

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes, merupakan penyakit kronis yang ditandai oleh peningkatan kadar glukosa dalam darah. Diabetes merupakan salah satu permasalahan kesehatan global yang cukup tinggi [1]. Berdasarkan laporan dari *International Diabetes Federation*, jumlah penderita diabetes di seluruh dunia diperkirakan mencapai sekitar 537 juta pada tahun 2021 dan akan meningkat menjadi 783 juta pada tahun 2045. Indonesia sendiri berada di peringkat ke-5 dengan jumlah penderita diabetes terbanyak di dunia. Dari data tersebut dapat disimpulkan bahwa penderita diabetes terus meningkat dari tahun ke tahun, sehingga diperlukan tindakan pencegahan yang lebih baik untuk mengatasi permasalahan diabetes secara global[2].

Diabetes sering kali tidak teridentifikasi, yang menyebabkan penyakit tersebut berkembang secara bertahap sehingga memicu komplikasi yang lebih serius. Meskipun sudah ada metode diagnosis klinis yang menjadi standar pemeriksaan, tidak semua masyarakat dapat melakukan skrining kesehatan secara rutin karena keterbatasan seperti lokasi, waktu, dan biaya. Oleh karena itu, dibutuhkan alternatif lain yang dapat membantu masyarakat untuk mengetahui risiko diabetes secara mandiri [3]. Salah satu upaya pencegahan peningkatan penyakit diabetes dapat dilakukan dengan menggunakan *machine learning* untuk mendeteksi lebih awal risiko diabetes. Jika disosialisasikan dengan baik, sistem prediksi ini dapat membuat masyarakat menjadi lebih sadar untuk melakukan pencegahan diabetes lebih awal dan mencegah penyakit tersebut berkembang menjadi lebih serius.

Setiap algoritma *machine learning* memiliki keunggulannya masing-masing. *Logistic Regression* terbukti cocok untuk pengembangan model prediksi, menggunakan probabilitas untuk mengidentifikasi hubungan antara variabel independen, terutama pada data yang dapat dipisahkan secara *linear* [4]. *Support Vector Machines* (SVM) memiliki keunggulan untuk menangani data *non-linear* dengan menggunakan kernel trick [5]. *Random forest* merupakan metode *ensemble*

berbasis decision tree, yang dapat meningkatkan akurasi dengan menggabungkan prediksi dari berbagai tree untuk mengurangi *overfitting* [6]. Selain itu, ada teknik penggabungan beberapa model atau bisa disebut dengan *ensemble*. Metode ini menggunakan kombinasi dari berbagai model dasar yang saling melengkapi kelebihan dan kekurangan dari setiap model lainnya [7].

Penelitian mengenai prediksi diabetes telah dilakukan oleh Priyanka Rajendra dan Shahram Latif pada tahun 2021, menggunakan algoritma *Logistic Regression* sebagai metode utama, serta menerapkan teknik *ensemble* seperti *Max Voting* dan *Stacking*. Pada penelitian ini dataset yang digunakan merupakan dataset klinis dari PIMA *Indian Diabetes* yang mencakup beberapa faktor seperti berapa nilai glukosa, tekanan darah, umur dan sebagainya. Pada dataset PIMA *Indian Diabetes*, *Logistic Regression* mencapai akurasi 75,32%, *Decision Tree* 72,29%, SVM 74,89%, KNN 67,97% *Naive bayes* 74.03%, yang meningkat menjadi 76,35% dengan *Stacking*, dan lebih meningkat lagi menjadi 77,7% menggunakan *max voting* [8].

Penelitian serupa dilakukan oleh Vilorio, Herazo-Beltran dan Cabrera, menggunakan metode *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF) untuk memprediksi diabetes. Data klinis pada penelitian ini diambil dari 500 pasien di Kolombia yang mencakup faktor, seperti usia, BMI, kadar glukosa, serta riwayat kesehatan yang dikategorikan menjadi tidak memiliki diabetes, berisiko diabetes, atau menderita diabetes. Model SVM mendapatkan hasil akurasi sebesar 95,36% ketika digunakan pada dataset pasien lokal. Namun, akurasi menurun menjadi 66,25% saat diterapkan pada dataset *Pima Indian*, yang memiliki karakteristik populasi berbeda [9].

