Implementasi SMOTE untuk mengatasi *Imbalance Class* pada Klasifikasi *Car Evolution* menggunakan K-NN

Femi Dwi Astuti*[1], Febri Nova Lenti*[2]
[1,2] Informatika, STMIK AKAKOM, Yogyakarta
e-mail: ***[1]femi@akakom.ac.id**, [2]febri@akakom.ac.id

Abstrak

Permasalahan ketidakseimbangan kelas akan terus ada karena data tidak dapat dipaksa untuk selalu seimbang. Ketidakseimbangan kelas memberikan dampak yang tidak baik pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas. Hal ini dapat menurunkan nilai accuracy hasil klasifikasi. SMOTE merupakan salah satu turunan teknik over-sampling untuk menanggulangi ketidakseimbangan kelas dengan menyeimbangkan dataset dengan meningkatkan ukuran kelas minor. SMOTE diterapkan pada klasifikasi dataset car evolution menggunakan algoritma klasifikasi KNN. Hasil klasifikasi dievaluasi akurasinya menggunakan 10fold-cross validation dengan membandingkan hasil klasifikasi yang hanya menggunakan KNN dan menggunakan KNN dan SMOTE. Hasil penelitian menunjukkan bahwa penggunaan SMOTE mampu mengatasi imbalance class dengan menaikkan nilai akurasi rata-rata sebesar 9.97%. Semakin kecil nilai k pada klasifikasi K-NN maka semakin besar tingkat akurasinya Dari hasil uji dengan k=3, k=5 dan k=10, maka akurasi klasifikasi tertinggi K-NN dengan k=3 menggunakan SMOTE sebesar 93.11%.

Kata kunci*—klasifikasi,K-NN,SMOTE,Imbalance Class*

Abstract

The problem of imbalance class will continue to exist because the data cannot be forced to always balance. Imbalance class has an unfavorable impact on the classification results where the minority class is often misclassified as the majority class. This can lower the accuracy value of the classification results. SMOTE is a derivative of the over-sampling technique to overcome class imbalances by balancing the dataset by increasing the size of the minor class. SMOTE is applied to the classification of the car evolution dataset using the KNN classification algorithm. The classification results are evaluated for their accuracy using 10fold-cross validation by comparing the classification results using only KNN and using KNN and SMOTE. The results showed that the use of SMOTE was able to overcome the imbalance class by increasing the average accuracy value by 9.97%. The smaller the k value in the K-NN classification, the greater the level of accuracy. From the test results with k = 3, k = 5 and k = 10, the highest K-NN classification accuracy with k = 3 uses SMOTE of 93.11%

Keywords— *classification,K-NN,SMOTE,Imbalance Class*

1. PENDAHULUAN

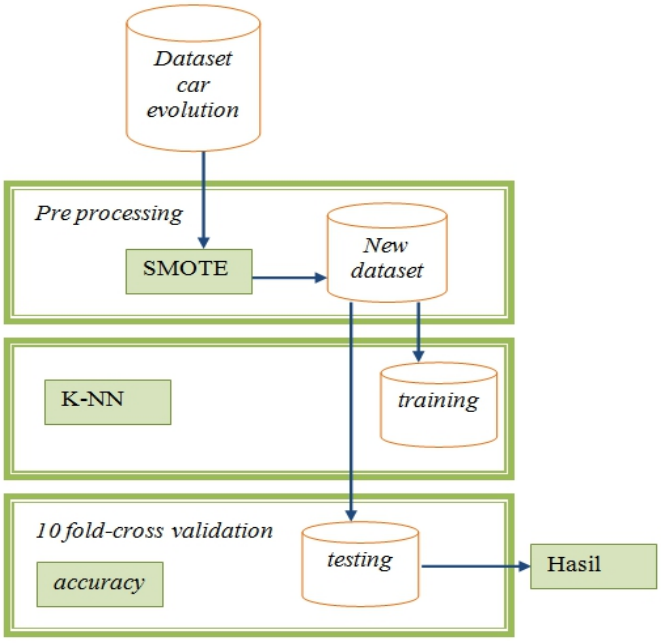
Kelas yang tidak seimbang *(imbalanced class)* merupakan kondisi dimana terdapat dataset yang jumlah kelas datanya tidak merata atau bisa diartikan adanya perbedaan yang signifikan terhadap jumlah kelas. Jumlah kelas ada yang mayor (jumlah data signifikan banyak) dan ada yang minor (jumlah data signifikan sedikit) [1]. Kelas yang tidak seimbang dapat memberikan dampak yang tidak baik pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas[2] karena secara teori mayoritas classifier

