterapkan dalam segala jens bidang ilmu seperti ilmu managemen perekonomian, ilmu dalam bidang bisnis, ilmu geologi, geografi, pertanian, hingga ilmu dalam dunia medis. Salah satunya dalam dunia medis sangat bermanfaat bagi banyak kalangan, karena penerapan ini sangat banyak digunakan terhadap penyakit-penyakit serius sehingga dapat memperlambat bahkan menghilangkan penyakit yang telah didiagnosa sebelumnya (Ma, 2020).

Kinerja machine learning dalam perkembangan teknologi informasi sangat membantu dalam dunia medis, dimana kalangan dokter dan tim medis lainnya mendapatkan kesempatan dalam meningkatkan kualitas kerja tim medis. Salah satu jenis penyakit berbahaya yang dapat diprediksi merupakan penyakit diabetes dimana penyakit ini merupakan jenis penyakit yang banyak menyebabkan kelumpuhan, kerusakan yang mengerikan pada penderita lanjutan hingga kematian. Hal tersebut terjadi karena keterlambatan diagnosis. Sehingga pentingnya dilakukan diagnosis untuk membantu proses medis yang lebih komplek lagi dan merupakan proses yang penting dalam sebuah peneltian dibidang medis. Permasalahan yang sangat sering dihadapi dan sering terjadi dalam hal diagnosis waktu yang sangat lama dalam proses prediksi terhadap gejala dan kurangnya tingkat akurasi dalam proses klasifikasi (Lv and Qiao, 2020).

Dalam proses penerapan sesuatu maka dibutuhkan machine learning dan machine learning tidak dapat beroperasi tanpa algoritma, seleksi fitur merupakan salah satu komponen yang sangat penting dalam menentukan akurasi dalam proses menjalankan algoritma klasifikasi sehingga pentingnya mengetahui atribut utama pada sebuah penyakit. Sering didapatkan hasil yang berbeda ketika mendiagnosis penyakit diabetes awal, sehingga dipilih fitur atribut penting yang didasarkan pada efisiensi diagnosis (California, Irvine and Serikat, 2021). Keputusan diagnosis penyakit yang berbeda merupakan salah satu tantangan yang paling penting untuk analisis data diabetes tahap awal. Umumnya keputusan diagnosis didasarkan kepada pemeriksaan kepada pasien dan juga pengalaman dokter. Pada pemeriksaan yang dilakukan tidak semua kondisi atribut terpenuhi, tetapi tetap harus dilakukan proses diagnosis (Hendro, Adji and Setiawan, 2012).

Kesalahan pada proses diagnosis akan mempengaruhi hasil pada pasien ataupun dokter, jadi diagnosis merupakan tugas kompleks yang membutuhkan keahlian tinggi dan pengalaman. Sistem komputer dapat membantu para dokter sebagai alat memprediksi dan diagnosis penyakit diabetes, penelitian medis selalu berususan dengan data yang besar, penanganan data yang besar dengan tidak benar dapat mempengaruhi keakuratan terhadap hasil. Dataset yang digunakan merupakan hasil dari informasi-informasi yang disediakan oleh situs yang akurat pada bidangnya (Albahri *et al.*, 2020). Pemilihan fitur yang tepat dapat mempersingkat waktu dan menghemat biaya dalam melakukan diagnosis salah satu proses pemilihan fitur dapat dilakukan reduksi atribut yang digunakan untuk mengurangi data dan menghapus atribut yang tidak relevan, berlebihan, dalam sebuah dataset. Pengurangan atribut telah menjadi langkah penting dalam pengenalan pola dan tugas Machine Learning (C, 2020). Hal ini menginspirasi penulis sehingga menambahkan proses reduksi atribut di dalam pengambilan keputusan pada proses klasifikasi dalam meningkatkan diagnosis penyakit diabetes tahap awal.

Seleksi fitur dilakukan untuk mengidentifikasi *subset* atribut yang optimal (Thaseen, Kumar and Ahmad, 2019). Seleksi fitur dapat mengurangi dimensi data dan waktu pelatihan (AlShboul *et al.*, 2018). (Thakkar and Lohiya, 2021)

