Analisis Perbandingan Metode-metode Rebalancing Dalam Menangani Imbalanced Data Pada Klasifikasi Tingkat Keparahan Covid-19 Dengan Metode Random Forest = Comparative Analysis of Rebalancing Methods in Handling Imbalanced Data on COVID-19 Severity Classification with Random Forest

Muhammad Ilham Randi, author

Abstrak

Dalam melakukan klasifikasi, tidak jarang terdapat data dengan jumlah anggota kategori yang tidak seimbang. Khususnya dalam dunia kesehatan dimana kategori yang diamati umumnya lebih jarang terjadi. Jika ketidakseimbangan ini tidak ditangani terlebih dahulu maka dapat memberikan hasil klasifikasi yang bias dan kurang akurat. Terdapat beberapa metode rebalancing konvensional untuk menanganinya seperti random oversampling dan random undersampling, namun keduanya diklaim memiliki beberapa kelemahan sehingga beberapa metode yang lebih kompleks dikembangkan. Namun jumlah metode yang dapat digunakan untuk menangani data kategorik selain metode konvensional tersebut masih minim. Salah satu metode yang dapat menangani data kategorik adalah synthetic minority over sampling-technique nominal continuous atau SMOTE-NC yang merupakan ekstensi dari SMOTE yang dikembangkan untuk menangani dataset dengan variabel campuran. Skripsi ini membahas perbandingan dari metode random oversampling dan SMOTE-NC juga metode gabungannya dengan undersampling yaitu random oversampling + undersampling dan SMOTE-NC + undersampling untuk menangani ketidakseimbangan data. Masing-masing metode tersebut akan diterapkan untuk klasifikasi tingkat keparahan COVID-19 berdasarkan urgensi perawatan rumah sakit dengan menggunakan metode random forest dimana selanjutnya dapat dilihat kombinasi metode yang menghasilkan performa terbaik. Penelitian ini juga bertujuan untuk melihat faktor-faktor manakah yang paling penting dalam memprediksi tingkat keparahan COVID-19 berdasarkan urgensi rumah sakit. Digunakan metode Leave-One-Out Cross-Validation untuk mengukur konsistensi model. Diperoleh hasil bahwa metode SMOTE-NC dengan undersampling memberikan performa terbaik dengan komorbid paru-paru, kadar c-reactive protein dan prokalsitonin merupakan variabel terpenting dalam model. Selain itu diperoleh kesimpulan bahwa pemilihan metode rebalancing yang tepat bergantung pada karakteristik data yang dimiliki.

.....

In conducting classification, it is not uncommon for data with an unbalanced number of category members. Especially in the world of health where the categories we observe are generally less common. If this imbalance is not handled first, it can give biased and less accurate classification results. There are several conventional rebalancing methods to handle it, such as random oversampling and random undersampling, but both are claimed to have several weaknesses so that several more complex methods were developed. However, the number of methods that can be used to handle categorical data other than the conventional methods is still minimal. One method that can handle categorical data is synthetic minority over sampling-technique nominal continuous or SMOTE-NC which is an extension of SMOTE which was developed to handle datasets with mixed variables. This thesis discusses the comparison of random oversampling and SMOTE-NC methods as well as their combined methods with undersampling, namely random oversampling

+ undersampling and SMOTE-NC + undersampling to handle data imbalances. These methods will be applied to the classification of the severity of COVID-19 based on the urgency of hospital care using the random forest method, wherein the combination of methods that produces the best performance will be seen. This study also aims to see which factors are the most important in predicting the severity of COVID-19 based on hospital urgency. The Leave-One-Out Cross-Validation method is used to measure the consistency of the model. It was found that the SMOTE-NC method with undersampling gave the best performance with lung comorbidities, c-reactive protein and procalcitonin levels were the most important variables in the model. In addition, it can be concluded that the selection of the right rebalancing method depends on the characteristics of the data held.