mengamsusikan distribusi yang relative seimbang [3]. Nilai accuracy hasil klasifikasi bisa menjadi kecil apabila menggunakan dataset yang tidak seimbang kelasnya. Selain permasalahan ketidakseimbangan kelas, permasalahan lain yang sering muncul adalah jumlah atribut yang banyak pada dataset. Jumlah atribut yang banyak pada dataset dapat mempengaruhi hasil performa klasifikasi. Berdasarkan permasalahan tersebut maka penelitian ini mencoba menggunakan salah satu teknik untuk menangani ketidakseimbangan kelas yaitu over-sampling dengan menggunakan SMOTE *(Synthetic Minority Over-sampling Technique)*. SMOTE merupakan salah satu algoritma turunan dari *over-sampling*. Penggunaan SMOTE ini dapat meningkatkan nilai sensitivitas lebih dari 50% walaupun nilai akurasi dan spesifitasnya menurun dibandingkan dengan pemodelan sebelum SMOTE [4]. SMOTE dapat menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan yang tanpa SMOTE [5][6]. Proses evaluasi dilakukan untuk melihat pengaruh penggunaan SMOTE dalam mengatasi ketidakseimbangan kelas sebelum proses klasifikasi dengan K-NN. K-NN digunakan karena merupakan metode klasifikasi yang populer tetapi belum ada kemampuan untuk menangani data yang tidak seimbang [5].

2. METODE PENELITIAN

2.1 Alur Penelitian

Alur penelitian untuk mengimplementasikan algoritma SMOTE untuk ketidak seimbangan Kelas serta penggunaan algoritma K-NN untuk klasifikasi dapat dilihat pada Gambar 1.



Gambar 1 Alur Penelitian

Pada Gambar 1 dapat dilihat alur penelitian yang akan dilakukan, dimulai dengan dataset yang sudah diambil sampai keluar hasilnya. Tahap pertama yang dilakukan dalam penelitian ini yaitu tahap pengumpulan data sebagai dataset. Dataset yang digunakan adalah data *public car evolution* yang diambil dari *UCI machine learning repository*. Setelah dataset ditentukan,

kemudian dilakukan *pre processing* untuk mengatasi ketidakseimbangan kelas menggunakan teknik SMOTE. Dataset yang digunakan untuk proses klasifikasi adalah dataset yang baru setelah dilakukan *pre processing.* Dataset hasil *pre processing (new dataset*) selanjutnya akan diklasifikasikan menggunakan metode KNN. Evaluasi/pengujian penelitian dilakukan menggunakan 10-*fold cross validation* untuk menghitung nilai *accuracy*. Pengujian dilakukan dengan menggunakan nilai k=3, k=5 dan k=10 untuk klasifikasi KNN nya, kemudian dibandingkan dengan menggunakan nilai k=3, k=5 dan k=10 untuk klasifikasi KNN dengan diawali penggunaan teknik SMOTE untuk menangani *imbalance class* nya. Indikator ketercapaian dalam penelitian ini diawali dengan ditunjukkannya hasil *accuracy* yang berbeda antara klasifikasi dengan KNN saja dan *accuracy* KNN dengan SMOTE.

2.2 K-NN

Salah satu metode klasifikasi sederhana yaitu metode *K-Nearest Neighbor* (K-NN) Metode K-NN merupakan *Instance Based Learning*, metode ini melakukan klasifikasi terhadap objek berdasarkan jarak antara objek tersebut dengan objek lain[7]. Metode K-NN menggunakan prinsip ketetanggaan *(neighbor)* untuk memprediksi kelas yang baru. Jumlah tetangga yang dipakai adalah sebanyak *k* tetangga. Perhitungan jarak dapat dilakukan dengan menggunakan persamaan (1), sedangkan algoritma perhitungan K-NN sebagai berikut :

1. Menentukan k sebagai banyaknya jumlah tetangga terdekat dengan objek baru.
2. Menghitung jarak antar objek/data baru terhadap semua objek/data yang telah di training.
3. Mengurutkan hasil perhitungan.
4. Menentukan tetangga terdekat berdasarkan jarak minimum ke k.
5. Menentukan kategori dari tetangga terdekat dengan objek/data.
6. Menggunakan kategori mayoritas sebagai klasifikasi objek/data baru.

