Penanganan Data Ketidakseimbangan dalam Pendekatan SMOTE Guna Meningkatkan akurasi Algoritma K-NN

[1] Oktriana Siboro, [2] Yunita Pricilia Banjarnahor, [3] Anita Gultom, [4] Novriadi Antonius Siagian, [5] Parasian DP. Silitonga

[1] Universitas Katolik Santo Thomas medan, Fakultas Ilmu Komputer, Jl. Setiabudi,Kampung Tengah, Kec. Medan Tuntungan, Kota Medan, Sumatera Utara, Indonesia

E-Mail: oktrianasiboro@gmail. com[1], yunitabanjarnahor03@gmail.com[2], anitagultom048@gmail. com[3], novriadi.antonius95@gmail.com[4], parasianirene@gmail.com[5]

Abstrak

Klasifikasi data tidak seimbang merupakan masalah yang sering pada bidang *machine learning* dan data *mining*. Pada penelitian ini, diterapkan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk mengatasi permasalahan ketidakseimbangan kelas pada dataset. Berfokus pada algoritma K-Nearest Neighbors (K-NN) dan menganalisis peningkatan akurasi setelah implementasi SMOTE. Data yang direkam sebanyak 8.545 data, jumlah atribut sebanyak 6 atribut dan jumlah kelas sebanyak 2 kelas. Hasil penelitian ini menunjukkan akurasi dengan teknik SMOTE meningkat dibanding tanpa menggunakan SMOTE, misalnya dengan K = 11 akurasi algoritma K-NN teknik SMOTE sebesar 0, 8742 lebih tinggi dibandingkan akurasi algoritma K-NN tanpa SMOTE sebesar 0,8683. Hal ini menunjukkan bahwa penggunaan SMOTE dapat menjadi solusi efektif untuk meningkatkan akuasi algoritma K-NN pada dataset yang tidak seimbang. Oleh karena itu, penelitian ini menyimpulkan bahwa penerapan SMOTE dapat meningkatkan akurasi algoritma K-NN pada data tidak seimbang.

Kata Kunci: Akurasi; Data Tidak Seimbang; Klasifikasi; K-Nearest Neighbor; Smote.

Abstract

Classification of unbalanced data is a frequent problem in the field of machine learning and data mining. In this research, SMOTE (Synthetic Minority Oversampling Technique) is applied to solve the problem of class imbalance in the dataset. It focuses on K-Nearest Neighbors (K-NN) algorithm and analyzes the accuracy improvement after SMOTE implementation. The data recorded was 8,545 data, the number of attributes was 6 attributes and the number of classes was 2 classes. The results of this study show that the accuracy with SMOTE technique increases compared to without using SMOTE, for example with K = 11 the accuracy of the K-NN algorithm with SMOTE technique is 0.8741 higher than the accuracy of the K-NN algorithm without SMOTE of 0.8683. This shows that the use of SMOTE can be an effective solution to improve the accuracy of the K-NN algorithm on unbalanced datasets. Therefore, this study concludes that the application of SMOTE can improve the accuracy of the K-NN algorithm on unbalanced data.

Keywords: Accuracy; Imbalance Data; Classification; K-Nearest Neighbor; Smote.

PENDAHULUAN

Klasifikasi adalah proses pengelompokan objek atau data kedalam ketegori atau kelas berdasarkan atribut dari suatu dataset. Tujuan utama dari klasifikasi adalah untuk membangun model atau aturan yang digunakan untuk memprediksi kelas atau label dari data berdasarkan informasi yang telah di ketahui.

Salah satu masalah dalam machine learning dan data mining adalah ketidakseimbangan data. Data tidak seimbang merupakan suatu keadaan dimana distribusi kelas atau label di dalam dataset tidak merata, jumlah sampel dalam satu kelas lebih besar atau lebih kecil daripada kelas lainnya. Dalam kelas data tidak seimbang, kelas dengan jumlah yang lebih besar disebut kelas mayoritas dan kelas dengan jumlah yang lebih kecil sebagai kelas minoritas *(R. Siringoringo, 2018)* [1]. Salah satu teknik yang dapat mengatasi ketidakseimbangan data dengan menggunakan teknik oversampling seperti *Synthetic Minority Over-sampling Technique* (SMOTE). Metode ini cocok digunakan untuk data yang besar atau scale data. Teknik SMOTE bekerja dengan memperbanyak data minoritas sebanyak data mayoritas sehingga menyebabkan keseimbangan kelas dalam dataset. SMOTE menciptakan sampel sintetis dari kelas minoritas, sehingga menciptakan keseimbangan antara kelas dalam dataset *(A. N. Kasanah, et al., 2019)* [2].

