Jurnal Teknologi Terpadu

https://journal.nurulfikri.ac.id/index.php/jtt
ISSN : 2477-0043 ISSN ONLINE : 2460-7908

PENGARUH KESEIMBANGAN DATA TERHADAP AKURASI MODEL *SUPPORT VECTOR MACHINE* PADA *DATA SET* DONOR DARAH

Agung Widyanto[1], Kusrini[2], Kusnawi[3]

[1,2,3] Teknik Informatika, Universitas Amikom Yogyakarta
Kabupaten Sleman, Daerah Istimewa Yogyakarta, Indonesia 55281
agungwidyanto@students.amikom.ac.id, kusrini@amikom.ac.id, kusnawi@amikom.ac.id

Abstract

In classification, unbalanced data is expected. Unbalanced data has an inequality ratio between the majority and minority classes. Models trained with unbalanced data tend to predict the minority class as the majority class. This study aims to determine the effect of data balance on the accuracy of the Support Vector Machine (SVM) classification model. The data set used is the blood donor data set downloaded from the repository belonging to the University of California, Irvine (UCI). The Waikato Environment for Knowledge Analysis (WEKA) tool was chosen to present the results of training development and model testing. The research framework scheme is used as a reference for knowledge flow. In scenario 1, data pre-processing includes handling missing values using mean-impulse and normalizing MinMax scaling. With a data set that has an inequality ratio of 1:3, the SVM classifier gets an accuracy performance of 76.7%. In scenario 2, post-pre-processing is done by balancing the data using the Synthetic Minority Oversampling Technique (SMOTE). SVM classifier gets 69.8% accuracy performance. Model performance is evaluated using confusion metrics. The gap in recall values for each class is very high in scenario 1 (2.8% and 99.8%). Things are different in scenario 2 (75.6% and 64%). The test results of 748 samples obtained an accuracy of 76.7% for the scenario-1 model and 93.2% for the scenario-2 model. This proves that the balance of data influences the accuracy of the SVM classification model.

Keywords: *Data imbalance, Min-Max scaling, SMOTE, SVM, WEKA*

Abstrak

Pada klasifikasi, data yang tidak seimbang menjadi hal yang umum ditemukan. Data yang tidak seimbang memiliki rasio ketimpangan kelas mayoritas dan minoritas. Model yang dilatih dengan data yang tidak seimbang mengakibatkan model cenderung memprediksi kelas minoritas sebagai kelas mayoritas. Penelitian ini memiliki tujuan untuk mengetahui pengaruh keseimbangan data terhadap akurasi model klasifikasi *Support Vector Machine (SVM)*. *Data set* yang digunakan adalah *data set* donor darah yang diunduh dari repositori milik *University of California,Irvine (UCI)*. Alat *Waikato Environment for Knowledge Analysis (WEKA)* dipilih untuk menyajikan hasil pengembangan pelatihan dan pengujian model. Skema kerangka kerja penelitian digunakan sebagai acuan *Knowledge Flow*. Pada skenario-1, pra-pemrosesan data mencakup penanganan *missing value* menggunakan *mean-impulse* dan normalisasi *MinMax Scaling*. Dengan *data set* yang memiliki rasio ketimpangan 1:3, pengklasifikasi *SVM* mendapatkan performa akurasi sebesar 76.7%. Sedangkan pada skenario-2, pasca pra-pemrosesan dilakukan penyeimbangan data menerapkan *Synthetic Minority Oversampling Technique (SMOTE)*. Pengklasifikasi *SVM* mendapatkan performansi akurasi 69.8%. Kinerja model dievaluasi menggunakan *confusion metric*. Gap nilai *recall* tiap kelas sangat tinggi pada skenario-1 (2.8% dan 99.8%). Hal yang berbeda pada skenario-2 (75.6% dan 64%). Hasil uji 748 sampel, didapatkan akurasi 76.7% model skenario-1, dan akurasi 93.2% model skenario-2. Hal ini membuktikan bahwa keseimbangan data memiliki pengaruh terhadap akurasi model klasifikasi *SVM*.

Kata kunci: *Data imbalance, Min-Max scaling, SMOTE, SVM, WEKA*

1. PENDAHULUAN

Permasalahan ketidakseimbangan pada kumpulan data memberikan kecenderungan hasil latih pembelajaran mesin melakukan prediksi sering kali kelas minoritas di salah klasifikasikan sebagai kelas mayoritas. Hal ini tentunya mengakibatkan penurunan performa akurasi pada *class* yang diprediksi.

Pada penelitian [1][2] transfusi/donor darah, perilaku data donor dapat dilihat dari *Recency* adalah berapa bulan