2. 3 Ketidak seimbangan Kelas (Imbalanced Class)

Imbalanced class dapat diartikan sebagai adanya rasio yang tidak proporsional di setiap kelas di dalam dataset. Macam-macam algoritma *imbalanced class* yaitu :

a. Under sampling
Menyeimbangkan dataset dengan mengurangi ukuran kelas mayor. Metode ini digunakan ketika jumlah data mencukupi. Dengan menjaga semua sampel di kelas minor dan secara acak memilih jumlah sampel yang sama di kelas mayor, dataset baru yang seimbang dapat diambil untuk pemodelan lebih lanjut.

b. Over sampling
Metode ini digunakan jika jumlah data tidak mencukupi. Menyeimbangkan dataset dengan meningkatkan ukuran kelas minor. Salah satu teknik turunan dari *over sampling* yaitu SMOTE *(Syntetic Minority Oversampling Technique).* Teknik ini juga menambah jumlah kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data baru (data sintetis).

3. HASIL DAN PEMBAHASAN

Data yang dipakai dalam penelitian ini adalah data *car evolution* yang diambil dari UCI *machine learning repository*. Data *car evolution* ini memiliki 1728 *record* data. Kriteria yang digunakan sebanyak 6 kriteria meliputi :

1. Buying price
Nilai atribut yang mungkin untuk kriteria buying price yaitu : v-high, high, med, low

2. *Price of maintance*
   Nilai atribut yang mungkin untuk kriteria price of maintance yaitu : v-high, high, med, low
3. *Doors*
   Nilai atribut yang mungkin untuk kriteria doors yaitu : 2, 3, 4, 5-more
4. *Person capacity*
   Nilai atribut yang mungkin untuk kriteria persons capacity yaitu : 2, 4, more
5. *Lug boot*
   Nilai atribut yang mungkin untuk kriteria lug boot yaitu : small, med, big
6. *Safety*
   Nilai atribut yang mungkin untuk kriteria safety yaitu : low, med, high

Distribusi kelas data dari data car evolution ini meliputi :

1. Unacc sebanyak 1210 (70%)
2. Acc sebanyak 384 (22,22%)
3. Good sebanyak 69 (3,993%)
4. V-good sebanyak 65 (3,762%)

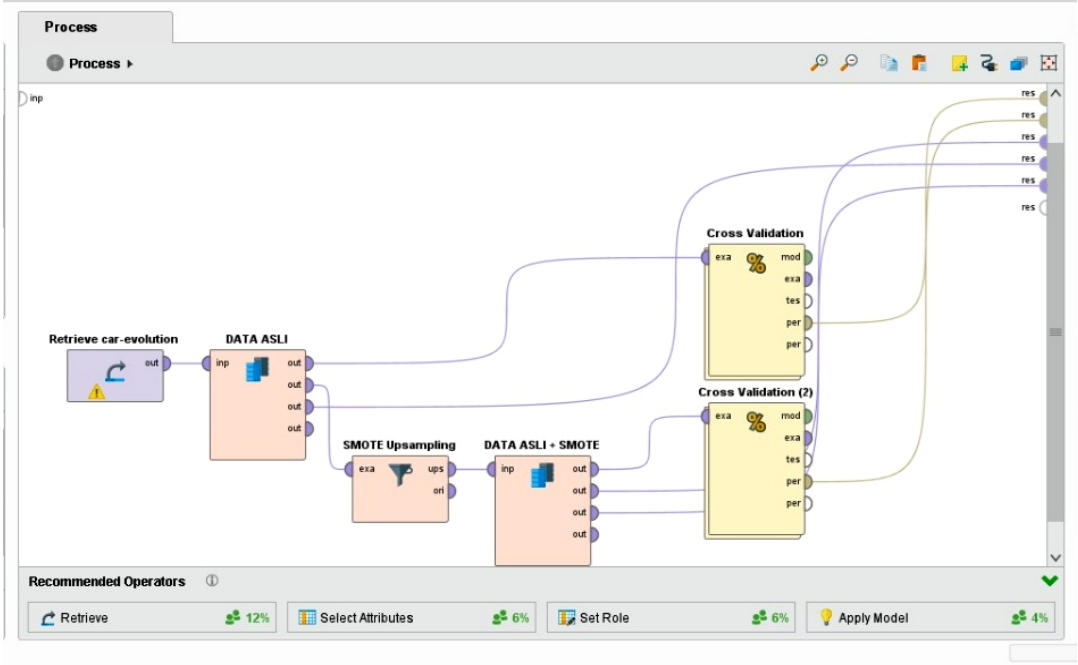
Berdasarkan data distribusi kelas, dapat dilihat bahwa terdapat ketidakseimbangan kelas untuk dataset car evolution kelas unacc dapat mencapai angka 70% sedangkan v-good hanya 3,762%. Contoh data yang digunakan pada penelitian ini dapat dilihat pada Tabel 4.1

