Available at https://jtsiskom.undip.ac.id (15 February 2020)
DOI:10.14710/jtsiskom.8.2.2020.89-93

Jurnal Teknologi dan Sistem Komputer, 8(2), 2020, 89-93

K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes

K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes

Hairani[1*)], Khurniawan Eko Saputro[2)], Sofiansyah Fadli[3)]

[1)]*Program Studi Ilmu Komputer, Fakultas Teknik dan Kesehatan, Universitas Bumigora*
Jl. Ismail Marzuki No.22, Cilinaya, Kota Mataram, Nusa Tenggara Barat, Indonesia 83127

[2)]*Program Studi Teknologi Informasi, Fakultas Teknik dan Kesehatan, Universitas Bumigora*
Jl. Ismail Marzuki No.22, Cilinaya, Kota Mataram, Nusa Tenggara Barat, Indonesia 83127

[3)]*Program Studi Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer Lombok*
Jl. Basuki Rahmat, Praya, Mataram, Nusa Tenggara Barat, Indonesia 83511

Cara sitasi: H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89-93, 2020. doi: 10.14710/jtsiskom.8.2.2020.89-93, [Online].

Abstract – The occurrence of imbalanced class in a dataset causes the classification results to tend to the class with the largest amount of data (majority class). A sampling method is needed to balance the minority class (positive class) so that the class distribution becomes balanced and leading to better classification results. This study was conducted to overcome imbalanced class problems on the Indian Pima diabetes illness dataset using k-means-SMOTE. The dataset has 268 instances of the positive class (minority class) and 500 instances of the negative class (majority class). The classification was done by comparing C4.5, SVM, and naïve Bayes while implementing k-means-SMOTE in data sampling. Using k-means-SMOTE, the SVM classification method has the highest accuracy and sensitivity of 82 % and 77 % respectively, while the naive Bayes method produces the highest specificity of 89 %.

Keywords – *k-means-SMOTE; SMOTE; classification performance; class imbalance*

Abstrak – Kemunculan kelas yang tidak seimbang dalam suatu dataset akan menghasilkan kecenderungan klasifikasi ke kelas yang memiliki jumlah instance terbesar (majority class). Metode sampling dibutuhkan untuk menyeimbangkan kelas minoritas (kelas positif) sehingga distribusi kelas menjadi seimbang dan memperoleh hasil klasifikasi yang lebih baik. Penelitian ini dilakukan untuk menyelesaikan permasalahan ketidakseimbangan kelas pada dataset penyakit diabetes Pima Indian menggunakan k-means-SMOTE. Dataset tersebut memiliki 268 data dari kelas positif (kelas minoritas) dan 500 data dari kelas negatif (kelas mayoritas). Tahap klasifikasi dilakukan dengan membandingkan penerapan algoritma C4.5, SVM, dan naïve Bayes pada hasil sampling k-means-SMOTE. Kombinasi k-means-SMOTE dengan metode klasifikasi SVM memiliki akurasi dan sensitivitas terbaik, yaitu sebesar 82 % dan 77 %, sedangkan dengan metode naive Bayes menghasilkan spesifisitas terbaik sebesar 89 %.

Kata Kunci – *k-means-SMOTE; SMOTE; kinerja klasifikasi; ketidakseimbangan kelas*

I. PENDAHULUAN

Permasalahan ketidakseimbangan kelas (*class imbalance*) merupakan permasalahan penting untuk diatasi. Ketidakseimbangan kelas merupakan kondisi dimana jumlah *instance* kelas mayoritas lebih banyak dibandingkan dengan jumlah *instance* kelas minoritas. Permasalahan ketidakseimbangan kelas menyebabkan metode klasifikasi lebih mudah mengklasifikasikan kelas mayoritas dibandingkan kelas minoritas. Salah satu dataset yang mempunyai ketidakseimbangan kelas adalah dataset penyakit diabetes Pima Indian. Dataset tersebut diperoleh dari repositori *UCI* yang memiliki jumlah *instance* kelas positif (kelas minoritas) 268 *instance* dan 500 untuk kelas negatif (kelas mayoritas).

