INTISARI

PERBANDINGAN KINERJA *MACHINE LEARNING* XGBOOST DAN *RANDOM FOREST* DALAM DETEKSI *FRAUD* ASURANSI KENDARAAN BERMOTOR DENGAN INTERPRETASI SHAP

Oleh

Alexander Budiman

21/473884/PA/20440

Fraud atau juga sering disebut dengan kecurangan dalam asuransi kendaraan bermotor merupakan masalah serius yang dapat menyebabkan kerugian finansial pada perusahaan asuransi. Untuk mengatasi hal tersebut, pemanfaatan *machine learning* sebagai alat deteksi dapat digunakan untuk meminimalisir tingkat *fraud*. Metode berbasis pohon seperti XGBoost dan *Random Forest* menawarkan kinerja dan efisiensi yang baik dalam permodelan. Dilakukan pembagian data latih sebesar 70% dan data uji sebesar 30%. Untuk menangani data tidak seimbang dilakukan teknik pembobotan dan menggunakan pencarian acak dalam menetapkan parameter terbaik pada kedua model. Berdasarkan metrik evaluasi, akurasi dan akurasi seimbang pada *Random Forest* sedikit lebih unggul daripada XGBoost. Namun, XGBoost memiliki nilai AUC dan waktu pelatihan model yang lebih baik daripada *Random Forest*. Dengan bantuan interpretasi SHAP, kedua model menunjukkan hal yang sama bahwa variabel yang menjadi indikasi terjadinya *fraud* adalah *incident_severity* dan *insured_hobbies*. Polis asuransi dengan *incident_severity "major damage*" serta seseorang pemilik asuransi dengan hobi *"chess"* dan *"cross-fit"* cenderung diklasifikasikan melakukan *fraud*.

ABSTRACT

PERFORMANCE COMPARISON OF XGBOOST AND RANDOM FOREST MACHINE LEARNING IN DETECTING MOTOR VEHICLE INSURANCE FRAUD WITH SHAP INTERPRETATION

By

Alexander Budiman

21/473884/PA/20440

Fraud in motor vehicle insurance, often referred to as insurance fraud, is a serious issue that can result in significant financial losses for insurance companies. To mitigate this risk, machine learning can be utilized as a detection tool to reduce fraud rates. Tree-based methods such as XGBoost and Random Forest offer good performance and efficiency in modeling. In this study, the data was divided into 70% for training and 30% for testing. To address data imbalance, weighting techniques were applied, and *random search* was used to find the best parameters for both models. Based on evaluation metrics, the accuracy and balanced accuracy of Random Forest slightly outperformed XGBoost. However, XGBoost showed better AUC scores and faster model training times compared to Random Forest. With SHAP interpretation, both models revealed that the variabel indicators of fraud are *incident_severity* and *insured_hobbies*. Insurance policies with an *incident_severity* of “major damage” and policyholders with hobbies such as “chess” and “cross-fit” tend to be classified as fraud cases.