Tabel 1 *Sample* Data car evolution

Buying Price	Price Of Maintance	Doors	Person Capacity	Lug_Boot	Safety	Class
vhigh	vhigh	2	2	small	low	unacc
vhigh	vhigh	2	2	small	med	unacc
vhigh	vhigh	2	2	small	high	unacc
vhigh	vhigh	2	2	med	low	unacc
vhigh	vhigh	2	2	med	med	unacc
vhigh	vhigh	2	2	med	high	unacc
vhigh	med	4	4	big	med	acc
vhigh	med	4	4	big	high	acc
…	…	…	…	…	…	…
…	…	…	…	…	…	…
med	med	2	4	big	high	vgood

3.1 Implementasi

Penelitian dilakukan dengan menggunakan tools rapid miner machine learning untuk melihat akurasi dari penggunaan smote untuk mengatasi imbalance class pada klasifikasi K-NN. Klasifikasi K-NN dengan SMOTE menggunakan rapid miner dapat dilihat pada Gambar 4.1.



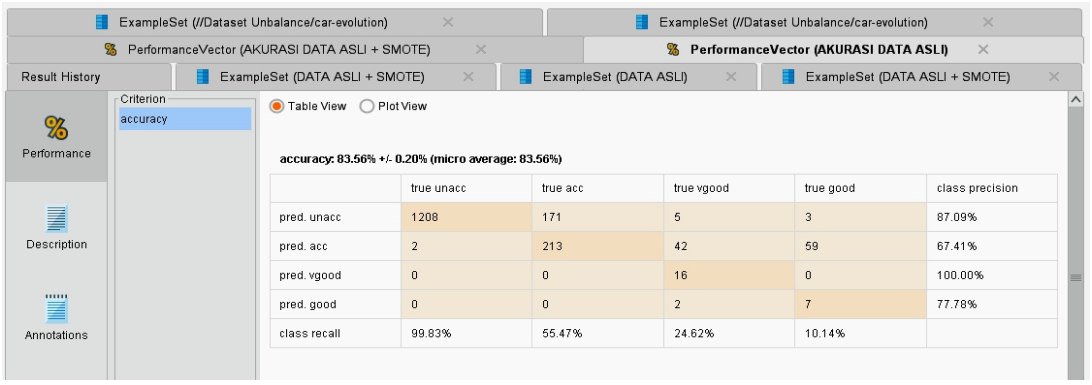
Gambar 2 Klasifikasi K-NN dengan SMOTE menggunakan rapid miner

Implementasi di rapid miner dilakukan sekaligus untuk membandingkan antara data asli yang di klasifikasi dengan K-NN dibandingkan dengan data asli yang diterapkan SMOTE, hasil data yang sudah di terapkan SMOTE baru diklasifikasi dengan K-NN.

3.2 Pengujian

Pengujian penelitian dilakukan dengan menggunakan *cross validation* untuk melihat nilai *accuracy*. Pengujian klasifikasi dilakukan dengan menggunakan nilai k=3, k=5 dan k=10. Hasil pengujian smote pada K-NN dengan k=3, k=5 dan k=10 dapat dilihat pada Gambar 4.1 sampai Gambar 4.6.

3.2.1 Pengujian K-NN dengan k=3

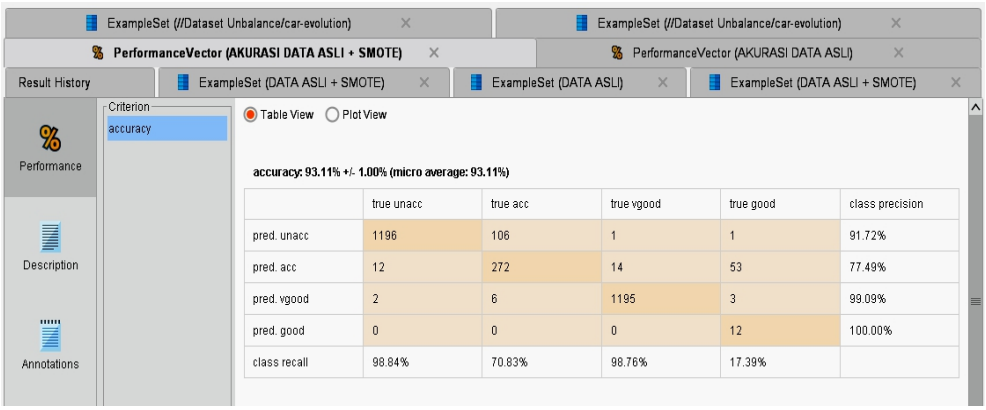


Gambar 3 K-NN dengan k=3

Gambar 3 menunjukkan hasil pengujian *accuracy* terhadap hasil klasifikasi K-NN tanpa menggunakan SMOTE untuk nilai k=3. Dapat dilihat bahwa nilai accuracy sebesar 83.56%. class recall untuk unacc sebesar 99.83%, acc sebesar 55.47%, vgood sebesar

