

INTISARI

KLASIFIKASI DATA TAK SEIMBANG STATUS *LAPSE* POLIS ASURANSI KENDARAAN MENGGUNAKAN REGRESI LOGISTIK BINER DAN *SUPPORT VECTOR MACHINE*

Oleh

Ahmad Ziyang Faradis Mumtaza

20/459241/PA/19902

Polis asuransi adalah kontrak penting antara perusahaan asuransi dan tertanggung yang mengatur berbagai ketentuan. Salah satu ketentuan yang diatur seperti jatuh tempo pembayaran kontribusi. Apabila kontribusi tidak dibayarkan tepat waktu, status polis akan berubah dari aktif menjadi tidak aktif (*lapse*). Hal ini dapat merugikan kedua belah pihak dan mempengaruhi likuiditas perusahaan. Berdasarkan hal tersebut, prediksi potensi *lapse* di masa mendatang menjadi penting. Pada perusahaan asuransi yang sehat, jumlah polis *lapse* jauh lebih sedikit dibandingkan polis aktifnya. Hal tersebut mengindikasikan adanya ketidakseimbangan data. Pada kasus data tidak seimbang, model klasifikasi yang dihasilkan cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Untuk mengatasi hal ini dapat dilakukan *resampling* data untuk menyeimbangkan data. Metode *oversampling* seperti *Synthetic Minority Oversampling Technique* (SMOTE) dan metode *undersampling* seperti *Tomek Links* dapat digunakan. Pada metode SMOTE ditambahkan data sintetis untuk kelas minoritas, sementara pada metode *Tomek Links* data mayoritas yang serupa dengan kelas minoritas akan dihapus. Pada penelitian ini dilakukan penerapan metode *resampling* SMOTE, *Tomek Links*, dan kombinasi keduanya pada data tak seimbang status *lapse* polis asuransi kendaraan untuk dilakukan pemodelan klasifikasi Regresi Logistik Biner dan *Support Vector Machine* (SVM). Metode SVM digunakan karena metode ini efektif pada data dengan banyak fitur atau dimensi. Pada metode Regresi Logistik Biner, *odds ratio* dari model dapat digunakan untuk mengukur dan mengkomunikasikan efek variabel. Berdasarkan data yang digunakan, dapat ditunjukkan bahwa hasil klasifikasi menggunakan metode SVM memberikan performa lebih baik dibandingkan metode Regresi Logistik Biner.

ABSTRACT

CLASSIFICATION OF IMBALANCED DATA ON VEHICLE INSURANCE POLICY LAPSE STATUS USING BINARY LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINE

By

Ahmad Ziyang Faradis Mumtaza

20/459241/PA/19902

An insurance policy is a crucial contract between the insurer and the insured that establish various terms and conditions. One such condition is the payment due date for contributions. If contributions are not paid on time, the policy status changes from active to inactive (lapse). This situation can disadvantage both parties and affect the company's liquidity. Therefore, predicting the potential for lapse in the future becomes crucial. In a healthy insurance company, the number of lapsed policies is significantly lower than the number of active policies. This indicates the presence of an imbalanced dataset. In cases of imbalanced data, the resulting classification models tend to predict the majority class while neglecting the minority class. To address this issue, data resampling can be applied to balance the dataset. Oversampling methods, such as the Synthetic Minority Oversampling Technique (SMOTE), and undersampling methods, such as Tomek Links, can be used. The SMOTE method generates synthetic data for the minority class, whereas the Tomek Links method removes majority class data that are similar to the minority class. In this study, the SMOTE, Tomek Links, and their combination are applied as resampling methods on imbalanced data related to the lapse status of vehicle insurance policies for classification modeling using Binary Logistic Regression and Support Vector Machine (SVM). The SVM method is employed because it is effective for datasets with many features or dimensions. In Binary Logistic Regression, the odds ratios from the model can be used to quantify and communicate the effects of the variables. Based on the data used, it is demonstrated that the classification results using the SVM method provide better performance than those using the Binary Logistic Regression method.