Permasalahan ketidakseimbangan kelas dapat ditangani dengan 2 pendekatan, yaitu pendekatan level data dan pendekatan level algoritma [1]. Beberapa metode *sampling* pada pendekatan level data yang bisa digunakan untuk menyelesaikan permasalahan ketidakseimbangan kelas adalah *oversampling*, *undersampling*, dan hibrida (kombinasi keduanya). *Oversampling* bekerja menyeimbangkan kelas minoritas dengan cara menduplikasi kelas minoritas yang sama persis sehingga terjadi *overfitting*. *Undersampling* bekerja menyeimbangkan kelas minoritas dengan menghapus kelas mayoritas sampai distribusinya

[*)]Penulis korespodensi (Hairani)
Email: hairani@universitasbumigora.ac.id

Submitted: 6 November 2019; Revised: 10 February 2020; Accepted: 14 February 2020; Published: 30 April 2020

seimbang. Kelemahan *undersampling* adalah banyaknya kehilangan informasi data berguna. Metode *sampling* hibrida bekerja dengan cara menambahkan kelas minoritas dan menghapus data kelas mayoritas sehingga distribusi kelas seimbang.

Chawla dkk. [2] mengembangkan metode *sampling* SMOTE untuk mengatasi kelemahan yang ada pada metode *oversampling*. Jika metode *oversampling* melakukan duplikasi data pada kelas minoritas sehingga terjadinya *overfiting*, metode SMOTE menambahkan kelas minoritas dengan membangkitkan data buatan atau sintesis berdasarkan k-tetangga terdekat (*k-nearest neigbor*) antar kelas minoritas.

Metode SMOTE digunakan untuk mengatasi permasalahan ketidakseimbangan kelas dalam beragam metode klasifikasi dan prediksi, di antaranya SVM dalam [3]-[9], naive Bayes dalam [7]-[10], kNN dalam [7], [11], dan C4.5 dalam [3], [8]. Penerapan SMOTE dengan beragam metode tersebut telah dilakukan dalam berbagai aplikasi, di antaranya dalam deteksi penyalahgunaan kartu kredit [3], [11], prediksi pengambil mata kuliah [5], diagnosis medis [6], klasifikasi kinerja siswa di kelas [7], segmentasi pelanggan pada industri perbankan [8], analisis sentimen [9], dan analisis sentimen [10]. Namun, metode SMOTE memiliki kelemahan, yaitu secara acak memilih *instance* kelas minoritas untuk di-*oversample* dengan menggunakan *uniform probability* sehingga rentan menghasilkan data *noise* karena tidak membedakan area kelas yang tumpang tindih [12].

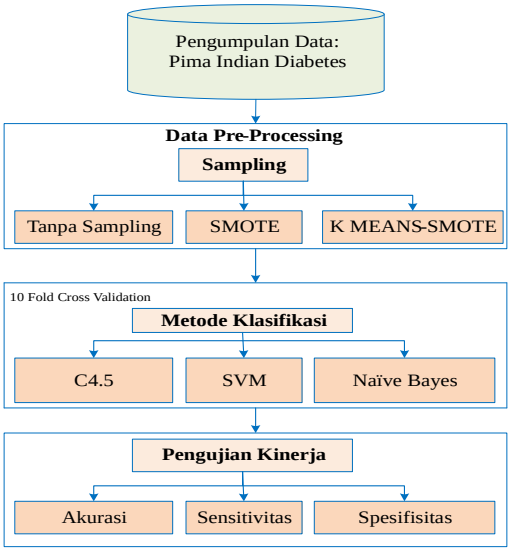
Secara khusus, klasifikasi penyakit diabetes menggunakan dataset diabetes Pima Indian telah dikaji dalam [13], [14]. Hairani dkk. [13] menggunakan metode klasifikasi *correlated naive Bayes classifier* dan naive Bayes untuk klasifikasi penyakit diabetes, sedangkan Nass dkk. [14] menggunakan C4.5, SMO, naive Bayes, kNN, dan random forest. Namun, penelitian tersebut tidak menyelesaikan permasalahan ketidakseimbangan kelas pada dataset yang digunakan. Penelitian ini bertujuan mengaplikasikan algoritma k-means-SMOTE seperti dalam [12] untuk menyelesaikan permasalahan ketidakseimbangan kelas pada dataset diabetes Pima Indian yang dikombinasikan dengan metode klasifikasi C4.5, SVM, dan naive Bayes.

II. METODE PENELITIAN

Kerangka kerja penelitian ini ditunjukkan pada Gambar 1. Tahapan pertama pada penelitian ini adalah pengumpulan dataset diabetes Pima Indian yang di peroleh dari repositori UCI. Dataset diabetes Pima Indian terdiri dari 768 data dan 9 atribut. Atribut diabetes Pima Indian ditunjukkan pada Tabel 1.

