

ABSTRAK

Varian genetik manusia merupakan perbedaan rangkaian *Deoxyribonucleic Acid* (DNA) dalam genom individu dari suatu populasi manusia. Penelitian varian genetik menghasilkan informasi penting untuk mengukur *clinical significance* suatu varian. Banyak laboratorium klinis melakukan pengklasifikasian *clinical significance* varian genetik secara manual tanpa teknologi kecerdasan seperti *machine learning*. Hasil klasifikasi tersebut kemudian diunggah ke arsip publik seperti *ClinVar*. Permasalahan timbul dari hasil yang diunggah ke *ClinVar* ketika teridentifikasi perbedaan kesimpulan *clinical significance* pada varian genetik sejenis antar laboratorium. Inkonsistensi ini dapat menimbulkan kebingungan dan kesalahan dalam pengambilan keputusan medis yang tepat, sehingga berpotensi merugikan berbagai pihak. Pada penelitian ini, dikembangkan suatu model klasifikasi berbasis algoritma *Gradient Boosting Classifier* untuk klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar*. *Dataset* yang digunakan adalah *ClinVar* yang tidak seimbang dari *Kaggle*. Terdapat dua skenario utama pada penelitian ini, yaitu *hyperparameter tuning* dengan dan tanpa pembobotan sampel saat pelatihan untuk menemukan model terbaik, serta penerapan *feature importance* berdasarkan model terbaik. Pada tahap *data preprocessing*, dilakukan penghapusan fitur *redundant*, pembagian data dengan *stratified split*, imputasi *missing values*, penanganan *outlier*, *data encoding*, dan *data scaling*. *Grid Search Cross-Validation* digunakan untuk menemukan kombinasi *hyperparameter* terbaik meliputi *n_estimators*, *max_depth*, *subsample*, dan *max_features*. Metode pembobotan sampel yang digunakan adalah *Sample Based Class Weight*, yang memanfaatkan *library Scikit-learn*. Berdasarkan hasil pengujian, model terbaik menghasilkan nilai *F1-score* untuk kelas 1 (minor) = 57,00%, kelas 0 = 75,00%, dengan nilai *G-mean* = 72,41%, dan nilai *balanced_accuracy* = 73,05%. Model tersebut dibangun menggunakan *class weight* = 'balanced' dengan kombinasi *hyperparameter* *n_estimators* = 128, *max_depth* = 7, *subsample* = 1,0, *max_features* = 'sqrt', dan tanpa penerapan *feature importance*.

Kata kunci : *ClinVar*, Genetika, *Gradient Boosting Classifier*, Inkonsistensi, Klasifikasi