Diterima 25 Mei 2023, Direvisi 4 Juni 2023, Diterima untuk publikasi 13 Juni 2023

dihitung dari sukarelawan mendonorkan darah terakhir kalinya, *Frequency* adalah berapa kali sukarelawan telah mendonorkan darah, *Monetary* adalah berapakah jumlah(cc) darah yang telah sukarelawan donorkan, *Time* adalah berapakah bulan akumulasi sukarelawan mendonor. Studi kasus donor darah ini dipilih dengan pendekatan *Recency*;*Frequency;Monetary*;*Time;Churn (RFMTC)* sebagai modifikasi dari *Recency of purchase, Frequency of purchase, dan Monetary value of purchase (RFM)* yang digunakan untuk memprediksi perilaku pendonor darah sehingga diperoleh formula yang dapat memperkirakan probabilitas pendonor diklasifikasikan apakah akan mendonorkan darahnya atau tidak. Penelitian [2] mendapatkan akurasi 78.13% dengan menerapkan pengklasifikasi *Naïve Bayes*, dan kenaikan akurasi 80.8% setelah dilakukan kombinasi *Naïve Bayes* dan *K-Mean*. Perbandingan kelas mayoritas dan minoritas pada kelas biner mendonorkan atau tidak mendonorkan menunjukkan bahwa *data set* ini memiliki ketidakseimbangan data dengan rasio ketimpangan 1 berbanding 3. Ketimpangan ini merupakan bentuk ketidakseimbangan pada kumpulan data dengan kondisi kelas mayoritas memiliki jumlah sampel yang tinggi dibandingkan dengan kelas lainnya (minoritas) secara kuantitas[3].

Beberapa penelitian sebelumnya telah dilakukan untuk mengidentifikasi ketidakseimbangan data. Penelitian [4] menggambarkan hasil deteksi peristiwa suara (*SED*) lebih cenderung ke *frame* yang tidak aktif daripada *frame* aktif. Hal ini didapatkan, karena perbedaan durasi waktu antar *class event* suara membuat perbedaan jumlah secara signifikan sampel data antar *class event*. Hasil latih yang tidak seimbang ini memberikan pengaruh model kinerja deteksi peristiwa suara (*SED)* yang dihasilkan. *Imbalanced ratio* menunjukkan ukuran ketidakseimbangan *data set* [5][6].

Dalam penelitian [7] diterapkan algoritma *SMOTE* untuk mendapatkan keseimbangan kumpulan data yang dipergunakan. Penelitian tersebut menggunakan beberapa *data set. Data set* pertama sebanyak 68 data, *data set* kedua sejumlah 180 data, dan *data set* ketiga dengan 371 data. Pasca dilakukan *SMOTE treatment*, seluruh kumpulan data tersebut menjadi *balance data set*. Teknik *oversampling* menunjukkan perbaikan terhadap akurasi [8][9][10].

Penelitian lainnya [11] memprediksi probabilitas pendonor darah akan berdonor kembali atau tidak dengan menggunakan algoritma *Random Forest*, *SVM* dan *Regresi Logistik*. *Instance* dari *data set* terdapat 9 fitur di antaranya memiliki kesamaan dengan pendekatan *rfmt* yakni tanggal terakhir berdonor, jumlah akumulasi berdonor, tanggal pertama berdonor darah dan periodenya.

Dengan demikian, penelitian yang dilakukan ini adalah upaya untuk mendapatkan *data set* donor darah yang seimbang yang tidak dilakukan pada penelitian sebelumnya[2]. Teknik penambahan data sintesis *SMOTE* pada penelitian [7] dengan pengaturan standar digunakan untuk mendapatkan *data set* dengan rasio ketimpangan 1:1. Upaya lainnya untuk mencapai tujuan penelitian ini adalah dengan melakukan perbandingan hasil evaluasi model klasifikasi *SVM* yang diterapkan *data set* sebelum seimbang dan sesudah seimbang. Pengukuran performa tiap model mencakup evaluasi seluruh *class* dan evaluasi masing-masing *class.* Untuk mendapatkan kesimpulan pengaruh keseimbangan data pada penelitian ini dilakukan simulasi prediksi dengan 748 sampel pada model yang telah dihasilkan untuk mengetahui performa akurasinya.

2. METODE PENELITIAN

2.1 Jenis, sifat, pendekatan dan pengumpulan data penelitian

Jenis penelitian adalah studi kasus, penelitian dengan melakukan pengamatan pengaruh keseimbangan data dengan pengklasifikasi *Support Vector Machine* pada kasus probabilitas pendonor untuk melakukan donor darah kembali atau tidak melakukan donor darah.

Sifat dari penelitian yang dilakukan adalah kausal, dengan sebab keseimbangan data maka didapatkan akibat dalam data latih pada algoritma klasifikasi pendonor darah. Variabel bebas yakni *class* variabel adalah variabel yang dimanipulasi dalam kondisi terkontrol.

Pendekatan penelitian ini adalah pendekatan kuantitatif dengan melakukan penelitian sesuai dengan tahap-tahap atau alur penelitian yang telah dibuat. Dilakukan secara objektif dan verifikatif terhadap hasil pengukuran secara numerik.

Data set yang diunggah di repositori milik *UCI* ini adalah kumpulan data yang berasal dari Taiwan, di mana di Kota Hsin-Chu saat itu menyelenggarakan kegiatan Layanan Transfusi Darah. Terdapat sebanyak 748 *instances*, dengan 5 atribut (variabel) lihat Tabel 1, yakni *Recency* adalah berapa bulan sejak dihitung dari terakhir sukarelawan mendonorkan darah, *Frequency* adalah berapa kali sukarelawan telah mendonorkan darah, *Monetary* adalah berapakah jumlah (cc) darah yang telah sukarelawan donorkan, *Time* adalah berapakah bulan akumulasi sukarelawan mendonor dan sebuah kelas biner yang merepresentasikan berdonor(2) atau tidak berdonor(1). Penelitian sebelumnya menggunakan *data set* yang sama.