24.62% dan good sebesar 10.14%. Sedangkan *class precision* untuk unacc sebesar 87.09%, acc sebesar 67.41%, vgood sebesar 100.00% dan good sebesar 77.78%.

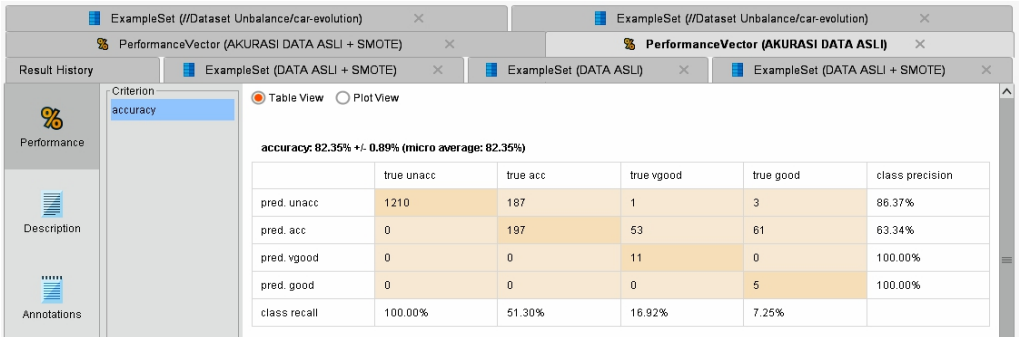
3.2.2 Pengujian SMOTE dan K-NN dengan k=3



Gambar 4 SMOTE dan K-NN dengan k=3

Gambar 4.3 menunjukkan hasil pengujian accuracy terhadap penggunaan smote pada klasifikasi K-NN untuk nilai k=3. Dapat dilihat bahwa nilai accuracy termasuk tinggi yaitu sebesar 93.11%. class recall untuk unacc sebesar 98.84%, acc sebesar 70.83%, vgood sebesar 98.76% dan good sebesar 17.39%. Sedangkan class precision untuk unacc sebesar 91.72%, acc sebesar 77.49%, vgood sebesar 99.09% dan good sebesar 100.00%. berdasarkan hasil ini dapat dilihat bahwa penggunaan SMOTE untuk mengatasi imbalance class dapat menaikkan nilai accurasy dari hasil klasifikasi sampai sebesar 9.55%.

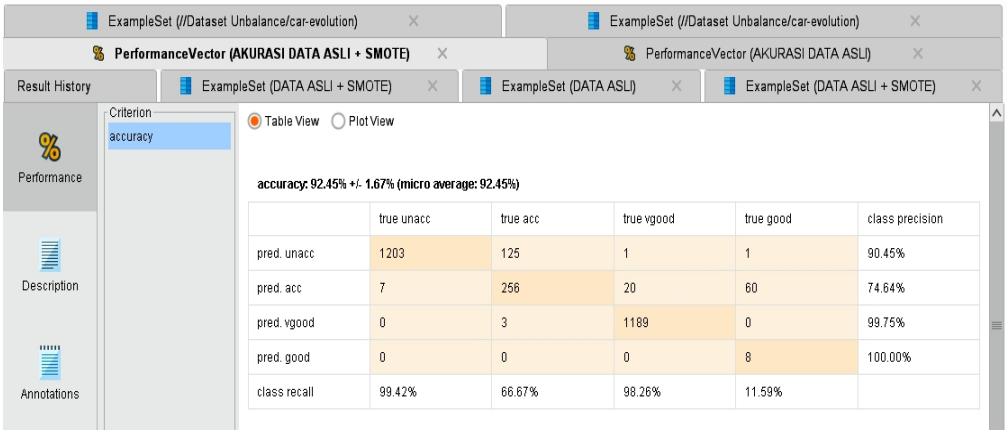
3.2.3 Pengujian K-NN dengan k=5



Gambar 5 K-NN dengan k=5

Gambar 5 menunjukkan hasil pengujian *accuracy* terhadap hasil klasifikasi K-NN tanpa menggunakan SMOTE untuk nilai k=5. Dapat dilihat bahwa nilai *accuracy* sebesar 82.35%. class recall untuk unacc sebesar 100.00%, acc sebesar 51.30%, vgood sebesar 16.92% dan good sebesar 7.25%. Sedangkan class precision untuk unacc sebesar 86.37%, acc sebesar 63.34%, vgood sebesar 100.00% dan good sebesar 100.00%. berdasarkan hasil tersebut dapat disimpulkan bahwa dengan nilai K lebih besar maka *accuracy* cenderung lebih kecil.

