Abstract

Imbalanced data presents significant challenges in machine learning, leading to biased classification outcomes favoring the majority class. This issue is especially pronounced in financial distress classification, where data imbalance is common due to the scarcity of such instances in real-world datasets. This study aims to mitigate data imbalance in financial distress companies using the Kmeans-SMOTE approach by combining Kmeans clustering and the Synthetic Minority Oversampling Technique (SMOTE). Various classification approaches, including Naïve Bayes and Support Vector Machine (SVM) are implemented on a financial distress dataset from Kaggle to evaluate the effectiveness of Kmeans-SMOTE. Experimental results show that SVM outperforms Naïve Bayes with impressive accuracy (99.1%), f1-score (99.1%), Area Under Precision-Recall (AUPRC) (99.1%), and Geometric-mean (Gmean) (98.1%).