Tabel 1. Deskripsi

No	Nama Variabel	Deskripsi Variabel
1	*Recency*	adalah berapa bulan sejak dihitung dari terakhir sukarelawan mendonorkan darah
2	*Frequency*	adalah berapa kali sukarelawan telah mendonorkan darah
3	*Monetary*	adalah berapakah jumlah (cc) darah yang telah sukarelawan donorkan
4	*Time*	adalah berapakah bulan akumulasi sukarelawan mendonor

No	Nama Variabel	Deskripsi Variabel
5	*Class*	representasi mendornorkan darah di bulan Maret 2007, 1 tidak mendonorkan, 2 mendonorkan.

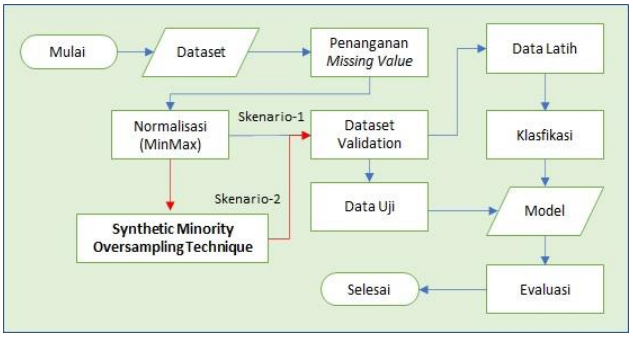
Tabel 2 menyajikan isi (sampel) dari *data set* yang dipergunakan. Variabel fitur bertipe numerik, sedangkan variabel target (*class*) bertipe nominal.

Tabel 2. Contoh Sampel *Data set*

Recency	*Frequency*	*Monetary*	*Time*	*Class*
2	50	12500	98	1
0	13	3250	28	1
1	16	4000	35	1
2	20	5000	45	1
1	24	6000	77	2

2.2 Tahapan penelitian

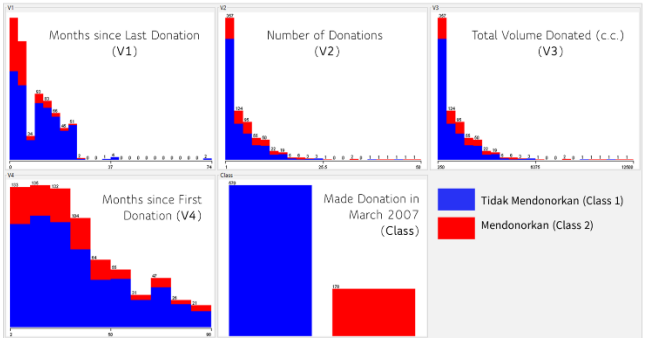
Tahapan penelitian terdiri dari beberapa proses yang ditunjukkan Gambar 1. Terdapat dua skenario. Pada skenario 1, dimulai dengan mengumpulkan *data set*, penanganan *missing value*, normalisasi, validasi, dan mengklasifikasikan menggunakan *Support Vector Machine (SVM)*. Hasilnya adalah model klasifikasi, yang dievaluasi menggunakan akurasi, presisi, *recall* dan *F-Measure* berdasarkan *confusion matrix*. Sedangkan pada skenario 2, yang membedakan sebelum proses normalisasi data, dilakukan *balancing data set*. Dilanjutkan dengan proses seperti pada skenario 1 untuk mendapatkan hasil evaluasi skenario 2. Selanjutnya dilakukan analisis perbandingan hasil evaluasi skenario 1 dan skenario 2 dengan melihat performa pada tiap *class* dan reratanya (*weighted average*).



Gambar 1. Kerangka Kerja Penelitian

2.2.1 *Data set*

Di awal aliran kerangka kerja penelitian pada gambar 1 adalah *Data set. Data set* yang telah didapatkan dieksplorasi untuk lebih mengenal karakter data. Dilakukan pengecekan terhadap kesesuaian deskripsi *data set* dengan data yang diunduh. Dilakukan pemastian kesesuaian pada format *type file data set*(.data), jumlah (748 sampel), identitas fitur (rfmt), atribut(*class*). Terdapat sebanyak 748 sampel, dengan 5 atribut yakni *Recency* adalah berapa bulan sejak dihitung dari terakhir sukarelawan mendonorkan darah (**V1**), *Frequency* adalah berapa kali sukarelawan telah mendonorkan darah (**V2**), *Monetary* adalah berapakah jumlah (cc) darah yang telah sukarelawan donorkan (**V3**), *Time* adalah berapakah bulan akumulasi sukarelawan mendonor (**V4**) dan sebuah kelas biner yang merepresentasikan berdonor(2) atau tidak berdonor(1) (**V5**). Pada Gambar 2, adalah visualisasi *data set* yang dipergunakan.



Gambar 2. Visualisasi *Data set*

2.2.2 Penanganan *Missing Value*

Pada penelitian ini jumlah sampel *data set* sebanyak 748, ditemukan pada V1 sebanyak 5 (~1%) data dengan status *missing value*. Yakni pada *record* ke 2;12;68;107;5055. Penanganan *missing value* ini dilakukan dengan teknik *impulse mean*[12]. Hasilnya, pada kelima *record* yang kosong tersebut di V1, telah disesuaikan dengan nilai 9.570659 yakni nilai *mean* dari data V1.