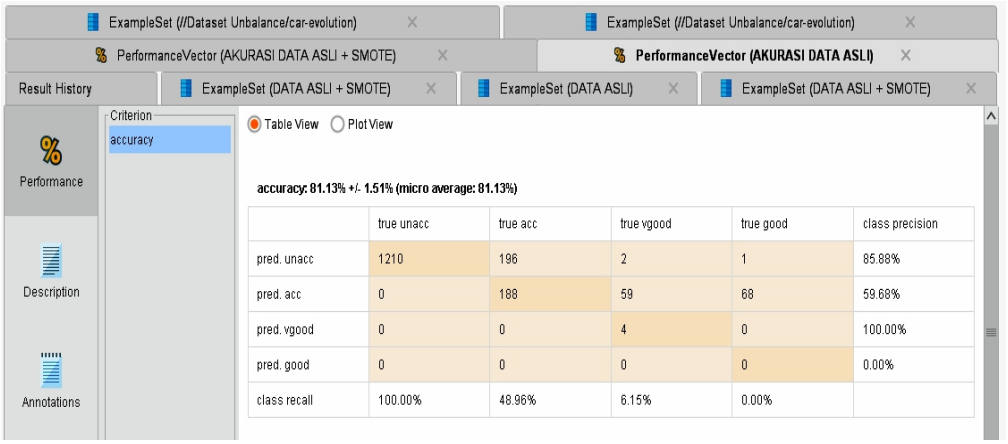
3.2.4 Pengujian SMOTE dan K-NN dengan k=5



Gambar 6 SMOTE dan K-NN dengan k=5

Gambar 6 menunjukkan hasil pengujian accuracy terhadap penggunaan smote pada klasifikasi K-NN untuk nilai k=5. Dapat dilihat bahwa untuk penggunaan SMOTE untuk nilai k yang berbeda, maka hasil *accuracy* juga berbeda. Nilai *accuracy* turun dari k=3, nilai *accuracy* k=3 sebesar 93.11% sedangkan k=5 hanya mencapai 92.45% sehingga penurunan sebesar 0.66%. Sedangkan jika sama-sama dibandingkan dengan k=5 tetapi tanpa menggunakan SMOTE, dapat dilihat bahwa penggunaan SMOTE untuk mengatasi imbalance class dapat menaikkan nilai *accuracy* dari hasil klasifikasi sampai sebesar 10.1%. kenaikan nilai ini lebih tinggi jika dibandingkan kenaikan k=3 tanpa SMOTE dan dengan SMOTE.

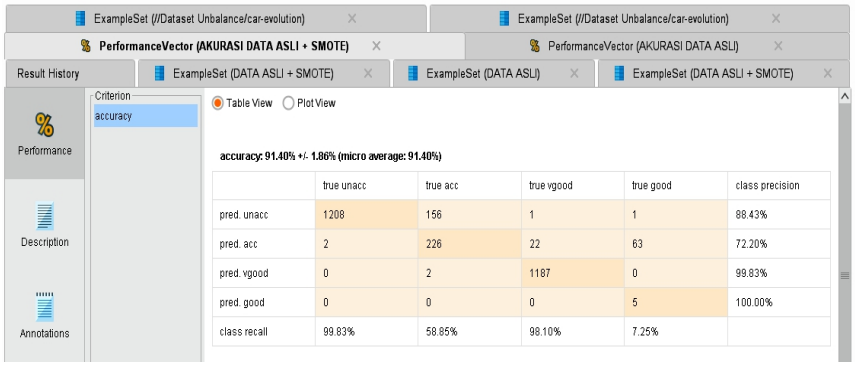
3.2.5 Pengujian K-NN dengan k=10



Gambar 7 K-NN dengan k=10

Gambar 7 menunjukkan hasil pengujian accuracy terhadap hasil klasifikasi K-NN tanpa menggunakan SMOTE untuk nilai k=10. Dapat dilihat bahwa nilai *accuracy* sebesar 81.13%, nilai ini turun jika dibandingkan dengan k=3 maupun k=5. Berdasarkan hasil tersebut dapat disimpulkan bahwa dengan nilai K lebih besar maka *accuracy* cenderung lebih kecil.

