A comparative study of over-sampling techniques as applied to seismic events

Mpho Mokoatle[1,2][0000−0001−9252−3914], Toshka Coleman[1][1111−2222−3333−4444], and Paul Mokilane[1][0000−0002−2649−8711]

[1] Council for Scientific and Industrial Research
Cluster: Next Generation Enterprises and Institutions, Data Science
https://www.csir.co.za/
[2] University of Pretoria, Pretoria, South Africa

Abstract. The likelihood that an earthquake will occur in a specific location, within a specific time frame, and with ground motion intensity greater than a specific threshold is known as a **seismic hazard**. Predicting these types of hazards is crucial since doing so can enable early warnings, which can lessen the negative effects. Research is currently being executed in the field of machine learning to predict seismic events based on previously recorded incidents. However, because these events happen so infrequently, this presents a class imbalance problem to the machine learning or deep learning learners. As a result, this study provided a comparison of the performance of popular over-sampling techniques that seek to even out class imbalance in seismic events data. Specifically, this work applied SMOTE, SMOTENC, SMOTEN, BorderlineSMOTE, SVMSMOTE, and ADASYN to an open source Seismic Bumps dataset then trained several machine learning classifiers with stratified K-fold cross-validation for seismic hazard detection. The SVMSMOTE algorithm was the best over-sampling method as it produced classifiers with the highest overall accuracy, F1 score, recall, and precision of 100%, respectively, whereas the ADASYN over-sampling methodology showed the lowest performance in all the reported metrices of all the models. To our understanding, no research has been done comparing the effectiveness of the aforementioned over-sampling techniques for tasks involving seismic events.