# SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

**Informasi Pelaksanaan :**
SNESTIK II - Surabaya, 26 Maret 2022
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

**Informasi Artikel:**
DOI : 10.31284/p.snestik.2022.2868

**Prosiding ISSN 2775-5126**

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id


# Implementasi Algoritma SMOTE Sebagai Penyelesaian *Imbalance Hight Dimensional Datasets*


Rinci Kembang Hapsari[1*], Tutuk Indriyani[2]

Jurusan Teknik Informatika, Institut Teknologi Adhi Tama Surabaya[1,2]
*e-mail: *rincikembang@itats.ac.id*



## *ABSTRACT*

*In real life, especially in the medical field, multiclass classifications are often encountered with unbalanced input data imbalanced datasets. The major class is the larger data, while the minor class is the small. The imbalanced condition of the dataset dramatically affects the accuracy of the classification process. The classification algorithm will experience a decrease in performance if it is given imbalanced input data. Therefore, it is necessary to balance the input data to maintain the performance of the classification algorithm. So, in this study, the SMOTE algorithm was applied to solve the problem of unbalanced class distribution in the imbalanced dataset. This study uses three datasets: Dataset 1 consisting of 68 data, Dataset 2 consisting of 180 data, and Dataset 3 consisting of 371 data. The three datasets become balanced data after being operated with the SMOTE algorithm.*



***Keywords:*** *High Dimensional Datasets; Imbalanced; SMOTE Algorithm.*



## ABSTRAK

Dalam kehidupan nyata, khususnya di bidang medis, sering dijumpai klasifikasi *multiclass* dengan data input yang tidak seimbang, *imbalanced dataset*. Kelas mayor merupakan jumlah data yang lebih banyak, sedangkan kelas minor jumlahnya sedikit. Kondisi *dataset* yang *imbalanced* sangat mempengaruhi hasil akurasi proses klasifikasi. Algoritma klasifikasi akan mengalami penurunan performa jika diberikan input data yang *imbalanced*. Oleh karena itu, diperlukan penyeimbangan data input untuk mempertahankan performa algoritma klasifikasi. Sehingga, dalam penelitian ini diterapkan algoritma SMOTE untuk menyelesaikan permasalahan distribusi kelas yang tidak seimbang pada *imbalanced dataset*. Penelitian ini menggunakan 3 *dataset*, yaitu Dataset 1 yang terdiri dari 68 data, Dataset 2 terdiri dari 180 data, dan Dataset 3 terdiri dari

371 data. Setelah dioperasikan dengan algoritma SMOTE, ketiga *dataset* tersebut menjadi data yang seimbang.

**Kata kunci:** Algoritma SMOTE; *High Dimensional Datasets*; *Imbalanced.*

## PENDAHULUAN

Penelitian di bidang medis telah banyak dilakukan untuk memprediksi kasus penyakit yang jarang dibandingkan dengan populasi normal. Sehingga ketidakseimbangan kelas menjadi masalah umum di sebagian besar dataset medis. Dataset yang antar kelas data satu dengan kelas yang lain tidak berimbang, yang disebut *imbalanced data*. Dimana kelas yang memiliki data paling banyak disebut kelas mayoritas, sedangkan kelas yang memiliki data sedikit disebut kelas minoritas. Dengan adanya ketidakseimbangan kelas, kelas minoritas memiliki jumlah *instance* yang signifikan lebih rendah dibandingkan dengan kelas lainnya.

Kebanyakan pengklasifikasi bertujuan untuk mencapai kinerja yang optimal di seluruh kelas. Kondisi *imbalanced* data menjadi masalah dalam proses klasifikasi, hal ini akan menyulitkan metode klasifikasi pada saat melakukan fungsi generaliasai pada proses *machine learning* [1]. Hampir semua algoritma klasifikasi akan condong memprediksi kelas mayoritas dibandingkan dengan kelas minoritas. Sehingga menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas dari pada kelas minoritas [2][3][4].