3.2.6 Pengujian SMOTE dan K-NN dengan k=10



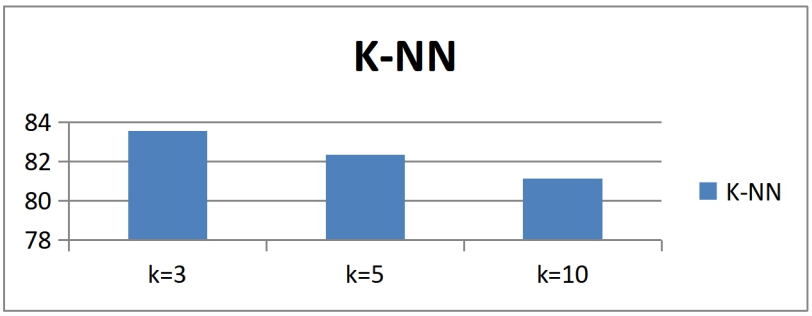
Gambar 8 SMOTE dan K-NN dengan k=10

Gambar 8 menunjukkan hasil pengujian *accuracy* terhadap penggunaan smote pada klasifikasi K-NN untuk nilai k=10. Dapat dilihat bahwa untuk penggunaan SMOTE untuk nilai k yang berbeda, maka hasil *accuracy* juga berbeda. Nilai *accuracy* turun dari k=5, nilai *accuracy* k=5 sebesar 92.45% sedangkan k=10 hanya mencapai 91.40% sehingga penurunan sebesar 1.05%. Sedangkan jika sama-sama dibandingkan dengan k=10 tetapi tanpa menggunakan SMOTE, dapat dilihat bahwa penggunaan SMOTE mampu menaikkan nilai *accuracy* dari hasil klasifikasi sampai sebesar 10.27%. kenaikan nilai ini lebih tinggi jika dibandingkan kenaikan k=3 maupun k=5 tanpa SMOTE dan dengan SMOTE. Untuk melihat tabel perbandingan penggunaan K-NN dengan SMOTE dan K-NN untuk k=3, k=5 dan k=10 dapat dilihat pada tabel 2.

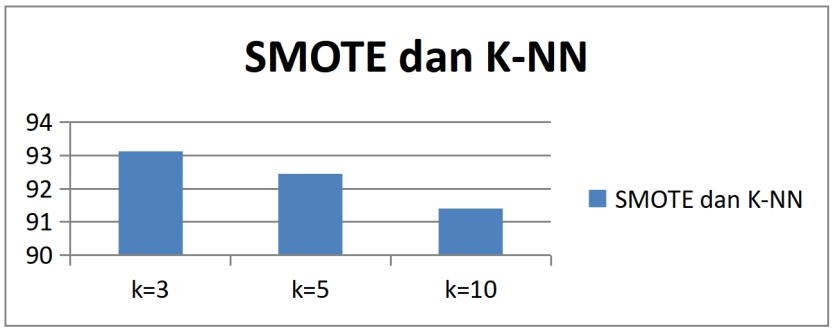
Tabel 2 perbandingan hasil akurasi

Metode	Akurasi (%)		
	k=3	k=5	k=10
K-NN	83.56	82.35	81.13
SMOTE dan K-NN	93.11	92.45	91.40
Kenaikan akurasi	9.55	10.1	10.27
Rata-rata kenaikan	**9.97**		

Tabel 2 menunjukkan hasil perbandingan hasil accuracy menggunakan K-NN dan hasil accuracy penggunaan SMOTE sebelum menggunakan K-NN. Gambar 9 menunjukkan grafik perbandingan nilai akurasi untuk hasil klasifikasi menggunakan K-NN dengan k=3, k=5 dan k=10. Dari grafik terlihat bahwa k=3 memiliki nilai akurasi tertinggi jika dibandingkan dengan k=5 dan k=10.