Menjelaskan bahwa mengikutsertakan seleksi fitur dapat memberikan dua keuntungan yaitu dapat membantu mengurangi *curse of dimensionality* dan mengkomputasikan fitur yang penting dapat membantu interpretasi data. Contoh metode untuk seleksi fitur seperti, *Principal Componen Analysis* (PCA), *Chi-square, Wrapper based feature selection* dan *filter*. dimana berinteraksi dengan pengklasifikasi untuk mengevaluasi kegunaan fitur sehingga menghasilkan *subset* fitur yang lebih baik. Metode ini lebih lambat dibandingkan dengan metode *filter based feature selection, Neural Network* (Islam Ayon and Milon Islam, 2019) (Balasaraswathi, Sugumaran and Hamid, 2017). Principal component analysis (PCA) adalah teknik untuk mengurangi dimensi atribut dari dataset, meningkatkan interpretabilitas tetapi pada saat yang sama meminimalkan hilangnya informasi, banyak penelitian dan buku yang telah ditulis dan juga bahkan ada buku tentang varian PCA untuk tipe data tertentu (Gárate-Escamila, Hajjam El Hassani and Andrès, 2020). PCA dapat digabungkan dengan metode lain yang akurasinya dikenal keakurasiannya cukup tinggi, dimana PCA berfungsi untuk melakukan reduksi dimensionalitas dan dengan pemilihan fitur pada PCA yang hanya memilih nilai eigen yang paling besar maka ketika digabungkan dengan metode lain akan menaikan tingkat akurasi klasifikasi dan terutama sekali mempercepat hasil pengenalannya (Lei *et al.*, 2019)

Pada penemuan tingkat akurasi dari proses klasifikasi terhadap penyakit diabetes, penelitian terdahulu melakukan pengujian terhadap diabetes yang memiliki banyak perkembangan variasi sehingga diilakukan ekstraksi terhadap pemiihan data set meggunakan metode beberapa metode klasifikasi secara bersamaan seperti Decision tree, Naïve Bayes, K- Nearest Neighbor memiliki tingkat akurasi terhadap ekstraksi fitur data set sebesar 75,65% (Azrar *et al.*, 2018). Dan pada penelitian lainnya, memperlihatkan metode decition tree dalam menghasilkan nilai akurasi yang lebih tinggi terhadap pemilihan fitur dengan menggeneralisasi fitur yang optimal dari dataset, sehingga dapat meningkatkan akurasi klasifikasi dengan pencapaian 98,00% (Sneha and Gangil, 2019). Pernyataan beberapa penelitian terdahulu menyebutkan dalam proses klasifikasi menggunakan random forest lebih memiliki tingkat akurasi yang tinggi daripada

algoritma klasifikasi naïve bayes, decision tree, K-nearest neighbor (Xaverius et al., 2020).

Metode *Radom forest* telah membuktikan keberhasilannya dalam prediksi penyakit diabetes tahap awal dengan melakukan perbandingan tiga algoritma klasifikasi machine learning yaitu Suport Vector Machine, Naive Bayes dan Random Forest digunakan dalam percobaan ini untuk mendeteksi diabetes secara dini. Hasil penelitian terlihat masalah regresi dan klasifikasi data dan merupakan salah satu algoritma machine learning terbaik adalah algoritma *random forest* yang digunakan di berbagai bidang dengan pencapaian dengan tingkat akurasi 97,88 %, dan tool yang digunakan adalah WEKA (Apriliah *et al.*, 2021). Berdasarkan dari penjelasan permasalahan di atas maka penulis bermaksud untuk melakukan penelitian dalam mengoptimalkan kinerja data mining klasifikasai Random Forest dalam memprediksi penyakit diabetes tahap awal dimana akan berfokus kepada peningkatan akurasi dan ketepatan waktu dengan cara penggunaan ensamble random forest digunakan untuk proses enlarge dataset, sedangkan Principal Component Analysis (PCA) digunakan untuk seleksi fitur. Maka penelitian ini diberi judul "**Peningkatan Kinerja Random Forest Melalui Seleksi Fitur Secara PCA Untuk Mendeteksi Penyakit Diabetes Tahap Awal".**

1.2 Masalah Penelitian

Pada penelitian ini dalam membahas masalah penelitian dilakukan dengan dua bagian yaitu identifikasi masalah yang dilakukan untuk mendefinisikan masalah dan bagian selanjutnya adalah menentukan rumusan masalah.

1.2.1 Identifikasi Masalah

Pada penelitian ini berdasarkan latar belakang masalah ditemukan identifikasi masalah sebagai berikut ini:

1. Terlalu banyak Atribut yang dimiliki oleh dataset sehingga sulit mengetahui atribu-atibut utama untuk meningkatkan hasil prediksi penyakit diabetes tahapan awal.

2. Waktu yang dibutuhkan lebih lama jika pada dataset memiliki atribut atau record yang terlalu banyak.

1.2.2 Rumusan Masalah

Berdasarkan dengan latar belakang masalah yang sudah dijelaskan di atas, maka yang menjadi rumusan masalah dalam penelitian ini adalah pada proses diagnosis yang dilakukan sering pertanyaan atau atribut tidak terpenuhi secara keseluruhan. Sehingga perlu dilakukan reduksi atribut pada dataset penyakit diabetes tahapan awal dan melakukan peningkatan akurasi terhadap atribut dalam seleksi fitur pada dataset penyakit diabetes awal dengan cara penggunaan ensamble random forest digunakan untuk proses enlarge dataset, sedangkan Principal Component Analysis (PCA) digunakan untuk seleksi fitur.