Ada beberapa penyebab buruknya hasil algoritma pembelajaran pada klasifikasi kelas minoritas. Sampel minoritas mungkin diperlakukan sebagai noise, ukuran sampel kecil dapat menyebabkan tantangan bagi model untuk mendeteksi pola minoritas dan metrik evaluasi bias terhadap kelas mayoritas [5][6]. Dalam aplikasi medis, kesalahan klasifikasi kelas minoritas pasien, membebankan biaya lebih dari kesalahan dalam mengklasifikasikan orang sehat. Namun, algoritma pembelajaran standar sebagian besar mengasumsikan kesalahan klasifikasi yang sama dan distribusi kelas yang seimbang. Jika kondisi *imbalance class* diabaikan, algoritma klasifikasi akan mengalami penurunan performa [7].

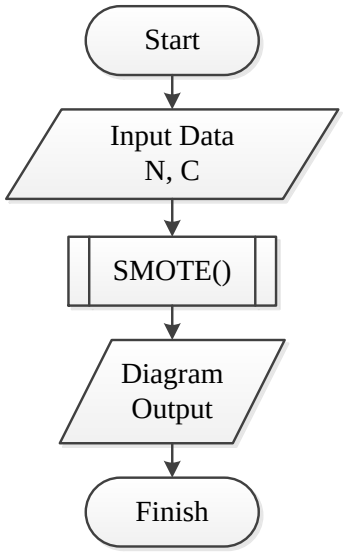
Dalam mengatasi masalah ketidakseimbangan kelas, terdapat dua pendekatan utama[8]. Pada level data, distribusi kelas data menjadi cukup seimbang dengan teknik sampling. Pada tingkat algoritma, distribusi data tetap tidak berubah, tetapi dengan memodifikasi biaya kesalahan klasifikasi di kelas minoritas, model telah disesuaikan untuk lebih fokus pada pembelajaran kelas minoritas [9][10]. Dalam pergerakan *threshold* yang dikategorikan dalampendekatan level algoritma, prediksi label kelas didasarkan pada *threshold* optimal, dengan *default threshold* (0,5) yang rutin digunakan [8].

Teknik sampling pada *imbalanced* data mengakibatkan tingkat *imbalanced data* semakin kecil dan proses klasifikasi dapat dilakukan dengan tepat[11]. Pendekatan *samping* dibedakan menjadi dua, yaitu pertama, *oversampling* dilakukan untuk menyeimbangkan jumlah distribusi data dengan cara jumlah data pada kelas minoritas ditingkatkan. Kedua adalah *undersampling*, yaitu dilakukan dengan melakukan pengurangan jumlah data pada kelas mayoritas agar data seimbang. Salah satu metode *oversampling* adalah metode *Synthetic Minority Oversampling Technique* (SMOTE). Metode ini melakukan pembuatan "*synthetic"* data, yang merupakan data replikasi dari data kelas monoritas. Penerapan SMOTE dapat memperbaiki kualitas klasterisasi, dengan berhasil menghilangkan noise serta menyelesaikan masalah *imbalanced* pada 71 dataset[12]. Klasifikasi pada data *imbalanced* menghasilkan nilai akurasi yang lebih tinggi dengan menggunakan algoritma SMOTE dari pada klasifikasi yang ridak menggunakan algoritma SMOTE [13][14].

Pada penelitian ini, peneliti melakukan penanganan data *imbalanced* terhadap kelas minor menggunakan pendekatan teknik *sampling*, yaitu *oversampling*. Teknik *oversampling* dipilih karena dilakukan dengan menambahkan dataset pada kelas minoritas, sehingga tidak mengalami kehilangan informasi dari dataset. Berdasarkan dari penelitian sebelumnya, makaalgoritma *oversampling* yang digunakan adalah SMOTE, karena bisa menghasilkan akurasi yang lebih baik dan efektif dalam menangani kelas minoritas dan bisa mengurangi *overfitting.*

## METODE

Dalam penelitian ini untuk menyelesaikan permasalahan data yang tidak seimbang (*imbalance*) digunakan mtode SMOTE. Alur proses sistem ditunjukkan pada Gambar 1.




Gambar 1. Flowchart proses sistem

### *Imbalance Class*

Kelas yang tidak seimbang adalah masalah umum dalam klasifikasi pembelajaran mesin. Dimana imbalance class merupakan kondisi distribusi antar kelas yang tidak proporsioanal pada sebuah dataset, dimana terdapat salah satu kelas yang memiliki jumlah data sangat besar (kelas mayoritas) dibandingkan dengan kelas lainnya (kelas minoritas) [9]. Perbedaan jumlah data yang sangat besar antar kelas dapat mengakibatkan model klasifikasi sering tidak mampu memprediksikan kelas minoritas dengan tepat sehingga banyak data pengujian yang seharusnya berada pada kelas minoritas diprediksikan salah oleh model klasifikasi [15].