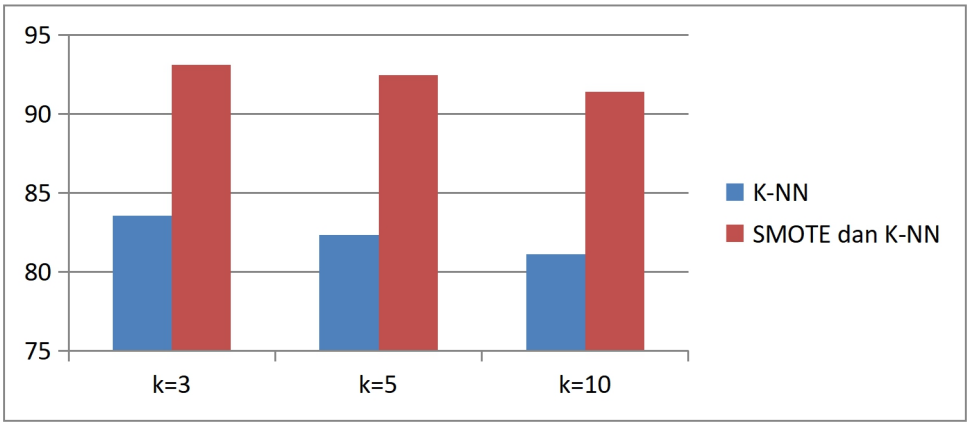


Gambar 9 Perbandingan akurasi K-NN



Gambar 10 Perbandingan akurasi K-NN dengan SMOTE

Gambar 10 menunjukkan grafik perbandingan nilai akurasi untuk hasil klasifikasi menggunakan SMOTE dan K-NN dengan k=3, k=5 dan k=10. Dari grafik terlihat bahwa k=3 memiliki nilai akurasi tertinggi jika dibandingkan dengan k=5 dan k=10.



Gambar 4.10 Grafik perbandingan hasil akurasi

Gambar 4.10 menunjukkan grafik perbandingan nilai akurasi untuk hasil klasifikasi menggunakan K-NN dengan k=3, k=5 dan k=10 dan K-NN menggunakan SMOTE dengan k=3, k=5 dan k=10. Dari grafik terlihat bahwa baik untuk k=3, k=5 dan k=10 nilai akurasi semua meningkat jika menggunakan SMOTE.

4. KESIMPULAN

Berdasarkan pembahasan yang telah diuraikan, maka dapat diambil kesimpulan sebagai berikut :

1. penerapan SMOTE pada K-NN dapat meningkatkan kenaikan akurasi rata rata 9.97% , dan peningkatan paling tinggi sebesar 10.27% terjadi pada k=10, yaitu k dengan akurasi paling rendah di KNN.
2. Semakin kecil nilai k maka semakin besar tingkat akurasinya
3. Dari hasil uji dengan k=3, k=5 dan k=10, maka akurasi klasifikasi tertinggi K-NN dengan k=3 menggunakan SMOTE sebesar 93.11%
4. Selisih akurasi antar k dengan K-NN lebih besar jika dibandingkan sesisih akurasi antar k dengan SMOTE dan K-NN

5. SARAN

Untuk menguji efektifikas penggunaan SMOTE, perlu dilakukan pengujian dengan k yang beragam serta metode klasifikasi yang lain.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada STMIK AKAKOM yang telah memberi dukungan **financial** terhadap penelitian ini.

DAFTAR PUSTAKA

[1] Naufal A R, Wahono R S, Syukur A, 2015, Penerapan Bootstraping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan, *Journal of Intelligent Systems,* Vol.1, No.2, ISSN : 2356-3982, hal. 98-108.

[2] Siringoringo R, 2018, Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan KNearest Neighbor, *Jurnal ISD,* Vol.3, No.1, pISSN : 2477-863X, eISSN:2528-5114, hal. 4-49.

[3] Pristyanto, Y, 2019, Penerapan Metode Ensemble untuk Meningkatkan Kinerja Algoritme Klasifikasi pada Imbalanced Dataset, *Jurnal Teknoinfo,* Vol. 13, No.1, ISSN:2615-224X, hal. 11-16.

[4] Wijaya J, Soleh A M, Rizki A, 2018, Penanganan Data Tidak Seimbang Pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB, *XPlore ,* vol.2, No.2, ISSN:2302-5751, hal. 32-40.

[5] Kasanah, A,N., Muladi, Pujianto U, 2017, Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN, *Jurnal RESTI,* Vol.3, No.2, ISSN:2580-0760, hal. 196-201.

[6] Astuti T, Adipurwoko S P, Diyani R, Santosa R A, Permadi B, 2018, Pengaruh Seleksi Fitur dan SMOTE terhadap Performa Klasifikasi Ranking Mobile Legends, *CITISEE,* ISBN : 978-602-60280-1- 3, hal. 114-117

[7] E. M. El Houby, N. I. Yassin, & S. Omran, 2017, A Hybrid Approach from Ant Colony Optimization and K-nearest Neighbor for Classifying Datasets Using Selected Features, *Informatica*, Vol. 41, No. 4.