

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus merupakan salah satu penyakit kronis yang menjadi masalah kesehatan global karena jumlah penderitanya yang terus meningkat dari tahun ke tahun. Penyakit ini terjadi ketika pankreas tidak mampu memproduksi insulin dalam jumlah yang cukup atau ketika tubuh tidak dapat menggunakan insulin yang dihasilkan secara efektif. Kondisi tersebut menyebabkan peningkatan kadar glukosa dalam darah yang dalam jangka panjang dapat menimbulkan berbagai komplikasi serius, seperti penyakit jantung, kerusakan ginjal, gangguan penglihatan, hingga kerusakan saraf. Menurut data Organisasi Kesehatan Dunia (WHO), lebih dari 422 juta orang di seluruh dunia hidup dengan diabetes, dan sekitar 1,5 juta kematian setiap tahunnya secara langsung disebabkan oleh penyakit ini. Angka tersebut diperkirakan akan terus meningkat seiring dengan perubahan pola hidup masyarakat, meningkatnya angka obesitas, kurangnya aktivitas fisik, serta bertambahnya jumlah penduduk usia lanjut [1].

Tingginya prevalensi diabetes menunjukkan pentingnya upaya deteksi dini guna mengurangi risiko komplikasi dan meningkatkan kualitas hidup penderita. Deteksi dini memungkinkan tenaga kesehatan untuk memberikan penanganan yang lebih cepat dan tepat sebelum kondisi pasien berkembang menjadi lebih serius. Dalam era transformasi digital saat ini, pemanfaatan teknologi informasi di bidang kesehatan menjadi salah satu solusi yang efektif untuk meningkatkan kualitas layanan medis. Salah satu teknologi yang berkembang pesat adalah *machine learning*, yaitu cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan sistem komputer mempelajari pola dari data historis dan menghasilkan prediksi atau keputusan tanpa harus diprogram secara eksplisit. *Machine learning* telah banyak diterapkan dalam bidang kesehatan, seperti klasifikasi penyakit, prediksi risiko kesehatan, analisis data pasien, hingga pengolahan citra medis, dengan tingkat akurasi yang semakin baik [2].

Salah satu algoritma *machine learning* yang banyak digunakan untuk tugas klasifikasi adalah *Random Forest*. Algoritma ini merupakan metode *ensemble*

learning yang bekerja dengan menggabungkan sejumlah pohon keputusan (*Decision Tree*) untuk menghasilkan prediksi yang lebih akurat dan stabil. Setiap pohon keputusan dibangun menggunakan subset data dan subset fitur yang dipilih secara acak, kemudian hasil prediksi dari seluruh pohon digabungkan melalui mekanisme voting untuk menentukan kelas akhir. Pendekatan tersebut membuat *Random Forest* memiliki kemampuan yang baik dalam mengurangi risiko overfitting, meningkatkan kemampuan generalisasi model, serta mampu menangani dataset dengan jumlah fitur yang banyak dan kompleks. Selain itu, algoritma ini juga dikenal cukup tangguh terhadap data yang mengandung noise maupun data yang tidak lengkap [3].

Dalam penelitian ini digunakan dataset diabetes yang diperoleh dari platform *Kaggle* dengan jumlah 2.768 data pasien. Dataset tersebut dipilih karena memiliki jumlah data yang cukup besar dan memuat berbagai atribut kesehatan yang relevan dalam proses identifikasi risiko diabetes. Beberapa atribut yang terdapat dalam dataset meliputi usia (*Age*), jenis kelamin (*Gender*), kadar glukosa darah (*Blood Glucose Levels*), tekanan darah (*Blood Pressure*), indeks massa tubuh atau *Body Mass Index* (BMI), riwayat merokok (*Smoking History*), kadar hemoglobin terglikasi (HbA1c), serta berbagai indikator kesehatan lainnya yang diketahui memiliki hubungan dengan risiko terjadinya diabetes. Atribut-atribut tersebut merepresentasikan kondisi fisiologis dan gaya hidup individu yang dapat digunakan sebagai dasar dalam menentukan kemungkinan seseorang menderita diabetes.