Tahapan kedua adalah prapemrosesan. Tahapan prapemrosesan yang digunakan pada penelitian ini adalah metode *sampling* karena dataset diabetes Pima Indian memiliki jumlah *instance* kelas positif (kelas minoritas) sebanyak 268 dan kelas negatif (kelas mayoritas) 500. Metode *sampling* k-means-SMOTE digunakan untuk menyeimbangkan *instance* kelas positif dan negatif. Sebagai pembanding, metode *sampling* original, SMOTE, dan k-means-SMOTE diterapkan dalam tahap prapemrosesan. Perbandingan jumlah antara kedua kelas pada dataset diabetes Pima Indian ketiga metode *sampling* ditunjukkan pada Tabel 2.

Metode k-means-SMOTE terlebih dahulu melakukan pengelompokan pada kelas mayoritas dan minoritas. Jika suatu klaster terdapat nilai *ratio imbalance* > 1, maka kelas minoritas ditambahkan dengan metode SMOTE. Algoritme k-means-SMOTE dinyatakan dalam Algoritme 1. Parameter *X* merupakan data observasi atau fitur pada dataset, *y* target atau kelas,



Gambar 1. Kerangka kerja penelitian

Tabel 1. Atribut dataset diabet Pima Indian

No	Atribut	Label	Deskripsi
1	*Number of times pregnant*	NP	Jumlah kehamilan
2	*Plasma glucose concentration a 2 hours in oral glucose tolerance test (mg/dL)*	GTT	Kadar glukosa 2 jam setelah konsumsi larutan glukosa
3	*Diastolic blood pressure (mmHg)*	DBP	Tekanan darah *diastolic*
4	*Triceps skin fold thickness (mm)*	TSF	Ketebalan lipatan kulit *triceps*
5	*2-Hour serum insulin (mIU/ml)*	HSI	Kadar insulin dalam darah 2 jam setelah makan
6	*Body mass index (Kg/m2)*	BMI	Berat masa tubuh
7	*Diabetes pedigree function*	DPF	Riwayat penyakit keluarga
8	*Age (years)*	Age	Umur
9	*Tested Negative and Tested Positive*	Class	

Tabel 2. Rasio kelas dataset diabetes Pima Indian

Metode *sampling*	Jumlah *instance*	
	Negatif	Positif
Original	500	268
SMOTE	500	500
K-means-SMOTE	500	500

dan *n* merupakan jumlah sampel yang dihasilkan dari proses *oversampling* k-means-SMOTE. Parameter *k* adalah jumlah kluster digunakan pada k-means, *irt* adalah batas ambang rasio ketidakseimbangan, *knn* adalah jumlah tetangga terdekat yang digenerate metode SMOTE, dan *de* adalah eksponen yang digunakan untuk perhitungan densitas.

Algoritme 1. k-means-SMOTE

// Klaster ruang input dan saring klaster dengan instance
// minoritas lebih banyak dari instance mayoritas
Clusters ← *kmeans(X)*
Filtered clusters ← Θ
for *c* ε *clusters* ***do***
  $imbalanced\,ratio \leftarrow \frac{majority\,count(c)+1}{minority\,count(c)+1}$
  if *imbalanced ratio* ← *irt* ***then***
    Filtered cluster ← filtered clusters υ {*c*}
  end
end

// Setiap klaster yang disaring, hitung bobot sampel
// berdasarkan densitas kelas minoritas
for *f* ε *clusters* ***do***
  average minority dist (f) ← *mean (euclidean dist (f))*
  $density\,factor(f) \leftarrow \frac{majority\,count(c)}{average\,minority\,dist(f)^{de}}$
  $sparsity\,factor(f) \leftarrow \frac{1}{density\,factor(f)}$
end
$sparsity\,sum \leftarrow \sum_{f\,\varepsilon\,clusters} sparsity\,factor(f)$
$sampling\,weight(f) \leftarrow \frac{sparsity\,factor(f)}{sparsity\,sum}$

// Penambahan sampel setiap klaster yang disaring
// menggunakan SMOTE. Jumlah sampel yang dihasilkan
// dihitung menggunakan bobot sampel
Generated samples ← Θ
for *f* ε *filtered clusters* ***do***
  $number\,of\,samples \leftarrow \|n \times sampling\,weight(f)\|$
  generated samples ← *generated samples* υ { *SMOTE* (*f, number of samples, knn*)}
end
return generated samples

