ABSTRACT

Human genetic variants represent differences in Deoxyribonucleic Acid (DNA) sequences within individual genomes of a human population. Genetic variant research yields essential information for measuring the clinical significance of variants. Many clinical laboratories perform manual classification of genetic variant clinical significance without intelligent technologies such as machine learning. These classification results are subsequently uploaded to public archives like ClinVar. Problems arise from data uploaded to ClinVar when differences in clinical significance conclusions are identified for similar genetic variants across laboratories. These inconsistencies may cause confusion and errors in appropriate medical decision-making, potentially adversely affecting various parties. This research developed a classification model based on the Gradient Boosting Classifier algorithm to classify human genetic variant inconsistencies in the ClinVar public archive. The dataset used was an imbalanced ClinVar dataset from Kaggle. The study employed two main scenarios: hyperparameter tuning with and without sample weighting during training to find the best model, and feature importance implementation based on the best model. Data preprocessing involved redundant feature removal, stratified split data division, missing values imputation, outlier handling, data encoding, and data scaling. Grid Search Cross-Validation was utilized to determine the optimal hyperparameter combinations, including n_estimators, max_depth, subsample, and max_features. The sample weighting method utilized Sample Based Class Weight from the Scikit-learn library. Based on testing results, the best model achieved F1-scores of 57,00% for class 1 (minor) and 75,00% for class 0, with a G-mean score of 72,41% and balanced accuracy score of 73,05%. The best model was built using class weight = 'balanced' with hyperparameter combinations of n_estimators = 128, max_depth = 7, subsample = 1,0, max_features = 'sqrt', and without implementing feature importance.

Keywords : Classification, ClinVar, Genetic, Gradient Boosting Classifier, Inconsistency