2.2.3 Normalisasi *Data set*

Tahap berikutnya yang dilakukan pada skenario-1 adalah melakukan normalisasi data. Normalisasi adalah bagian dari teknik pra-pemrosesan data. Dengan kumpulan data yang sudah dinormalkan, hal tersebut berkontribusi dalam pembelajaran mesin untuk mendapatkan model yang baik[11]. Metode normalisasi yang digunakan penelitian adalah *Min-Max Scaling*.

Dalam *Min-Max Scaling* data disesuaikan dalam jangkauan rentang standar antara 0-1. Persamaan (1) merupakan formula untuk menghitung nilai yang diskalakan.

$$X_{std} = \frac{(X - X.\min(axis=0))}{(\mathrm{X.max(axis=0)} - X.min(axis=0)}$$

$$X_{scaled} = X_{std} \times (\max - min) + min \qquad (1)$$

2.2.4 *Data balancing*

Pada skenario 2, sebelum dilakukan normalisasi data terlebih dahulu dilakukan penyeimbangan data. Pada Gambar 2, perbandingan kelas mayoritas dan minoritas pada kelas biner mendonorkan atau tidak mendonorkan menunjukkan bahwa himpunan data ini memiliki ketidakseimbangan data dengan rasio ketimpangan 1:3. Ketimpangan kelas ini merupakan bentuk ketidakseimbangan dengan kondisi kelompok kelas mayoritas memiliki jumlah sampel berbeda , biasanya lebih

tinggi dibandingkan dengan kelas lainnya (minoritas) secara kuantitas [3].

Teknik yang digunakan untuk menyeimbangkan himpunan data adalah dengan menerapkan *random over sampling*, *Sythentic Minority Oversampling Technique (SMOTE)* [13][14]. Dengan menambahkan data sintesis melalui *SMOTE* maka didapatkan 570 sampel masing-masing *class.*

2.2.4 Distribusi data untuk *training* dan *testing*

Tahap selanjutnya data dibagi menjadi *data training* dan *data testing*. *Data training* digunakan untuk pemodelan dan *data testing* dilakukan untuk menguji kinerja dan kebenaran dalam model yang digunakan. Pada penelitian ini digunakan 10 *fold CV(Cross Validation)*, data dibagi 10 *fold* yang memiliki ukuran mendekati sama, terdapat 10 *subset* data untuk mengevaluasi model. Dalam prosesnya tiap 10 *subset*, digunakan 9 *training fold* dan 1 *testing fold* oleh *Cross Validation*.

Pengklasifikasi menggunakan algoritma *SVM* untuk mendapatkan *hyperplane* sebagai sebuah fungsi untuk mengoptimalkan separasi pada observasi dengan nilai variabel target yang berbeda [15]. Disebut h*yperplane* karena berupa line pada dua dimensi dan disebut *flatplane* pada *multiple* dimensi. Hasil prediksi dari model ditabulasikan ke dalam *confusion matrix* ditunjukkan pada Tabel 3. Evaluasi akurasi didapatkan menggunakan persamaan (2). Evaluasi presisi didapatkan menggunakan persamaan (3). Evaluasi *recall* didapatkan menggunakan persamaan (4), dan *F-1 Score* dengan persamaan (5).

Tabel 3. *Confusion Matrix*

		Aktual	
		***Class* 1**	***Class* 2**
Prediksi	***Class 1***	*True Positive (TP)*	*False Negative (FN)*
	Class 2	*False Positive (FP)*	*True Negative (TN)*

Akurasi dihitung menurut Persamaan (2).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

Presisi dihitung menurut Persamaan (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall dihitung menurut Persamaan (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

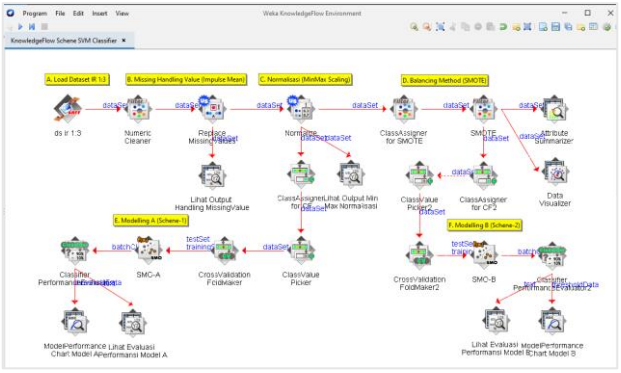
F-Measure dihitung menurut Persamaan 5.

$$FM = 2 \times \frac{Pr * Sn}{Pr + Sn} \quad (5)$$

Tabel evaluasi *confusion matrix* pada Tabel 3 adalah sebagai metrik untuk mendapatkan performa model sebagai hasil latih pembelajaran mesin. Dilakukan perhitungan dan pengamatan akurasi menggunakan persamaan (2), *precision* menggunakan persamaan (3), nilai hasil *recall* menggunakan persamaan (4), dan *F-Measure* menggunakan persamaan (5). Secara spesifik dengan mengamati nilai yang dihasilkan pada *precision* dan *recall*, pembelajaran yang dihasilkan digunakan untuk meningkatkan nilai *recall* tanpa mempengaruhi nilai *precision*, dan hal ini didapatkan melalui penerapan keseimbangan data.

