

ABSTRAK

Diabetes merupakan penyakit tidak menular yang memiliki tingkat kematian 70% di dunia. Mayoritas kasus diabetes, 90%-95%, merupakan kasus diabetes merupakan diabetes tipe 2 yang disebabkan oleh gaya hidup yang tidak sehat. Diabetes tipe 2 dapat dideteksi lebih dini dengan menggunakan pemeriksaan yang mengandung parameter terkait diabetes. Namun, dataset tidak selalu berisi informasi yang lengkap, distribusi antara kelas positif dan negatif sebagian besar tidak seimbang, dan beberapa parameter memiliki kepentingan yang rendah untuk kelas keputusan. Untuk mengatasi masalah tersebut maka perlu dilakukan preprocessing untuk meningkatkan presisi dan recall pada deteksi diabetes. Dalam makalah ini, kami mengusulkan pendekatan pada pemrosesan awal dataset, yang diterapkan pada deteksi diabetes. Pendekatan preprocessing terdiri dari proses berikut: proses nilai hilang, proses data tidak seimbang, proses kepentingan fitur, dan proses augmentasi data. Proses preprocessing data menggunakan median untuk nilai yang hilang, random oversampling untuk data yang tidak seimbang, skor Gini di random forest untuk kepentingan fitur, dan distribusi posterior untuk augmentasi data. Kami menggunakan random forest dan logistik regresi sebagai algoritma klasifikasi. Hasil percobaan menunjukkan bahwa klasifikasi meningkat dengan presisi 20% dan recall 24% dengan menerapkan metode preprocessing yang kami usulkan dengan klasifikasi random forest dibandingkan dengan metode preprocessing tanpa usulan dengan klasifikasi random forest.

Keywords: Diabetes Mellitus Type 2, Data Preprocessing, Data Augmentation, Random Forest, Classification