Tahapan selanjutnya adalah implementasi dan validasi metode klasifikasi. Metode klasifikasi yang digunakan penelitian ini adalah C4.5, SVM, dan naive Bayes. Ketiga metode klasifikasi tersebut termasuk 10 metode klasifikasi *data mining* yang memiliki performa yang bagus [15]. Validasi model metode klasifikasi yang digunakan pada penelitian ini adalah *10-fold cross validation.* Metode *10-fold cross validation* ini membagi dataset menjadi 10 kelompok data.

Pengujian kinerja dilakukan untuk mengetahui kinerja metode C4.5, SVM, dan naive Bayes yang dikombinasikan dengan metode tanpa sampling, dengan *sampling* SMOTE, dan k-means-SMOTE. Pengujian kinerja ini berdasarkan akurasi, sensitivitas, dan spesifisitas yang dinyatakan menggunakan matrik konfusi. Perhitungan kinerja dari metode (akurasi, sensitivitas, dan spesifisitas) dinyatakan dalam Persamaan 1 sampai dengan Persamaan 3. Parameter TP (*True Positive*) menunjukkan jumlah prediksi positif dari aktual kelas positif. FP (*False Positive*) menyatakan jumlah prediksi positif dari kelas aktual negatif. TN (*True Negative*) menyatakan jumlah prediksi negatif dari kelas aktual negatif. FN (False Negative) menunjukkan jumlah prediksi negatif dari kelas aktual positif.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Sensitivitas = \frac{TP}{TP+FN} \quad (2)$$

$$Spesifisitas = \frac{TN}{TN+FP} \quad (3)$$

III. HASIL DAN PEMBAHASAN

Implementasi metode C4.5, SVM, dan naive Bayes dilakukan tanpa *sampling*, dengan SMOTE, dan k-means-SMOTE dan membandingkan kinerjanya berdasarkan akurasi, sensitivitas, dan spesifisitas. Matriks konfusi hasil pengujian menggunakan validasi model *10-fold cross validation* untuk ketiga metode tersebut dinyatakan dalam Tabel 3 untuk metode C4.5, Tabel 4 untuk SVM, dan Tabel 5 untuk naive Bayes.

Dari matriks konfusi tersebut, diperoleh hasil akurasi, sensitivitas, dan spesifisitas dari metode klasifikasi dengan *sampling*-nya masing-masing seperti ditunjukkan pada Tabel 6, Tabel 7, dan Tabel 8. Metode *sampling* k-means-SMOTE memiliki akurasi lebih baik dibandingkan metode SMOTE dan tanpa *sampling* untuk semua metode klasifikasi yang digunakan (Tabel 6). Metode k-means-SMOTE ini dapat memperbaiki akurasi klasifikasi penyakit diabetes dalam [13] yang menggunakan SMOTE dan naive Bayes sebesar 6 % (dari 73 % menjadi 79 %).

Nilai akurasi terbaik diperoleh dari kombinasi metode SVM dan k-means-SMOTE. Hal ini sejalan dengan [8], [9], [14], [16] yang menyatakan bahwa metode SVM memiliki akurasi terbaik. Lebih lanjut, metode SMOTE dan k-means-SMOTE dapat memperbaiki akurasi dalam metode SVM tanpa *sampling* sesuai dengan [3]-[9].

Metode *sampling* k-means-SMOTE memiliki nilai sensitivitas lebih baik dibandingkan metode SMOTE pada semua metode klasifikasi yang digunakan (Tabel 7). Nilai sensitivitas terbaik diperoleh dari kombinasi metode SVM dan k-means-SMOTE. Peningkatan nilai sensitivitas terjadi setelah *oversampling* menggunakan SMOTE dan k-means-SMOTE karena ada penambahan kelas minoritas (kelas positif) sehingga distribusi kelasnya menjadi seimbang [5], [8], [10]-[12]. Selain peningkatan akurasi, metode naive Bayes dan k-means-SMOTE ini memiliki keunggulan daripada [13] dari sisi sensitivitas, yaitu kebenaran dalam memprediksi positif terhadap keseluruhan data positif.