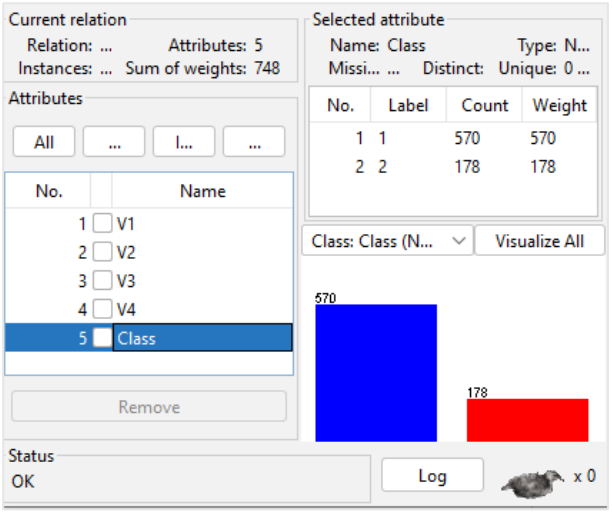
3. HASIL DAN PEMBAHASAN

Penyiapan lingkungan pengembangan dan pengujian model dilakukan dengan memanfaatkan *tool Waikato Environment for Knowledge Analysis (WEKA) - licensed under the GNU General Public License*. *WEKA* menerapkan berbagai algoritma pembelajaran mesin untuk melakukan beberapa proses yang berkaitan dengan penambangan data.



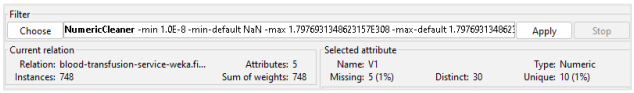
Gambar 3. *WEKA Knowledge Flow*

Kerangka kerja penelitian pada Gambar 1, ditransformasikan pada sebuah *layout Knowledge Flow WEKA* seperti yang diperlihatkan pada Gambar 3. Hal ini membuat tahapan penelitian yang dilakukan menjadi runut, utuh dan jelas. Penyiapan *data set* ditunjukkan pada blok "A. *Load Data set IR* 1:3". *Imbalanced Ratio* (*IR*) atau rasio ketimpangan kelas didapatkan dari perbandingan jumlah sampel minor 178 terhadap jumlah sampel mayor 570 (lihat Gambar 4).



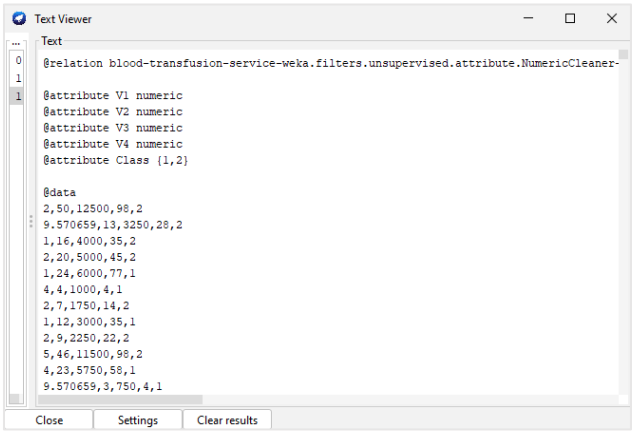
Gambar 4. Rasio Ketimpangan *Data set*

Tahapan *pre-processing* pada blok "B. *Missing Handling Value (Impulse Mean)*", Pada penelitian ini jumlah sampel sebanyak 748, pada *filter* yang dipilih gunakan *weka.filters.unsupervised.attribute.NumericCleaner* dengan pengaturan *minDefault* adalah *NaN* dan *minThreshold* 0.1E-7. Ditemukan pada V1 sebanyak 5 sampel (~1%) lihat Gambar 5, data dengan status *missing value*. Yakni pada baris ke 2;12;68;107;5055.



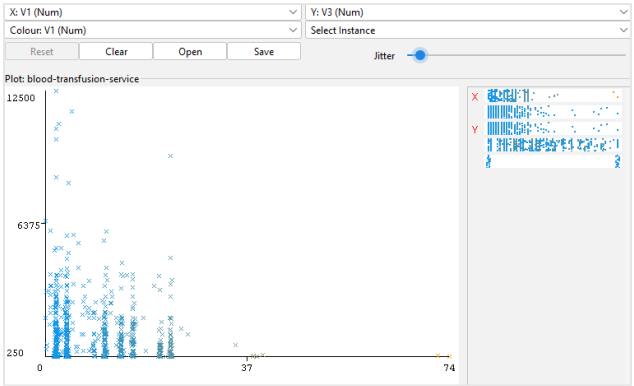
Gambar 5. *Status Missing Value*

Untuk penanganan *missing value* dengan *mean-impulse, weka.filters.unsupervised.attribute.ReplaceMissingValues.*



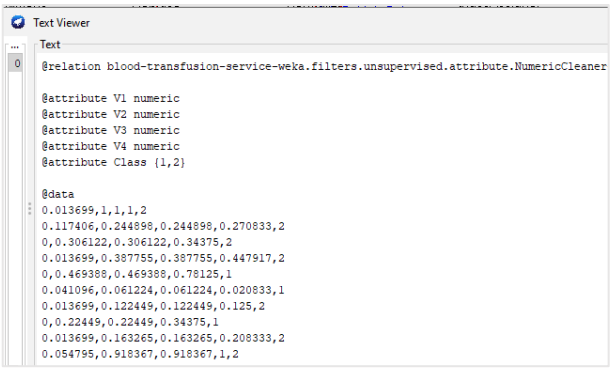
Gambar 6. Sampel Data Hasil *Impulse Mean*

Hasilnya, pada kelima *record* yang kosong tersebut di V1, telah disesuaikan dengan nilai 9.570659. Lihat pada Gambar 6 terdapat 12 baris sampel data. Pada data ke 2 dan 12 telah dilakukan penanganan *missing value* dengan *impulse mean*.