Ketidakseimbangan data dapat berdampak signifikan pada kinerja algoritma machine learning, terutama algoritma klasifikasi. Algoritma K-Nearest Neighbors (KNN), sebagai contoh algoritma pembelajaran berbasis instansi (instance-based learning), melakukan prediksi berdasarkan kesamaan (similaritas) dengan tetangga terdekat dalam ruang fitur (K). Oleh karena itu, kinerja KNN sangat dipengaruhi oleh representasi data dan metrik jarak yang digunakan.

Akurasi adalah salah satu metrik evaluasi umum yang digunakan untuk mengukur kinerja model machine learning, termasuk metode K-Nearest Neighbors (KNN). Akurasi mengukur sejauh mana model dapat memprediksi dengan benar pada seluruh data uji yang digunakan. Berikut adalah rumus dari accuracy dataset:

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Keterangan:
TP (True Positive): Jumlah prediksi positif yang benar. TN (True Negative): Jumlah prediksi negatif yang benar.
FP (False Positive): Jumlah prediksi negatif palsu (sebenarnya positif, tetapi diprediksi negatif). FN (False Negative): Jumlah prediksi positif palsu (sebenarnya negatif, tetapi diprediksi positif).

1.1 Tujuan Penelitian

Penelitian ini bertujuan untuk menangani ketidakseimbangan data dalam pendekatan SMOTE guna meningkatkan akurasi algortima K- NN (K-Nearest Neighbors)

BAHAN DAN METODE

1. Dataset
   Dataset yang digunakan adalah dataset anemia yang bersumber dari *kaggle*.com. Dataset ini memiliki jumlah 8545 data, dengan jumlah 2 kelas, dimana kelas mayoritas (0) memiliki jumlah 5640 dan jumlah kelas minoritas (1) sebanyak 2904.

2. SMOTE (Synthetic Minority Over-sampling Technique)
   SMOTE adalah teknik oversampling untuk menangani ketidakseimbangan kelas dalam dataset. Adapun cara kerja SMOTE:
   - Buat sampel sintetis baru dari kelas minoritas berdasarkan tetangga terdekatnya dalam ruangfitur
   - Meningkatkan distribusi kelas dalam dataset menjadi lebih seimbang
   - Meningkatkan kinerja model pada kelas minoritas tanpa kehilangan infomasi dari kelas mayoritas
3. Imbalance Data
   Imbalance data (data tidak seimbang) adalah suatu keadaan dimana distribusi kelas atau label di dalam dataset tidak merata, jumlah sampel yang lebih besar disebut kelas mayoritas dan kelas yang lebih kecil disebut minoritas
4. Algoritma K-NN (K-Nearest Neighbor)
   Algoritma K-NN adalah algoritma dalam *machine learning* yang digunakan untuk masalah klasifikasi dan regresi. Algoritma ini akan memprediksi kelas atau nilai data baru berdasarkan mayoritastetangga terdekatnya.

Adapun cara kerja algoritma K-NN:

- Tentukan nilai K: jumlah tetangga terdekat yang dianalisis
- Hitung jarak: jarak antara data baru dengan data di dataset
- Temukan tetangga terdekat: K tetangga terdekat berdasarkan jarak
- Prediksi Klasifikasi: kelas terbanyak diantara K tetangga terdekat