Metode k-means-SMOTE memiliki nilai spesifisitas lebih baik dibandingkan metode SMOTE pada semua

metode klasifikasi yang digunakan (Tabel 8). Nilai spesifisitas terbaik didapatkan kombinasi metode naive Bayes dan k-means-SMOTE. Selain peningkatan akurasi, metode naive Bayes dan k-means-SMOTE ini memiliki keunggulan daripada [13] dari sisi spesifisitas, yaitu kebenaran dalam memprediksi negatif terhadap keseluruhan data negatif. Namun, penggunaan sampling pada C4.5 membuat spesifisitasnya lebih kecil. Hal ini menunjukkan bahwa C4.5 mampu memprediksi kelas negatif lebih baik daripada dengan *sampling*, walaupun presisi kelas negatifnya lebih rendah daripada k-means-SMOTE.

Secara keseluruhan, metode *sampling* k-means-SMOTE memiliki nilai akurasi, sensitivitas, dan spesifisitas terbaik dibandingkan dengan SMOTE pada semua metode klasifikasi yang digunakan, seperti C4.5, SVM, dan naive Bayes. Sejalan dengan [12], metode k-means-SMOTE memiliki kinerja lebih baik dari SMOTE pada dataset diabetes Pima Indian. Hal ini terjadi karena kelas dalam k-means-SMOTE dikelompokkan terlebih dahulu jika sebuah klaster memiliki *rasio imbalance* lebih dari 1 pada kelas minoritas sebelum dilakukan *oversampling* menggunakan SMOTE, sedangkan metode SMOTE melakukan *oversampling* pada kelas minoritas berdasarkan jarak terdekat tanpa pengelompokkan terlebih dahulu.

IV. KESIMPULAN

Penggunaan metode *sampling* k-means-SMOTE untuk menangani ketidakseimbangan kelas pada dataset diabetes Pima Indian menunjukkan peningkatan kinerja berdasarkan akurasi, sensitivitas, dan spesifisitas dibandingkan metode SMOTE. Kombinasi k-means-SMOTE dengan metode klasifikasi SVM memiliki akurasi dan sensitivitas lebih baik, yaitu 82 % dan 77 %, sedangkan dengan metode naive Bayes menghasilkan spesifisitas terbaik, yaitu 89 %.

DAFTAR PUSTAKA

[1] B. Santoso, H. Wijayanto, K. Notodiputro, and B. Sartono, "Class imbalanced problems: a review," *Conference Series: Earth and Environmental Science.*, vol. 58, no. 1, pp. 427-436, 2017. doi: 10.1088/1755-1315/58/1/012031

[2] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341-378, 2002. doi: 10.1613/jair.953

[3] S. Sisodia, N. K. Reddy, and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," in *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering*, Chennai, India, Sept. 2017, pp. 2747–2752. doi: 10.1109/ICPCSI.2017.8392219

[4] L. Demidova and I. Klyueva, "SVM classification: optimization with the SMOTE algorithm for the

Tabel 3. Matriks konfusi metode C4.5

Metode *sampling*	Prediksi	*Aktual* Negatif	*Aktual* Positif	Presisi kelas
Tanpa *sampling*	Negatif	389	128	0,75
	Positif	111	140	0,56
	Recall kelas	*0,78*	*0,52*	
SMOTE	Negatif	368	137	0,73
	Positif	132	363	0,73
	Recall kelas	*0,74*	*0,73*	
k-means-SMOTE	Negatif	385	115	0,77
	Positif	115	383	0,77
	Recall kelas	*0,77*	*0,77*	

Tabel 4. Matriks konfusi metode SVM

Metode *sampling*	Prediksi	*Aktual* Negatif	*Aktual* Positif	Presisi kelas
Tanpa *sampling*	Negatif	438	117	0,79
	Positif	62	151	0,71
	Recall kelas	*0,88*	*0,56*	
SMOTE	Negatif	394	147	0,73
	Positif	106	353	0,77
	Recall kelas	*0,79*	*0,71*	
k-means-SMOTE	Negatif	437	114	0,79
	Positif	63	384	0,86
	Recall kelas	*0,87*	*0,77*	

Tabel 5. Matriks konfusi metode naive Bayes

Metode *sampling*	Prediksi	*Aktual* Negatif	*Aktual* Positif	Presisi kelas
Tanpa *sampling*	Negatif	422	109	0,79
	Positif	78	159	0,67
	Recall kelas	*0,84*	*0,59*	
SMOTE	Negatif	392	168	0,7
	Positif	108	332	0,75
	Recall kelas	*0,78*	*0,66*	
k-means-SMOTE	Negatif	443	151	0,75
	Positif	57	347	0,86
	Recall kelas	*0,89*	*0,7*	