1.3 Tujuan dan Manfaat Penelitian

Tujuan dan manfaat penelitian sangat penting diketahui dalam sebuah penelitian agar pembaca paham dan mengerti terhadap setiap rencana dan pencapaian yang diharapkan oleh penulis

1.3.1 Tujuan Penelitian

Berdasarkan rumusan yang telah diuraikan maka tujuan yang ingin dicapai pada penelitian ini yakni menemukan atribut utama yang paling berpengaruh pada proses diagnosis dalam pengambilan keputusan dengan melakukan seleksi fitur pada dataset penyakit diabetes awal dengan cara penggunaan ensamble random forest digunakan untuk proses enlarge dataset, sedangkan Principal Component Analysis (PCA) digunakan untuk seleksi fitur.

1.3.2 Manfaat Penelitian

Manfaat penelitian ini sebagai berikut:

Adapun manfaat yang diharapkan akan didapatkan dari penelitian ini adalah sebagai berikut :

1. Dengan melakukan reduksi dan seleksi fitur pada atribut dataset dapat mempersingkat waktu proses dan menemukan atribut-atribut utama dalam proses pengambilan keputusan/kalsifikasi pada dataset.
2. Sebagai bahan masukan bagi peneliti lainnya tentang reduksi atribut yang dilakukan pada dataset.

1.4 Batasan Masalah

Adapun batasan pembahasan masalah pada tesis ini sebagai berikut:

1. Penelitian ini menggunakan dataset penyakit diabetes yang berasal dari kaggle.com. Dataset tersebut dapat diakses pada *link* berikut ini.

   https://www.kaggle.com/datasets/mathchi/diabetes-data-set
2. Untuk melakukan pengujian pada penelitian ini menggunakan aplikasi tools rapidminer 9.8.1.
3. Penelitian ini dilakukan dengan membandingkan hasil random forest dengan PCA dan tanpa PCA dari hasil masing-masing tingkat akurasi.
4. Evaluasi dapat dilakukan berdasarkan *confusion matrix* untuk penilaian *accuracy, precision, recall/sensitivity/TPR, selectivity/specificity/TNR,* dan *f1 score*.

1.5 Metodologi Penelitian

Pada penyusunan tesis ini penulis melakukan beberapa tahapan dalam penerapan metodologi untuk menyelesaikan masalah. Berikut merupakan langkah-langkah penyelesaian ini adalah:

1. Studi Literatur

   Dalam tahapan ini dilakukan pengumpulan data sebagai bahan referensi yang berhubungan dengan penelitian yang dilakukan, seperti: penyakit Diabetes Tahapan Awal, dataset penyakit diabetes, algoritma Random forest dan Principal Component Analysis (PCA)
2. Analisis Masalah

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. Preprosesing data

   Pada tahapan preprosesing data ini digunakan untuk melakukan penyesuaian data terhadap algoritma ataupun metode yang akan digunakan, pada penelitian tahap *preprocessing* akan dilakukan dalam dua cara yaitu *cleaning* dan *binning* menggunakan IBM IBM SPSS Modeler 18.0, sehingga hasilnya dapat diproses untuk pengujian selanjutnya.

4. Tahap Pengujian

   Melakukan pengujian terhadap seleksi fitur penyakit diabetes tahap awal dengan menggunakan metode PCA. Kemudian menerapkan algoritma *Random Forest* dalam proses prediksi dengan dan tanpa seleksi fitur menggunakan RapidMiner 9.8.1. hasil sesuai pada tahapan analisa dengan melihat nilai dari *confusion matrix*.

5. Evaluasi

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

1.6 Sistematika Penulisan

Penulisan tesis ini ada beberapa tahapan pada penyelesaian penelitian dimana sistematika penulisan terdiri dari beberapa bagian sub-sub bab yang bertujuan agar mempermudah dalam memperkenalkan dan menjelaskan isi dan pembahasan yang dilaporkan, berikut merupakan sistematika penulisan yang dilakukan dalam pembuatan laporan tesis:

1. BAB I Pendahuluan

   Pada bab I pendahuluan berisikan tentang latar belakang terhadap pengambilan judul, menjelaskan tentang permasalahan penelitian, tujuan dan manfaat penelitian, batasan masalah, sistematika penulisan, daftar symbol/ istilah (jika ada)

2. BAB II Kajian Literatur

Pada bab ini menjelasakan tentang tinjauan terhadap penelitian mulai dari tinjauan pustaka, objek penelitian dan pola konsep pemikiran terhadap pemecahan masalah dan hipotesis penelitian.

3. BAB III Metodologi Penelitan

   Pada bab III metodologi penelitian menjelaskan tentang analisis masalah, pemilihan metode dan *sampling*, metode pengumpulan data, alat-alat penelitian teknik analisis dan pengujian data

4. BAB IV Hasil dan Pembahasan

   Pada bab ini merupakan penelitian yang membahas tentang hasil penelitian dan dilakukan pembahasan terhadap hasil penelitian

5. BAB V Penutup

   Bab V merupakan bab terakhir berisikan kesimpulan dan saran.