HASIL DAN PEMBAHASAN

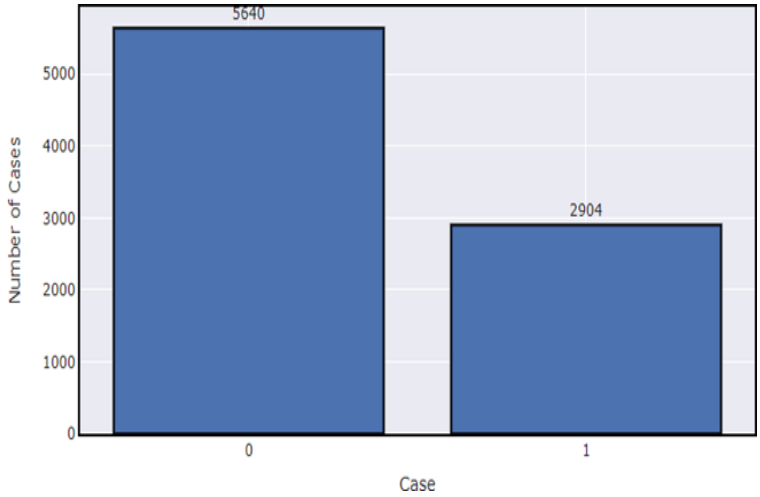
Pada penelitian ini, dengan menggunakan dataset anemia dari *kaggle.com*, Data yang digunakan adalahsebagai berikut:

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
0	0	70	115	39	271	0
1	0	37	126	30	332	1
2	0	90	127	48	283	0
3	0	47	96	45	199	0
4	1	63	122	35	306	0
...	...	...	...	...	...	...
8539	0	45	94	38	214	0
8540	0	67	109	34	267	0
8541	1	40	105	55	205	1
8542	1	40	108	33	269	1
8543	0	55	88	42	185	0

8544 rows × 6 columns

Tabel 1. Dataset pengujian

Dataset awal terdapat sebanyak 2904 positif anemia (1), dan sebanyak 5641 tidak menderita anemia (0).Berikut adalah grafik dataset:



Gambar 1. Grafik ketidakseimbangan awal

Membagi data latih dan data uji, dengan jumlah data sebanyak 8544 maka data latih senilai 80% dari 8544 yaitu sebanyak 6835 dan untuk data uji senilai 20% dari 8544 yaitu sebanyak 1709. Dari data latih maka dihitungbanyaknya label 1 dan label 0 adalah:

Untuk K- NN Convensional, dengan K = 1 maka tingkat akurasinya sebanyak:

```
KNeigbors :
[[1005  130]
 [ 128  446]]
Accuracy Score:  0.8490345231129316

K-Fold Validation Mean Accuracy: 83.83 %

Standard Deviation: 1.68 %

ROC AUC Score: 0.83 %

Precision: 0.77 %

Recall: 0.78 %

F1 Score: 0.78 %
```

Gambar 2. Hasil Akurasi K=1 K-NN convensional

Untuk menangani ketidakseimbangan data, maka dilakukan algoritma K-NN dengan teknik SMOTE.Algoritma K-NN dengan teknik SMOTE menyeimbangkan kelas mayoritas dengan kelas minoritas.

```
After OverSampling, counts of label 1: 4505
After OverSampling, counts of label 0: 4505
```

Setelah data sudah seimbang, maka dihasilkan tingkat akurasi algoritma K -NN dengan teknik SMOTE,dimana K=1:

```
KNeigbors :
[[996 139]
 [107 467]]
Accuracy Score:  0.8560561732007022

K-Fold Validation Mean Accuracy: 89.84 %

Standard Deviation: 1.67 %

ROC AUC Score: 0.85 %

Precision: 0.77 %

Recall: 0.81 %

F1 Score: 0.79 %
```

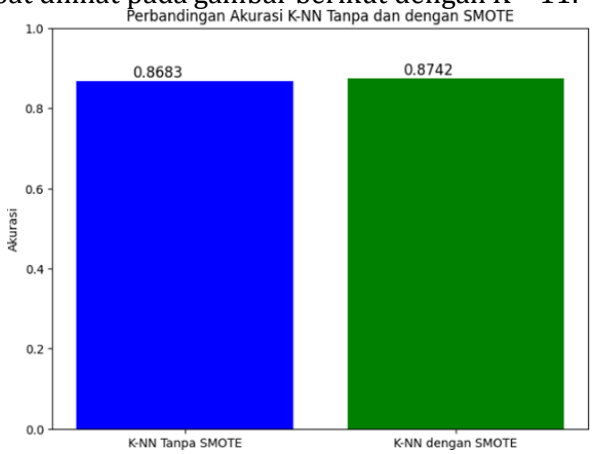
Gambar 3. Hasil Akurasi K=1 K-NN + SMOTE

