ABSTRACT

ANALYSIS OF SMOTE, TOMEK LINKS, AND HYBRID SMOTE+TOMEK LINKS METHODS ON NAIVE BAYES CLASSIFICATION FOR ADDRESSING IMBALANCED DATA IN TUBERCULOSIS DIAGNOSIS

By

NAFLAH FAULINA

Naive Bayes classification is a method based on Bayes' theorem, utilizing probabilistic and statistical techniques. In practice, many datasets exhibit imbalanced class distributions. A common issue with imbalanced data is that the classifier tends to predict the class with the larger data composition. As a result, this leads to high prediction accuracy for the majority class in the training data but poor prediction accuracy for the minority class. One resampling technique to address this issue involves using oversampling, undersampling, or a combination of both. The aim of this study is to balance the training data by applying SMOTE (Synthetic Minority Oversampling Technique), Tomek Links, and a hybrid approach combining both methods. The performance of the Naive Bayes classifier on the original imbalanced data is compared with its performance on the balanced data in diagnosing tuberculosis at Mayjend HM Ryacudu Kotabumi Hospital. The results show that the hybrid approach, combining SMOTE for oversampling and Tomek Links for undersampling, demonstrates the best performance in balancing the data and improving the accuracy of the Naive Bayes model. Specifically, the hybrid method achieved an average accuracy of 93%, an average sensitivity of 88%, an average specificity of 96%, an average False Positive Fraction (FPF) of 4%, and an average False Negative Fraction (FNF) of 12%.

Key words: Naive Bayes classification, SMOTE, Tomek Links, SMOTE+Tomek Links, Tuberculosis

ABSTRAK

ANALISIS METODE *SMOTE, TOMEK LINKS,* DAN *HYBRID SMOTE+TOMEK LINKS* TERHADAP KLASIFIKASI NAIVE BAYES UNTUK MENGATASI DATA TIDAK SEIMBANG PADA DIAGNOSA PENYAKIT TUBERKULOSIS

OLEH

NAFLAH FAULINA

Klasifikasi naive bayes adalah metode yang didasarkan pada teorema Bayes dengan metode probabilitas dan statistik. Dalam penerapannya banyak data yang ditemui memiliki distribusi tidak seimbang di setiap kelasnya. Permasalahan yang sering terjadi pada data tidak seimbang, klasifikasi cenderung memprediksi kelas yang memiliki komposisi data lebih besar. Akibatnya dihasilkan hasil akurasi prediksi yang baik pada kelas data *training* yang mayoritas, sedangkan akan dihasilkan akurasi prediksi yang buruk pada data *training* yang minoritas. Salah satu teknik *resampling* dapat dilakukan untuk menanganinya yaitu dengan metode *oversampling*, *undersampling*, dan gabungan keduanya. Tujuan dari penelitian ini adalah mengaplikasikan metode *Synthetic Minority Oversampling Technique* (SMOTE), Tomek Links, dan *hybrid* SMOTE+Tomek Links untuk mengatasi data tidak seimbang dan membandingkan kinerja (performa) klasifikasi naive bayes pada data original dan data yang telah seimbang dalam diagnosa penyakit Tuberkulosis di RSD. Mayjend HM Ryacudu Kotabumi. Hasil yang diperoleh *hybrid* antara teknik *oversampling* dengan SMOTE dan teknik *undersampling* dengan Tomek Links menunjukkan kinerja terbaik dalam menyeimbangkan data dan meningkatkan akurasi model naive bayes, dengan akurasi rata-rata mencapai 93%, *sensitivity* rata-rata 88%, *specificity* rata-rata 96%, rata-rata *False Positive Fraction(*FPF*)* 4%, dan rata-rata *False Negative Fraction (*FNF*)* 12%.

Kata Kunci: Klasifikasi Naive Bayes, SMOTE, Tomek Links, SMOTE+Tomek Links, Tuberkulosis