

ABSTRAK

Platform Android mendominasi pasar sistem operasi mobile global namun menjadi target utama serangan malware yang terus berkembang, sehingga diperlukan sistem deteksi yang andal dan efisien. Penelitian ini bertujuan menerapkan seleksi fitur Mutual Information pada dataset MH-100K dan mengevaluasi pengaruhnya terhadap performa model klasifikasi TabNet. Variabel bebas penelitian ini adalah metode seleksi fitur (tanpa seleksi fitur dan dengan seleksi fitur MI), sedangkan variabel terikat adalah performa model berdasarkan metrik accuracy, precision, recall, F1-score, dan AUPRC. Dataset MH-100K yang memuat 58.872 sampel unik dengan 24.833 fitur statis dibagi dengan rasio 70:15:15 secara stratified. Seleksi fitur MI diterapkan pada data latih menggunakan fungsi `mutual_info_classif`, menghasilkan empat skenario threshold (Min, Q1, Q2, Q3) yang masing-masing diuji pada model TabNet. Skenario terbaik diperoleh pada threshold Q3 yang mempertahankan 6.209 fitur atau setara reduksi 75% dari total fitur awal. Hasil evaluasi pada data uji menunjukkan model TabNet + MI-Q3 mencapai accuracy 96,69%, precision 83,65%, recall 93,46%, F1-score 88,28%, dan AUPRC 90,68%, mengungguli model baseline TabNet tanpa seleksi fitur pada empat dari lima metrik dengan peningkatan recall paling signifikan sebesar 9,43%. Penelitian ini membuktikan bahwa seleksi fitur Mutual Information efektif meningkatkan performa model TabNet sekaligus mereduksi dimensi fitur secara signifikan.

Kata kunci: TabNet, Mutual Information, Malware Android, Seleksi Fitur.

ABSTRACT

The Android platform dominates the global mobile operating system market but has become the primary target of increasingly sophisticated malware attacks, necessitating reliable and efficient detection systems. This study aims to apply Mutual Information feature selection on the MH-100K dataset and evaluate its impact on the performance of the TabNet classification model. The independent variable of this study is the feature selection method (without feature selection and with MI feature selection), while the dependent variables are the model performance metrics comprising accuracy, precision, recall, F1-score, and AUPRC. The MH-100K dataset containing 58,872 unique samples with 24,833 static features was split using a 70:15:15 stratified ratio. MI feature selection was applied to the training data using the `mutual_info_classif` function, producing four threshold scenarios (Min, Q1, Q2, Q3) each evaluated on the TabNet model. The optimal scenario was obtained at the Q3 threshold, retaining 6,209 features equivalent to a 75% reduction from the total original features. Evaluation results on the test data show that the TabNet + MI-Q3 model achieved an accuracy of 96.69%, precision of 83.65%, recall of 93.46%, F1-score of 88.28%, and AUPRC of 90.68%, outperforming the baseline TabNet model without feature selection on four out of five metrics with the most significant improvement in recall of 9.43%. This study demonstrates that Mutual Information feature selection effectively enhances TabNet model performance while significantly reducing feature dimensionality.

Keywords: TabNet, Mutual Information, Android Malware, Feature Selection.