Maka dilakukan analisis hasil perbandingan akurasi algoritma K-NN convensional dengan algoritma K-NNteknik SMOTE:

Tabel 2. Hasil Accuracy K-NN + SMOTE dan K-NN

k	K-NN	K-NN + SMOTE
1	0,8490	0,8561
3	0,8660	0,8672
5	0,8601	0,8648
7	0,8660	0,8648
9	0,8630	0,8689
11	0,8683	0,8742

Grafik perbandingan dapat dilihat pada gambar berikut dengan K = 11:



Gambar 4. Grafik hasil perbandingan dengan K = 11

Akurasi dataset algoritma K-NN dengan teknik SMOTE lebih tinggi dibandingkan algoritma K-NN convensional. Dengan menerapkan teknik SMOTE dapat menangani ketidakseimbangan data danmeningkatkan akurasi algoritma K-NN

KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa:

(1) Dengan menggunakan metode KNN + SMOTE, pengujian menunjukkan bahwa algoritma tersebut dapat menghasilkan akurasi yang lebih baik dalam mengatasi

masalah ketidakseimbangan kelas pada dataset. Hasil pengujian menunjukkan peningkatan signifikan dalam akurasi model, terutama untuk kelas minoritas. Penerapan metode SMOTE membantu dalam menghasilkan sampel sintetis untuk menyeimbangkan data. Akurasi dataset algoritma K-NN dengan teknik SMOTE lebih tinggi dibandingkan algoritma K-NN convensional.

(2) Setelah di lakukan pengujian, dengan nilai K1-K11 terdapat akurasi yg rendah yaitu di k7, tetapi dik1,3,5,9,11 akurasinya sangat baik, sehingga untuk penelitian ke depannya bisa mengoptimasi nilai k7, KNN + smote memiliki akurasi yang baik. Penelitian selanjutnya dapat lebih memperdalam analisis terhadap penyebab akurasi rendah pada K=7, mungkin dengan mengeksplorasi faktor-faktor seperti pola atau struktur data. Kemungkinan terdapat pola khusus dalam data yang membuat model kurang efektif pada K=7.

(3) Kelemahan utama KNN adalah kebutuhan untuk menentukan nilai K yang optimal. Pemilihan nilai K yang salah dapat mempengaruhi kinerja model secara signifikan.

UCAPAN TERIMAKASIH

Kami ingin mengucapkan terima kasih yang sebesar-besarnya atas publikasi jurnal yang mengulas “"Penerapan Metode SMOTE untuk Analisis Perbandingan Akurasi Data Tidak Seimbang Menggunakan Algoritma K-NN"”. Kami juga ingin menyampaikan terimakasih kepada dosen pembimbing yang telah memberikan bimbingan yang berharga serta dukungan yang tak ternilai selama proses penelitian ini. Keberhasilan kami adalah hasil dari kolaborasi dan semangat tim yangluar biasa, yang telah mendorong kami melewati setiap tantangan dan meraih pencapaian yang signifikan dalam bidang ini. Tanpa bantuan dan dorongan dari semua pihak yang terlibat, pencapaian kami dalam menganalisis Perbandingan Akurasi Data Tidak Seimbang tidak akan berjalan dengan baik.

DAFTAR PUSTAKA

[1] Chao-Ren Wang and Xin-Xue Zhao, "An Improving Majority Weighted Minority Oversampling technique for Imbalanced Classification Problem*." IEEE Access*, vol. 8, no. 1, pp. 14773-14783, 2020.

[2] Kasanah, A. N., Muladi, M., and Pujianto, U., "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196-201, 2019.

[3] R. Siringoringo, "Klasifikasi data tidak Seimbang menggunakan algoritma SMOTE dan k-nearestneighbor*," Journal Information System Development (ISD)*, vol. 3, no. 1, 2018.

[4] Siagian, N. A. (2021). Analisis Perbandingan Akurasi dalam Mengidentifikasi Jenis Kaca. **InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan**, 5(2), 283-294.