Penggunaan dataset dengan berbagai atribut kesehatan memberikan keuntungan dalam proses pembangunan model *machine learning* karena memungkinkan algoritma untuk mempelajari hubungan yang kompleks antarvariabel. Dalam dunia medis, faktor penyebab diabetes tidak hanya dipengaruhi oleh satu variabel saja, melainkan merupakan hasil interaksi dari berbagai faktor biologis, genetik, dan gaya hidup. Sebagai contoh, kadar glukosa darah yang tinggi dapat menjadi indikator utama diabetes, namun faktor lain seperti usia, obesitas, tekanan darah tinggi, serta riwayat kesehatan juga memiliki peran penting dalam meningkatkan risiko penyakit tersebut. Oleh karena itu, keberagaman atribut yang tersedia dalam *dataset* menjadi sumber informasi yang

sangat berharga bagi model *machine learning* untuk menemukan pola-pola tersembunyi yang sulit diidentifikasi melalui analisis manual.

Sebelum proses pemodelan dilakukan, data yang digunakan perlu melalui tahapan prapemrosesan (*preprocessing*) agar kualitas data menjadi lebih baik dan sesuai untuk digunakan dalam proses pembelajaran mesin. Tahapan ini meliputi pemeriksaan data yang hilang (*missing values*), penghapusan data duplikat, transformasi data kategorikal menjadi bentuk numerik, serta normalisasi atau standarisasi data apabila diperlukan. Proses *preprocessing* memiliki peran yang sangat penting karena kualitas data yang digunakan akan sangat memengaruhi performa model yang dihasilkan. Data yang bersih dan terstruktur memungkinkan algoritma machine learning untuk mempelajari pola dengan lebih efektif sehingga hasil prediksi yang diperoleh menjadi lebih akurat dan dapat diandalkan.

Penelitian ini menggunakan algoritma *Random Forest* sebagai metode utama untuk melakukan klasifikasi tingkat risiko diabetes. Pemilihan algoritma *Random Forest* didasarkan pada berbagai keunggulan yang dimilikinya dibandingkan metode klasifikasi lainnya. *Random Forest* merupakan algoritma berbasis *ensemble learning* yang menggabungkan banyak pohon keputusan (*Decision Tree*) untuk menghasilkan prediksi yang lebih stabil dan akurat. Setiap pohon keputusan dibangun menggunakan subset data yang berbeda melalui teknik *bootstrap sampling*, sedangkan pemilihan atribut pada setiap node dilakukan secara acak. Pendekatan tersebut memungkinkan setiap pohon menghasilkan perspektif yang berbeda terhadap data sehingga ketika hasil seluruh pohon digabungkan melalui mekanisme voting, model yang dihasilkan memiliki kemampuan generalisasi yang lebih baik.

Selain memiliki tingkat akurasi yang tinggi, *Random Forest* juga dikenal mampu mengatasi masalah *overfitting* yang sering terjadi pada algoritma *Decision Tree* tunggal. *Overfitting* merupakan kondisi ketika model terlalu mempelajari pola pada data pelatihan sehingga performanya menurun ketika digunakan untuk memprediksi data baru. Dengan mengombinasikan banyak pohon keputusan, *Random Forest* dapat mengurangi risiko tersebut dan menghasilkan model yang lebih robust. Algoritma ini juga memiliki kemampuan yang baik dalam menangani

dataset dengan jumlah atribut yang banyak, data yang mengandung noise, serta hubungan nonlinier antarvariabel yang sering ditemukan pada data kesehatan.

Penelitian ini tidak hanya berfokus pada penerapan algoritma *Random Forest* dalam klasifikasi diabetes, tetapi juga mengkaji pengaruh variasi proporsi pembagian data latih (training data) dan data uji (testing data) terhadap performa model yang dihasilkan. Pembagian data merupakan salah satu tahapan penting dalam pengembangan model *machine learning* karena menentukan seberapa banyak data yang digunakan untuk proses pembelajaran dan seberapa banyak data yang digunakan untuk proses evaluasi. Pemilihan proporsi pembagian data yang kurang tepat dapat menyebabkan model tidak mampu mempelajari pola data secara optimal atau menghasilkan evaluasi yang kurang representatif terhadap kondisi sebenarnya.