Gambar 7. Distribusi *monetary* terhadap *recency*

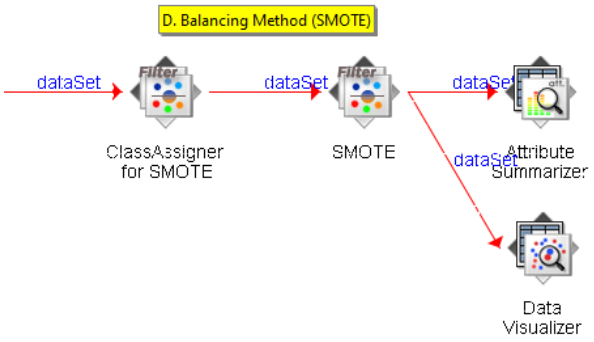
Gambar 7 memvisualkan sebaran data yang digunakan memiliki jangkauan distribusi nilai sangat jauh, dapat dicermati skala yang digunakan pada *monetary(V3)* sumbu-y adalah ribuan dan *recency(V1)* sumbu-x dengan skala satuan. Proses normalisasi yang digambarkan pada blok "C. Normalisasi (*MinMax Scaling*)", menghasilkan fitur berada dalam skala distribusi dengan rentang yang sama, gunakan *weka.filters.unsupervised.attribute.Normalize.* Hasilnya tanpa mengubah "makna data" himpunan data ini diskalakan pada setiap variabel *input*, lihat Gambar 8.



Gambar 8. Hasil Normalisasi

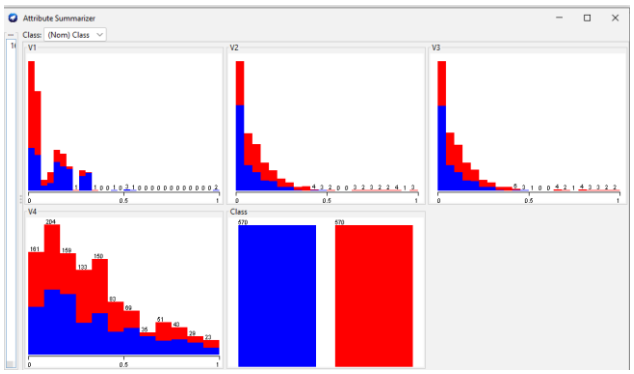
Pada Gambar 8, selain dari atribut *class* dilakukan normalisasi dengan menggunakan *MinMax Scaling.* Atribut V1-V4 memiliki rentang antara 0 hingga 1 setelah dilakukan normalisasi. *Data set* hasil normalisasi digunakan pada pembelajaran mesin model skenario-1.

Tahap penyeimbangan data digambarkan pada Gambar 9 *layout Knowledge Flow* label "D. *Balancing Method (SMOTE)*".



Gambar 9. *Knowledge Flow Balancing Method*

SMOTE meningkatkan kuantitas sampel minoritas dalam suatu himpunan data. Untuk mendapatkan rasio 1:1, gunakan pengaturan tetangga terdekat =5 dan persentase duplikasi =220. (*weka.filters.supervised.instance.SMOTE)*



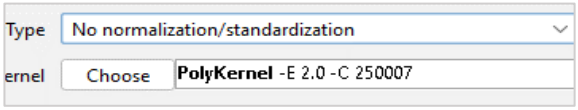
Gambar 10. Visualisasi *Data set* Hasil *SMOTE*

Algoritma *SMOTE* bekerja menggunakan sampel pada setiap kelas target dan tetangga terdekatnya untuk menghasilkan contoh data baru dengan menggabungkan fitur kasus target dan fitur tetangganya. Pendekatan ini meningkatkan fitur yang tersedia untuk setiap kelas dan membuat sampel sintesis menjadi lebih mirip. *Data set* hasil *SMOTE* dipergunakan pada pembelajaran mesin model skenario-2.

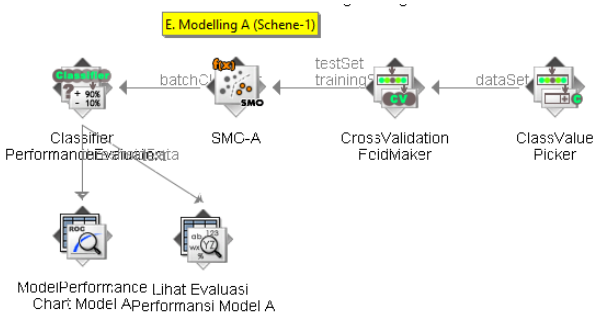
Klasifikasi dengan *SVM* yang dipergunakan pada *WEKA library(weka.classifiers.functions.SMO)* adalah dengan menerapkan *Sequential Minimum Optimation* (*SMO*) [16]. Pengaturan standar dilakukan proses normalisasi pada data latih. Karena sudah dilakukan tahap pra-pemrosesan pada himpunan data yang dipergunakan maka pilihan ini diabaikan seperti ditunjukkan pada Gambar 11. *Kernel* yang dipergunakan adalah *polynomial kernel (weka.classifiers.functions.supportVector.PolyKernel)* dengan nilai eksponen adalah 2 yang memperhitungkan Persamaan (6) dan Persamaan (7)

