

ABSTRAK

Diabetes mellitus merupakan salah satu penyakit kronis yang prevalensinya terus meningkat di seluruh dunia. Deteksi dini sangat penting untuk mengurangi risiko komplikasi yang ditimbulkan. Perkembangan teknologi *machine learning* telah memungkinkan penggunaan algoritma klasifikasi untuk membantu menganalisis data medis dalam rangka prediksi penyakit diabetes. Pada penelitian ini digunakan algoritma *Random Forest* untuk melakukan klasifikasi diabetes menggunakan *Dataset Healthcare Diabetes* dari *Kaggle*. Fokus penelitian adalah menganalisis pengaruh skema pembagian data latih dan uji (70:30, 80:20, dan 90:10) terhadap akurasi model. Proses penelitian meliputi tahapan pra-pemrosesan data, pemisahan fitur dan *label*, pembagian data latih dan uji sesuai skenario, pelatihan model menggunakan algoritma *Random Forest*, serta evaluasi model berdasarkan metrik akurasi dan *confusion matrix*. Variasi parameter yang digunakan dalam penelitian ini meliputi *criterion* (*gini* dan *entropy*) serta *n_estimators* (40, 60, 80, dan 100). Hasil penelitian menunjukkan bahwa algoritma *Random Forest* mampu mengklasifikasikan data diabetes dengan akurasi tinggi, yakni di atas 98% pada semua skenario. Skema pembagian 80:20 menghasilkan akurasi terbaik sebesar 99,64%, sedangkan skema 70:30 memberikan akurasi terendah sebesar 98,19%. Hal ini membuktikan bahwa proporsi pembagian data latih dan uji berpengaruh terhadap performa model, dengan skema 80:20 terbukti paling optimal.

Kata kunci: Diabetes Mellitus, *Machine learning*, *Random Forest*, Klasifikasi, Skema Pembagian Data