### Syntetics Minority Oversampling Technique (SMOTE)

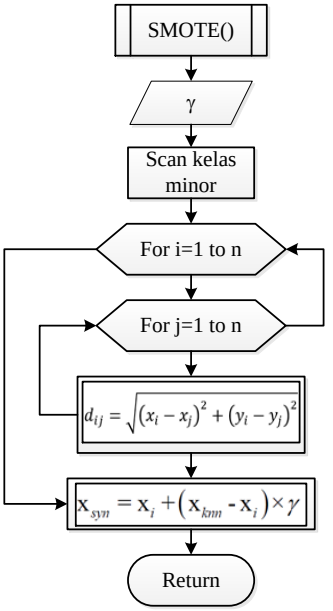
Algoritma SMOTE merupakan teknik *oversampling* dengan melakukan peningkatan jumlah data dalam kelas minoritas dengan cara melakukan replikasi jumlah data kelas minoritas secara random sehingga jumlahnya sama atau mendekati dengan data kelas mayoritas. Algoritma SMOTE bekerja dengan mencari K-Nearest Neighbor, yaitu mengelompokkan data berdasarkan pada tetangga terdekat.

Pemilihan tetangga terdekat dilakukan berdasarkan jarak *Euclidean* antara sepasang data. Diberikan data dengan $p$ variabel, yaitu $\boldsymbol{x}^T = [x_1, x_2, \cdots, x_n]$ dan $\boldsymbol{z}^T = [z_1, z_2, \cdots, z_n]$, sehingga jarak *Euclidean d(x,z)* dihitung dengan Persamaan 1.

$$d(x,z) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_n - z_n)^2} \quad (1)$$

Data *synthetics* dibangkitkan dengan menggunakan Persamaan 2.

$$x_{syn} = x_i + (x_{knn} - x_i) * \gamma \quad (2)$$




Gambar 2. Alur algoritma SMOTE

dimana, $x_{syn}$ adalah data hasil replikasi
$x_i$ adalah data ke-i dari kelas minor
$x_{knn}$ adalah data dari kelas minor yang memiliki jarak terdekat dari kelas $x_i$
$\gamma$ adalah bilangan random antara 0 dan 1

Tahap yang dilakukan pada algoritma SMOTE ditunjukkan pada Gambar 2. Untuk contoh simulasi algoritma SMOTE, maka diberikan simulasi data dan hasil algoritma SMOTE pada Tabel 1. Berdasarkan Tabel 1 terdapat 2 kelas minoritas, yaitu pada Y = 2 dan Y = 3. Pada kelas minoritas dilakukan replikasi mencari tetangga terdekat $x_{knn}$ dengan menggunkan jarak *Euclidean* untuk setiap data dalam kelas tersebut. Data *synthetics* dari kelas minoritas Y = 2 dibangkitkan dengan cara sebagai berikut.

Dari data ke-9 dan data ke-10

$$d\left(\begin{bmatrix}5\\7\\4\end{bmatrix}, \begin{bmatrix}5\\6\\5\end{bmatrix}\right) = \sqrt{(5-5)^2 + (7-6)^2 + (4-5)^2} = \sqrt{2}$$

Dari data ke-9 dan data ke-11

$$d\left(\begin{bmatrix}5\\7\\4\end{bmatrix}, \begin{bmatrix}5\\4\\5\end{bmatrix}\right) = \sqrt{(5-5)^2 + (7-4)^2 + (4-5)^2} = \sqrt{10}$$

Dari data ke-10 dan data ke-11

$$d\left(\begin{bmatrix}5\\6\\5\end{bmatrix}, \begin{bmatrix}5\\4\\5\end{bmatrix}\right) = \sqrt{(5-5)^2 + (6-4)^2 + (5-5)^2} = \sqrt{4}$$

Data *synthetics* dari kelas minoritas Y = 3, dibangkitkan dengan cara sebagai berikut.