Penelitian lain pada tahun 2019, yang dilakukan oleh Devika R dan Sai Vaishnav menggunakan algoritma *Naive Bayes*, *K-Nearest Neighbor* (KNN), dan *Random forest* sebagai model utama untuk memprediksi penyakit ginjal. Penelitian ini menggunakan dataset yang mencakup atribut kesehatan pasien, seperti usia, tekanan darah, kadar glukosa, dan sebagainya. Penelitian ini memberikan hasil bahwa model *Random forest* mencapai akurasi tertinggi yaitu 99,84%, sedangkan KNN dan *Naive Bayes* mencapai akurasi 99,63% dan 87,78% [10].

Berdasarkan penelitian terdahulu, terbukti bahwa model *Logistic Regression*, SVM, dan *Random forest* memiliki hasil evaluasi yang lebih baik dibandingkan model lainnya dalam memprediksi risiko penyakit. Penerapan *ensemble* juga dapat meningkatkan akurasi dibandingkan hanya menggunakan model tunggal. Penelitian sebelumnya yang menerapkan konsep *ensemble* masih memiliki kekurangan dalam memilih kombinasi model, di mana model dengan akurasi rendah ikut di *ensemble*, yang dapat mempengaruhi akurasi. Selain itu, sebagian besar penelitian sebelumnya menggunakan dataset klinis, dimana dalam penerapannya, membutuhkan pemeriksaan manual dengan alat medis dan butuh biaya tambahan.

Oleh karena itu, penelitian ini mengembangkan model prediksi dengan menerapkan *ensemble Weighted Voting*, yang mengkombinasikan tiga model dengan akurasi tertinggi berdasarkan penelitian sebelumnya. Penelitian ini menggunakan dataset survei yang mencakup faktor gaya hidup dan riwayat kesehatan, sehingga memungkinkan prediksi tanpa memerlukan pemeriksaan medis secara langsung. Dengan pemilihan model yang lebih baik, diharapkan dapat menghasilkan prediksi yang lebih akurat dan dapat diterapkan untuk deteksi dini risiko diabetes, dengan memanfaatkan model *ensemble*.

1.2 Rumusan Masalah

Berdasarkan latar belakang permasalahan tersebut, ada beberapa rumusan masalah sebagai berikut :

1. Bagaimana implementasi *Ensemble Weighted Voting* dengan menggabungkan algoritma *Logistic Regression*, SVM, dan *Random forest* dalam memprediksi risiko diabetes?
2. Bagaimana akurasi dan performa dari model *Ensemble Weighted Voting* yang menggabungkan algoritma *Logistic Regression*, SVM, dan *Random forest* dalam memprediksi risiko diabetes?
3. Bagaimana cara penerapan model prediksi risiko diabetes ke dalam bentuk sistem berbasis web?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang dijelaskan, ada beberapa tujuan penelitian sebagai berikut:

1. Membuat model prediksi risiko diabetes menggunakan metode *Ensemble* dengan teknik *Weighted Voting* yang menggabungkan algoritma *SVM*, *Logistic Regression*, dan *Random forest*.
2. Mengevaluasi hasil model prediksi risiko diabetes menggunakan metode *Weighted Voting* yang menggabungkan algoritma *SVM*, *Logistic Regression*, dan *Random forest*.
3. Mengimplementasikan model prediksi risiko diabetes ke dalam sistem berbasis web.

1.4 Manfaat Penelitian

Adapun manfaat yang akan didapatkan dari penelitian pengembangan model *ensemble* menggunakan algoritma *Logistic Regression*, *SVM*, dan *Random forest*, dalam melakukan prediksi risiko diabetes sebagai berikut :

1. Memberikan kontribusi terhadap pengembangan model dengan metode *Ensemble Weighted Voting* dalam prediksi risiko diabetes.
2. Membantu masyarakat mendeteksi potensi risiko diabetes secara mandiri sebelum melakukan diagnosis klinis.

1.5 Batasan Masalah

Berdasarkan latar belakang diatas, agar penelitian ini tidak terlalu luas dan hanya fokus pada permasalahan yang akan dikaji, maka batasan-batasan tersebut sebagai berikut :

1. Dataset yang digunakan mungkin tidak mencakup semua variabel yang relevan untuk memprediksi risiko diabetes, seperti data genetik atau faktor lingkungan.
2. Model dirancang untuk memprediksi 2 target hasil prediksi yaitu berisiko diabetes dan tidak berisiko diabetes.