Abstrak

Penggunaan Machine Learning pada dataset berupa source code telah dapat diterapkan untuk melakukan klasifikasi bahasa pemrograman. Akan tetapi, sampai saat ini belum ada penelitian yang secara langsung membahas dampak dari penerapan beberapa metode preprocessing. Penelitian ini bertujuan untuk menganalisis dampak dari penerapan beberapa metode preprocessing dan kombinasinya terhadap performa model machine learning yang digunakan pada dataset berupa source code. Untuk menyederhanakan penelitian ini dan meningkatkan fokus pada aspek metode preprocessing, maka pada penelitian ini dibangun sebuah model sederhana untuk melakukan klasifikasi source code berdasarkan code coverage-nya yaitu branch atau non-branch. Penelitian ini menggunakan Support Vector Machines (SVM) dan Multi-Layer Perceptron (MLP) pada model yang dibuat dan dianalisis. Metode preprocessing yang digunakan adalah beberapa jenis vectorization yaitu Vector Count, TF-IDF dan Word2Vec. Hasil penelitian ini menunjukkan bahwa penggunaan metode preprocessing dan kombinasinya mempengaruhi performa model secara signifikan. Pada model SVM, vectorization berupa TF-IDF memberikan hasil terbaik. Sedangkan pada model MLP, vectorization berupa Vector Count yang memberikan hasil terbaik. Selain itu, Metode preprocessing berupa ekstraksi karakter alfanumerik memberikan hasil yang maksimal terhadap performa model terlepas dari jenis model dan vectorization yang digunakan.

Kata Kunci : Machine Learning, Preprocessing, Source Code Dataset, Code Coverage, Klasifikasi.