RINGKASAN

METODE *NEAR MISS UNDERSAMPLING* UNTUK OPTIMALISASI *DATASET* PADA MODEL PREDIKSI PENYAKIT STROKE

Yusuf Muhammad Nur Zaman

Stroke merupakan kondisi yang terjadi ketika suplai darah ke otak terganggu atau berkurang akibat adanya penyumbatan atau pecahnya pembuluh darah. Pada tahun 2018 penyakit stroke menjadi penyebab kematian nomor satu di Indonesia, salah satu penyakit dengan tingkat prevalensi yang tinggi, dan juga biaya *pre-diagnosis* dan pengobatan yang tidak murah.

Beberapa data set medis yang memiliki dua kelas atau *binomial class* mengalami ketidakseimbangan kelas, hal ini disebut dengan *imbalanced dataset*, yaitu suatu kondisi ketika kelas tujuan yang akan diklasifikasi tidak memiliki rasio yang seimbang. Sehingga menyebabkan hasil klasifikasi yang bias karena *classifier* akan lebih condong mendeteksi kelas mayoritas dibanding dengan kelas minoritas. Kasus *imbalanced dataset* ini dapat diatasi dengan menerapkan metode *Near-Miss undersampling.*

Berdasarkan hasil pelatihan dan penelitian, model dengan *dataset* orisinil memiliki nilai skor f1, *precision*, dan *recall* bernilai 0.00, sedangkan model yang menggunakan *dataset* hasil dari *near-miss undersampling* versi 1 memiliki skor f1 sebesar 0.731, skor *precision* sebesar 0.717, dan skor *recall* sebesar 0.745, lalu model yang menggunakan *dataset* hasil dari *near-miss undersampling* versi 2 memiliki skor f1 sebesar 0.660, skor *precision* sebesar 0.673, dan skor *recall* sebesar 0.647, dan model yang menggunakan *dataset* hasil dari *near-miss undersampling* versi 3 memiliki skor f1 sebesar 0.747, skor *precision* sebesar 0.673, dan skor *recall* sebesar 0.725. Sehingga model dengan menggunakan *dataset* hasil *near-miss undersampling* memiliki performa yang jauh lebih baik dalam klasifikasi.

Kata kunci : *imbalanced dataset,* stroke, metode *near-miss undersampling*

SUMMARY

NEAR MISS UNDERSAMPLING METHOD FOR DATASET OPTIMIZATION IN STROKE DISEASE PREDICTION MODEL

Yusuf Muhammad Nur Zaman

Stroke is a condition that occurs when the blood supply to the brain is interrupted or reduced due to a blockage or rupture of a blood vessel. In 2018 stroke became the number one cause of death in Indonesia, one of the diseases with a high prevalence rate, and the cost of pre-diagnosis and treatment is not cheap.

Some medical data sets that have two classes or binomial classes experience a class imbalance, this is called an imbalanced data set, which is a condition when the destination class to be classified does not have a balanced ratio. This causes biased classification results because the classifier will be more inclined to detect the majority class than the minority class. This imbalanced dataset case can be overcome by applying the Near Miss undersampling method.

Based on the results of training and testing, the model with original dataset has a f1 score, precision score, and recall score are 0.00, while the model with dataset from near-miss undersampling version 1 has a f1 score of 0731, precision score of 0.717, and recall of 0.745, the the model with dataset from near-miss undersampling version 2 has a f1 score of 0.660, precision score of 0.673, and recall score of 0.647, and the model with dataset from near-miss undersampling version 3 has a f1 score of 0.747, precision score of 0.673, dan recall score of 0.725. So the model with undersampled using near-miss has a mush better performance in classification.

Keywords : imbalanced dataset, stroke, near miss undersampling method