$$K_{(x,y)} = < x, y >^p \quad (6)$$

$$K_{(x,y)} = (< x, y > +1)^p \quad (7)$$



Gambar 11. *Kernel SMO*



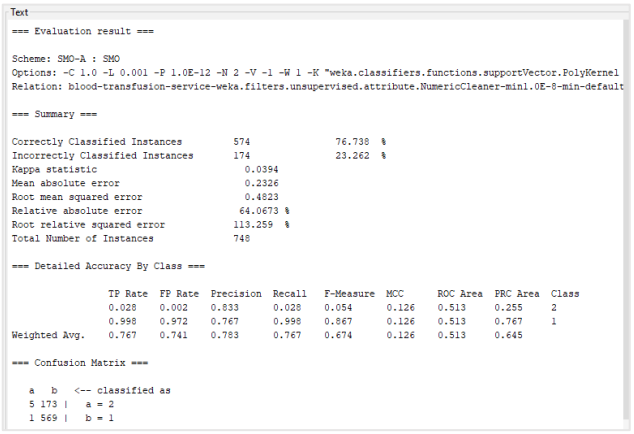
Gambar 12. *Layout* Pembelajaran Mesin Model A

Pada skenario-1 penyiapan pembelajaran mesin divisualkan pada Gambar 12, *data set* dengan rasio ketimpangan 1:3 dipergunakan sebagai data latih dan uji. Pada penelitian ini digunakan 10 *fold CV(Cross Validation)*. Dalam prosesnya tiap 10 *subset*, digunakan 9 *training fold* dan 1 *testing fold* oleh *Cross Validation*. Hasilnya didapatkan performa model seperti pada Tabel 4, dengan akurasi 76.7%, *precision* 78.3%, *recall* 76.7% F *score* 67.4%.

Tabel 4. *Confusion Matrix SVM* Skenario 1

	Aktual 1	**Aktual 2**
Prediksi 1	569	1
Prediksi 2	173	5

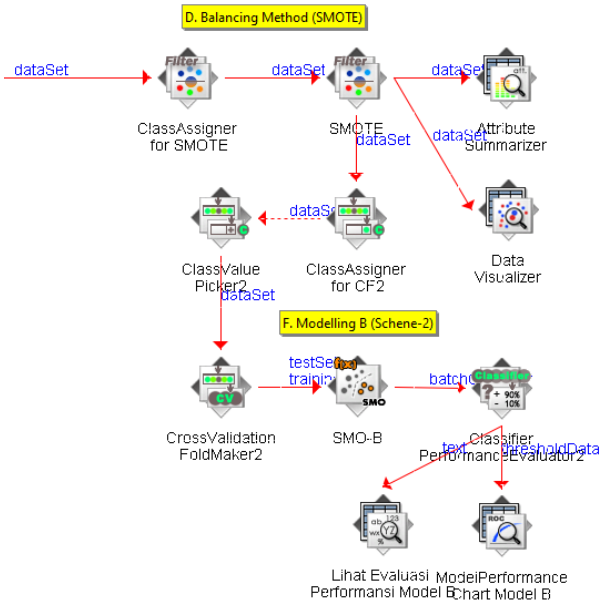
WEKA menampilkan *output* hasil perhitungan pada tiap *class* dan reratanya, lihat Gambar 13. Terdapat gap yang sangat jauh pada performa *recall* atau *sensitivity* (*True Positive Rate).* Bahwa *class* 1, memiliki *recall* 99.8% dan *class* 2 yang hanya memiliki *recall* 2.8%. Meskipun *weighted average* yang dihasilkan cukup tinggi sebesar 76.6%, namun model ini akan cenderung menghasilkan klasifikasi pada *class* 1 (tidak mendonorkan darah). Sementara *class* 2 (mendonorkan darah) cenderung diabaikan, karena *recall* pada *class* 2 ini sangat kecil 2.8%.



```
Text
=== Evaluation result ===

Scheme: SMO-A : SMO
Options: -C 1.0 -L 0.001 -P 1.0E-12 -N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel
Relation: blood-transfusion-service-weka.filters.unsupervised.attribute.NumericCleaner-min1.0E-8-min-default

=== Summary ===

Correctly Classified Instances         574               76.738  %
Incorrectly Classified Instances       174               23.262  %
Kappa statistic                          0.0394
Mean absolute error                      0.2326
Root mean squared error                  0.4823
Relative absolute error                 64.0673 %
Root relative squared error            113.259  %
Total Number of Instances              748

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.028    0.002    0.833      0.028     0.054      0.126     0.513     0.255     2
                 0.998    0.972    0.767      0.998     0.867      0.126     0.513     0.767     1
Weighted Avg.    0.767    0.741    0.783      0.767     0.674      0.126     0.513     0.645

=== Confusion Matrix ===

   a   b   <-- classified as
   5 173 |   a = 2
   1 569 |   b = 1
```

Gambar 13. Hasil Evaluasi Model A

Pada skenario-2, *data set* dengan rasio ketimpangan 1:1 dipergunakan untuk melatih dan menguji model yang dihasilkan, lihat Gambar 14.