Dari data ke-12 dan 13

$$d\left(\begin{bmatrix}7\\9\\8\end{bmatrix}, \begin{bmatrix}9\\8\\7\end{bmatrix}\right) = \sqrt{(7-9)^2 + (9-8)^2 + (8-7)^2} = \sqrt{6}$$

Dari perhitungan didapat dua jarak Euclidean yang terdekat pada kelas 2(Y = 2) yaitu $\sqrt{2}$ dan $\sqrt{4}$ maka kelas 2 akan dilakukan replikasi sebanyak dua kali. Jumlah data kelas 1 yang awalnya berjumlah 3 maka setelah dilakukan replikasi sebanyak 2 kali menjadi 9 data. Pada kelas 3 (Y = 3) terdapat satu jarak Euclidean, yaitu $\sqrt{6}$, kelas 2 dilakukan replikasi tiga kali. Jumlah data kelas 3 yang awalnya berjumlah 2 maka setelah dilakukan replikasi sebanyak 3 kali, jumlah data menjadi delapan data.

Tabel 1. Data asli Simulasi dan (*) data sintetik hasil SMOTE

| Data ke- | X1 | X2 | X3 | Y | Data ke- | X1 | X2 | X3 | Y | Data ke- | X1 | X2 | X3 | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 1,00 | 1,00 | 1 | 10 | 5,00 | 6,00 | 5,00 | 2 | 18* | 5,00 | 4,00 | 5,00 | 2 |
| 2 | 2,00 | 3,00 | 1,00 | 1 | 11 | 5,00 | 4,00 | 5,00 | 2 | 19* | 5,00 | 4.80 | 5,00 | 2 |
| 3 | 3,00 | 1,00 | 2,00 | 1 | 12 | 7,00 | 9,00 | 8,00 | 3 | 20* | 7,00 | 9,00 | 8,00 | 3 |
| 4 | 2,00 | 2,00 | 3,00 | 1 | 13 | 9,00 | 8,00 | 7,00 | 3 | 21* | 7,80 | 8,60 | 7,60 | 3 |
| 5 | 3,00 | 2,00 | 1,00 | 1 | 14* | 5,00 | 7,00 | 4,00 | 2 | 22* | 9,00 | 8,00 | 7,00 | 3 |
| 6 | 2,00 | 3,00 | 2,00 | 1 | 15* | 5,00 | 6,60 | 4,40 | 2 | 23* | 8,20 | 8,40 | 7,40 | 3 |
| 7 | 1,00 | 1,00 | 3,00 | 1 | 16* | 5,00 | 6,00 | 5,00 | 2 | 24* | 8,04 | 8,50 | 7,48 | 3 |
| 8 | 2,00 | 2,00 | 2,00 | 1 | 17* | 5,00 | 5,2 | 5,00 | 2 | 25* | 7,96 | 8,50 | 7,52 | 3 |
| 9 | 5,00 | 7,00 | 4,00 | 2 | | | | | | | | | | |

Cara menghitung data sintetik pada kelas 2, adalah :

$$x_{syn} = [5,7,4] + ([5,6,5] - [5,7,4]) \times 0,4 = [5; 6,6; 4,4]$$
$$x_{syn} = [5,6,5] + ([5,4,5] - [5,6,5]) \times 0,4 = [5; 5,2; 5]$$
$$x_{syn} = [5,4,5] + ([5,6,5] - [5,4,5]) \times 0,4 = [5; 4,8; 5]$$

Tabel 2 menunjukkan hasil distribusi dari data simulasi SMOTE.

Tabel 2. Distribusi Data Simulasi Sebelum dan setelah SMOTE

| Kelas mayor | Kelas minor | Replikasi | Kelas mayor | Kelas minor baru |
|---|---|---|---|---|
| 8 (61,54%) | 3 (23,08%) | 2 | 8 (32,00%) | 9 (36,00%) |
| | 2 (15,38%) | 3 | | 8 (32,00%) |

## HASIL DAN PEMBAHASAN

Data dalam penelitian ini diambil dari Kaggle.com. Dataset 1 terdiri dari 68 data yang dikelompokkan dalam 3 kelas, Dataset 2 terdiri dari 180 data yang dikelompokkan dalam 3 kelas, dan Dataset 3 terdiri dari 371 data yang dikelompokkan dalam 3 kelas. Hasil penyelesaian data *imbalanced* dengan menggunakan algoritma SMOTE ditunjukkan pada Tabel 3.