Tabel 6. Akurasi metode klasifikasi

Metode Klasifikasi	Metode *Sampling* Original	SMOTE	K-means-SMOTE
C4.5	0,69	0,73	0,77
SVM	0,77	0,75	**0,82**
Naïve Bayes	0,76	0,73	0,79

Tabel 6. Sensitivitas metode klasifikasi

Metode Klasifikasi	Metode *Sampling* Original	SMOTE	K-means-SMOTE
C4.5	0,52	0,73	**0,77**
SVM	0,56	0,71	**0,77**
Naïve Bayes	0,59	0,66	0,69

Tabel 7. Spesifisitas metode klasifikasi

Metode Klasifikasi	Metode *Sampling* Original	SMOTE	K-means-SMOTE
C4.5	0,78	0,74	0,77
SVM	0,88	0,79	0,87
Naïve Bayes	0,84	0,78	**0,89**

class imbalance problem," in *6th Mediterranean Conference on Embedded Computing (MECO)*, Bar, Montenegro, Jun. 2017, pp. 17–20. doi: 10.1109/MECO.2017.7977136

[5] F. A. Bachtiar, I. K. Syahputra, and S. A. Wicaksono, "Perbandingan algoritme machine learning untuk memprediksi pengambil matakuliah," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 5, pp. 543-548, 2019. doi: 10.25126/jtiik.2019651755

[6] L. Demidova and I. Klyueva, "Improving the classification quality of the SVM classifier for the imbalanced datasets on the base of ideas the SMOTE algorithm," *ITM Web of Conferences*, vol. 10, 2017, pp. 1-4. doi: 10.1051/itmconf/20171002002

[7] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, "Hybrid resampling to handle imbalanced class on classification of student performance in classroom," in *1st International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, Nov. 2017, pp. 207-212. doi: 10.1109/ICICOS.2017.8276363

[8] H. Hairani, N. A. Setiawan, and T. B. Adji, "Metode klasifikasi data mining dan teknik sampling SMOTE menangani class imbalance untuk segmentasi customer pada industri perbankan," in *Seminar Nasional Sains dan Teknologi*, Semarang, Indonesia, Aug. 2016, pp. 168-172.

[9] A. C. Flores, R. I. Icoy, C. F. Pena, and K. D. Gorro, "An evaluation of SVM and naive Bayes with SMOTE on sentiment analysis data set," in *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, Phuket, Thailand, Jul. 2018, pp. 1-4. doi: 10.1109/ICEAST.2018.8434401

[10] Z. Ulhaq and T. B. Adji, "Integrasi synthetic minority over-sampling technique (SMOTE) dengan correlated naïve Bayes pada klasifikasi siswa berkesulitan belajar," in *CITEE*, Yogyakarta, Indonesia, Jul. 2017, pp. 201-205.

[11] R. Siringoringo, "Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor," *Journal Information System Development*, vol. 3, no. 1, pp. 44-49, 2018.

[12] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information System*, vol. 465, pp. 1-20, 2018. doi: 10.1016/j.ins.2018.06.056

[13] H. Hairani, G. Nugraha, M. Nurkholis Abdillah, and M. Innuddin, "Komparasi akurasi metode correlated naive Bayes classifier dan naive Bayes classifier untuk diagnosis penyakit diabetes," *InfoTekJar (Jurnal Nasional. Informatika dan Teknologi Jaringan)*, vol. 3, no. 1, pp. 6-11, 2018.

[14] L. Nass, S. Swift, and A. Al Dallal, "Indepth analysis of medical dataset mining: a comparitive analysis on a diabetes dataset before and after preprocessing," *KnE Social Sciences*, vol. 3, no. 25, pp. 45-63, 2019. doi: 10.18502/kss.v3i25.5190

[15] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1-37, 2008.

[16] N. Nurajijah and D. Riana, "Algoritma naïve Bayes, decision tree, dan SVM untuk klasifikasi persetujuan pembiayaan nasabah koperasi syariah," *Jurrnal Teknologi dan Sistem Komputer.*, vol. 7, no. 2, pp. 77-82, 2019. doi: 10.14710/jtsiskom.7.2.2019.77-82