INTISARI

PENERAPAN *K-MEANS* SMOTE DENGAN *SILHOUETTE COEFFICIENT* UNTUK ANALISIS KLASIFIKASI DATA TIDAK SEIMBANG

oleh

Pingky Oktiawati

19/439215/PA/19038

Permasalahan data yang tidak seimbang antara satu kelas dengan kelas lainnya pada analisis klasifikasi menyebabkan hasil klasifikasi menjadi tidak akurat. Hal ini dapat terjadi karena model prediksi hanya akan memprediksi kelas mayoritas sehingga akurasi tinggi yang dihasilkan tidak merepresentasikan hasil yang sesungguhnya. Metode *resampling* atau modifikasi sampel dengan cara menyeimbangkan jumlah observasi kelas mayoritas dan kelas minoritas diterapkan untuk mengatasi permasalahan tersebut. *K-Means* SMOTE merupakan salah satu teknik *oversampling* modifikasi dari metode *Synthetic Minority Oversampling Technique* (SMOTE). Metode *K-Means* SMOTE bekerja dengan melakukan pengelompokkan data menggunakan algoritma *K-Means* kemudian dilanjutkan dengan *oversampling* menggunakan SMOTE pada klaster dengan lebih banyak kelas minoritas. Penentuan jumlah klaster optimal dilakukan dengan menggunakan *Silhouette Coefficient*. Selanjutnya, SMOTE akan menambah *instance* kelas minoritas untuk membuat jumlah sampel pada kelas minoritas dan kelas mayoritas menjadi lebih seimbang dengan cara membangkitkan *instance* sintetis, yaitu objek baru yang tidak ada dalam *dataset* tetapi memiliki kemiripan dengan objek yang ada dalam *dataset*. Pada penelitian ini dilakukan implementasi *K-Means* SMOTE dengan *Silhouette Coefficient* pada data yang tidak seimbang dari *dataset Pima Indians Diabetes, dataset Haberman's Survival,* dan data simulasi *Pima Indians Diabetes* dengan menggunakan metode klasifikasi *Random Forest*.

Kata kunci: klasifikasi, data tidak seimbang, *oversampling, K-Means*, SMOTE, *Silhouette Coefficient, Random Forest*

ABSTRACT

IMPLEMENTATION OF K-MEANS SMOTE WITH SILHOUETTE COEFFICIENT FOR CLASSIFICATION ANALYSIS OF IMBALANCED DATA

by

Pingky Oktiawati

19/439215/PA/19038

The problem of imbalanced data between one class and another in classification analysis causes an inaccurate result. This can happen because the prediction model will only predict the majority class so that the high accuracy does not represent the actual results. The resampling method or sample modification by balancing the number of observations of the majority class and the minority class is applied to overcome this problem. K-Means SMOTE is a modified oversampling technique from the Synthetic Minority Oversampling Technique (SMOTE) method. K-Means SMOTE method works by grouping data using the K-Means algorithm and then proceeding with oversampling using SMOTE on clusters with more minority classes. The optimal number of clusters is determined using the Silhouette Coefficient. Furthermore, SMOTE will add instances of the minority class to make the number of samples in the minority class and majority class be more balanced by generating synthetic instances, new objects that are not in the dataset but have a resemblance to the objects in the dataset. In this study, a K-Means SMOTE with Silhouette Coefficient is applied to imbalanced data from the Pima Indians Diabetes, Haberman's Survival, and simulation dataset using the Random Forest classification method.

Keywords: classification, imbalanced data, oversampling, K-Means, SMOTE, Silhouette Coefficient, Random Forest