Dalam penelitian ini digunakan tiga skema pembagian data yang berbeda, yaitu 70:30, 80:20, dan 90:10. Pada skema 70:30, sebanyak 70% data digunakan sebagai data latih dan 30% sisanya digunakan sebagai data uji. Skema ini memberikan jumlah data evaluasi yang relatif besar sehingga mampu memberikan gambaran yang lebih representatif mengenai kemampuan model dalam melakukan prediksi pada data baru. Pada skema 80:20, sebanyak 80% data digunakan untuk pelatihan dan 20% digunakan untuk pengujian. Skema ini merupakan salah satu metode yang paling sering digunakan dalam penelitian machine learning karena dianggap mampu memberikan keseimbangan antara proses pembelajaran dan evaluasi model. Sementara itu, pada skema 90:10, sebagian besar data digunakan untuk proses pelatihan sehingga model memiliki kesempatan lebih besar untuk mempelajari pola-pola yang terdapat dalam dataset. Namun, jumlah data pengujian yang lebih sedikit dapat memengaruhi tingkat representativitas hasil evaluasi yang diperoleh.

Perbandingan ketiga skema pembagian data tersebut dilakukan untuk mengetahui bagaimana perubahan jumlah data pelatihan dan data pengujian memengaruhi performa algoritma *Random Forest*. Secara teoritis, semakin banyak data yang digunakan untuk pelatihan, semakin banyak informasi yang dapat dipelajari oleh model. Akan tetapi, penggunaan data pelatihan yang terlalu besar

juga dapat mengurangi jumlah data yang tersedia untuk pengujian sehingga hasil evaluasi mungkin kurang mencerminkan kemampuan model secara keseluruhan. Sebaliknya, jika jumlah data pengujian terlalu besar, maka jumlah data pelatihan menjadi lebih sedikit sehingga model mungkin tidak dapat mempelajari pola data secara optimal. Oleh karena itu, diperlukan analisis yang lebih mendalam untuk menentukan proporsi pembagian data yang paling sesuai dalam kasus klasifikasi diabetes.

Untuk mengukur performa model, penelitian ini menggunakan metrik *accuracy* sebagai indikator utama evaluasi. *Accuracy* menunjukkan persentase jumlah prediksi yang benar dibandingkan dengan keseluruhan data yang diuji. Nilai *accuracy* yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan data pasien ke dalam kategori yang sesuai. Selain *accuracy*, hasil penelitian juga dapat memberikan gambaran mengenai konsistensi performa *Random Forest* pada berbagai skenario pembagian data sehingga dapat diketahui konfigurasi yang paling efektif untuk menghasilkan model klasifikasi diabetes yang optimal.

Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan (*Decision Support System*) berbasis *machine learning* untuk membantu proses deteksi dini diabetes. Sistem yang mampu melakukan prediksi risiko diabetes secara cepat dan akurat dapat membantu tenaga medis dalam melakukan *skrining* awal terhadap pasien sehingga tindakan pencegahan maupun penanganan dapat dilakukan lebih dini. Selain itu, penelitian ini juga diharapkan dapat memberikan manfaat bagi institusi kesehatan dalam memanfaatkan data medis yang tersedia sebagai sumber informasi untuk mendukung pengambilan keputusan yang lebih efektif.

Lebih lanjut, penelitian ini diharapkan dapat menjadi referensi bagi peneliti selanjutnya yang ingin mengembangkan model klasifikasi diabetes menggunakan algoritma *Random Forest* maupun algoritma *machine learning* lainnya. Hasil analisis mengenai pengaruh variasi pembagian data latih dan data uji dapat digunakan sebagai acuan dalam menentukan strategi pembagian data yang tepat untuk memperoleh performa model yang optimal. Dengan demikian, penelitian ini

tidak hanya berkontribusi pada peningkatan akurasi model prediksi diabetes, tetapi juga memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang memengaruhi kinerja model machine learning dalam bidang kesehatan. Pada akhirnya, pemanfaatan teknologi *machine learning* dalam deteksi dini diabetes diharapkan mampu mendukung upaya pencegahan, meningkatkan kualitas layanan kesehatan, serta mengurangi angka komplikasi dan kematian akibat penyakit diabetes mellitus.