Tabel 3. Hasil pengujian algoritma SMOTE terhadap *dataset*

| Nama | Kelas mayor | Kelas minor | Replikasi | Kelas mayor | Kelas minor baru |
|---|---|---|---|---|---|
| Dataset 1 | 48 (70,60%) | 8 (11,80%) | 5 | 48 (33,30%) | 48 (33,30%) |
| | | 12 (17,60%) | 33 | | 48 (33,30%) |
| Dataset 2 | 120 (66,70%) | 20 (11,10%) | 5 | 120 (33,33%) | 120 (33,30%) |
| | | 40 (22,20%) | 2 | | 120 (33,30%) |
| Dataset 3 | 262 (70,60%) | 65 (17,50%) | 3 | 262 (33,30%) | 260 (33,10%) |
| | | 44 (11,90%) | 5 | | 264 (33,60%) |

## KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa penyelesaian permasalahan distribusi kelas yang tidak seimbang pada *dataset imbalanced*

dapat diselesaikan dengan cara melakukan pembangkitan pada kelas data minoritas dengan algoritma SMOTE.

## DAFTAR PUSTAKA

[1] S Fotouhi, S Asadi, and M W Kattan 2019 A Comprehensive Data Level Analysis for Cancer Diagnosis on Imbalanced Data, *Journal of Biomedical Informatics*, vol. 90, no. October 2017, p. 103089.

[2] H Sain and S W Purnami 2015 Combine Sampling Support Vector Machine for Imbalanced Data Classification, *Procedia Computer Science*, vol. 72, pp. 59–66.

[3] Q Gu, X M Wang, Z Wu, B Ning, and C S Xin 2016 An Improved SMOTE Algorithm Based on Genetic Algorithm for Imbalanced Data Classification, *Journal of Digital Information Management*, vol. 14, no. 2, pp. 92–103.

[4] C Jian, J Gao, and Y Ao 2016 A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble, *Neurocomputing*, vol. 193, pp. 115–122.

[5] G Haixiang, L Yijing, J Shang, G Mingyun, H Yuanyue, and G Bing 2017 Learning from Class-Imbalanced Data: Review of Methods and Applications, *Expert Systems with Applications*, vol. 73, pp. 220–239.

[6] B Krawczyk 2016 Learning from Imbalanced Data: Open Challenges and Future Directions, *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232.

[7] V García, J S Sánchez, and R A Mollineda 2012 On the Effectiveness of Preprocessing Methods When Dealing with Different Levels of Class Imbalance, *Knowledge-Based Systems*, vol. 25, no. 1, pp. 13–21.

[8] M Buda, A Maki, and M A Mazurowski 2018 A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks, *Neural Networks*, vol. 106, no. March, pp. 249–259.

[9] A Fernández, S García, M Galar, and R C Prati 2019 *Learning from Imbalanced Data Sets (2018, Springer International Publishing).pdf*. Berlin: Springer, 2019.

[10] L M El Bakrawy, M A Cifci, S Kausar, and S Hussain 2022 A Modified Ant Lion Optimization Method and Its Application for Instance Reduction Problem in Balanced and Imbalanced Data, no. February.

[11] J Gao, L Gong, J Y Wang, and Z C Mo 2019 Study on Unbalanced Binary Classification with Unknown Misclassification Costs, *IEEE International Conference on Industrial Engineering and Engineering Management*, vol. 2019-December, pp. 1538–1542.

[12] E M F El Houby, N I R Yassin, and S Omran 2017 A Hybrid Approach from Ant Colony Optimization and K-Nearest Neighbor for Classifying Datasets Using Selected Features, *Informatica (Slovenia)*, vol. 41, no. 4, pp. 495–506.

[13] A Nikmatul Kasanah, M Muladi, and U Pujianto 2017 Penerapan Teknik SMOTE Untuk Mengatasi Imbalance Class Dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN, *RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201.

[14] T Astuti, S P Adipurwoko, R Diyani, R A Santosa, and B Permadi 2018 Pengaruh Seleksi Fitur Dan SMOTE Terhadap Performa Klasifikasi Ranking Mobile Legends, *CITISEE*, no. ISBN: 978-602-60280-1-3, pp. 113–117.

[15] G AlMahadin, A Lotfi, M M Carthy, and P Breedon 2022 Enhanced Parkinson's Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques, *SN Computer Science*, vol. 3, no. 1, pp. 1–21.