1.2 Rumusan Masalah

1.2.1 Metode Usulan

Berdasarkan latar belakang yang sudah dijelaskan sebelumnya, masalah utama dalam penelitian ini adalah bagaimana membuat model klasifikasi yang efektif dan akurat untuk mendeteksi risiko diabetes menggunakan algoritma *Random Forest*. Selain itu, terdapat beberapa cara membagi data latih dan data uji yang sering digunakan dalam proses pelatihan model *machine learning*, tetapi tidak semua cara tersebut memberikan hasil performa yang terbaik.

Perbedaan cara membagi data bisa memengaruhi tingkat *accuracy* dan kemampuan model dalam menggeneralisasi hasil. Karena itu, penting untuk melanjutkan penelitian tentang bagaimana variasi rasio antara data latih dan data uji dapat memengaruhi hasil klasifikasi.

1.2.2 Pertanyaan Penelitian

- 1) Bagaimana pengaruh skema pembagian data latih dan uji (80:20, 70:30, dan 90:10 terhadap *accuracy* model klasifikasi *Random Forest*?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Tujuan dari penelitian ini adalah sebagai berikut:

- 1) Mengetahui pengaruh pada skema pembagian data latih dan uji untuk klasifikasi *Random Forest*.

1.3.2 Manfaat

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan referensi bagi peneliti atau praktisi dalam memilih skema pembagian data yang optimal dalam membangun model klasifikasi berbasis *machine learning*, khususnya algoritma *Random Forest*.
2. Menunjukkan efektivitas algoritma *Random Forest* dalam klasifikasi data kesehatan, khususnya dalam mendeteksi risiko diabetes secara dini.
3. Memberikan kontribusi dalam pengembangan sistem pendukung keputusan (*decision support system*) untuk diagnosis penyakit berbasis data.

1.4 Batasan Masalah

Agar penelitian ini tetap terarah dan tidak keluar dari fokus utama, maka ruang lingkup penelitian dibatasi sebagai berikut:

- 1) Penelitian ini hanya menggunakan satu jenis algoritma klasifikasi, yaitu *Random Forest*, tanpa melakukan perbandingan dengan algoritma lain.
- 2) Hasil dari penelitian ini berupa hasil klasifikasi. Ada 2 class, yaitu diabetes dan tidak diabetes.
- 3) *Dataset* yang digunakan merupakan *dataset* publik terkait diabetes yang diambil dari platform *Kaggle*, yaitu "*Healthcare-Diabetes*":
”<https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes>”, yang berisi data rekam medis pasien dengan label kondisi diabetes.
- 4) Atribut pada *dataset* yaitu *Id*, *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, *Age*, *Outcome*.
- 5) Penilaian performa model dilakukan berdasarkan metrik evaluasi klasifikasi yaitu *accuracy*, *confusion matrix*, dan *loss*. Nilai *loss* dihitung berdasarkan persentase prediksi yang salah (*error rate*) dari hasil klasifikasi.

1.5 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun dalam lima bab utama agar pembahasan lebih terstruktur dan mudah dipahami. Adapun sistematika penulisan dalam laporan ini adalah sebagai berikut:

Bab I – Pendahuluan

Bab ini berisi latar belakang permasalahan, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup penelitian, serta sistematika penulisan.

Tujuannya adalah memberikan gambaran umum mengenai topik dan arah penelitian.

Bab II – Tinjauan Pustaka

Bab ini memuat teori-teori pendukung yang relevan dengan topik penelitian, seperti konsep diabetes mellitus, klasifikasi dalam *machine learning*, algoritma *Random Forest*, serta penelitian-penelitian terdahulu yang mendukung.

Bab III – Metodologi Penelitian

Bab ini menjelaskan tahapan-tahapan penelitian yang dilakukan, mulai dari deskripsi *dataset* yang digunakan, proses preprocessing data, pemisahan data latih dan uji, penerapan algoritma *Random Forest*, serta evaluasi kinerja model.

Bab IV – Hasil dan Pembahasan

Bab ini menjelaskan hasil dari penelitian yang dilakukan. Disini membahas mulai dari awal, yaitu pengunduhan dataset sampai hasil klasifikasinya.

Bab V – Kesimpulan dan Saran

Bab ini menyimpulkan hasil dari penelitian yang dilakukan. Dan juga menambahkan beberapa saran untuk penelitian kedepannya yang mungkin akan dilakukan oleh orang lain.