Gambar 14. *Layout* Pembelajaran Mesin Model A

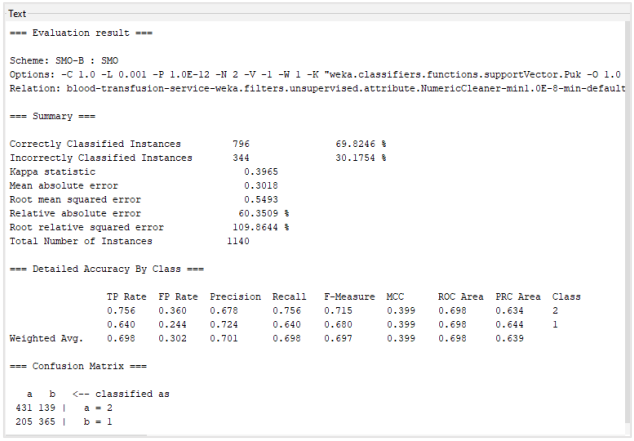
Data set tersebut dihasilkan melalui proses penambahan data sintesis pada *class* minor. Menerapkan teknik *Synthetic Minority Oversampling Technique (SMOTE)* salah satu varian *Random Over Sampling (weka.filters.supervised.instance.SMOTE)*[17]. Dengan pengaturan standar menggunakan nilai 5 pada jumlah tetangga terdekat (*nearestNeighbors)* dan peningkatan populasi minor 570 maka diberikan nilai 220 pada

pengaturan label prosentase. Diterapkan *CrossValidation* dengan 10 *fold*. Hasilnya didapatkan performa model, dengan akurasi 69.8%, presisi 70.1%, *recall* 69.8%, F *score* 69.7%. Nilai tersebut dengan memperhitungkan pada metrik *error* yang diperlihatkan pada Tabel 5.

Tabel 5. *Confusion Matrix SVM* Skenario 2

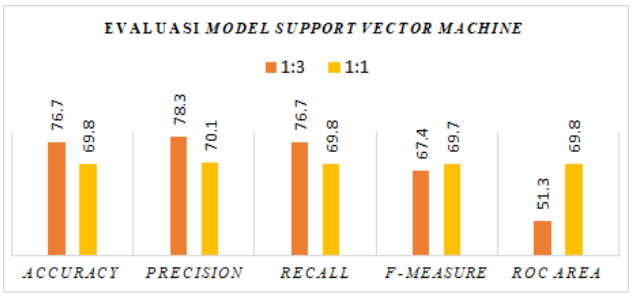
	Aktual 1	**Aktual 2**
Prediksi 1	365	205
Prediksi 2	139	431

Sepintas terlihat penurunan akurasi ~6% dibandingkan skenario 1. Namun jika dilihat lebih dalam pada Gambar 15, *class* 1 menunjukkan *recall* 64.0% dan *class* 2 dengan *recall* 75.6% Hal yang sangat berbeda pada skenario 1 adalah di mana *class* 1, memiliki *recall* 99.8% dan *class* 2 yang hanya memiliki *recall* 2.8%. Gap yang terlalu jauh ini (~97%) adalah fakta performa metrik *recall* pada skenario 1. Pada skenario 2, gap antar *class* pada metrik *recall* cenderung berimbang dan ketimpangannya tidak terlalu jauh seperti pada skenario 1.



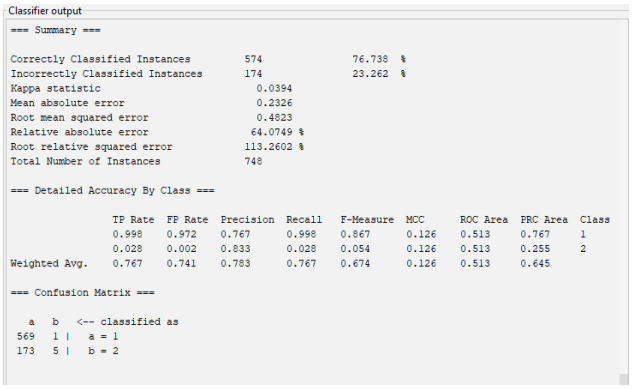
Gambar 15. Hasil evaluasi model B

Gambar 16 menampilkan perbandingan performa model klasifikasi yang dihasilkan oleh *SVM* dengan menggunakan *data set* dengan variasi *imbalanced ratio (IR)* yang berbeda. Untuk memudahkan penyebutan selanjutnya maka *IR* 1:3 adalah model A, *IR* 1:1 adalah model B. Dari 5 parameter performa, nilai model A lebih unggul ~6-8% di 3 parameter (akurasi, presisi dan *recall*). Namun pada 2 parameter lainnya *F-Score dan ROC Area* nilai model B lebih unggul.



Gambar 16. Grafik Performa Model

Weka menyediakan fitur *load/save* setelah model didefinisikan. Konfigurasi yang telah diatur di awal dapat dipanggil secara cepat tanpa melalui proses tahapan pengaturan awal untuk penyiapan model. Dalam simulasi diberikan data lokal pada kedua model tersebut dengan populasi 748 data sukarelawan yang telah memiliki label. Dalam simulasi pengujian, label dihilangkan untuk diprediksi oleh model A dan model B. Label sebagai fakta sekaligus pembanding terhadap hasil prediksi masing-masing model.



Gambar 17. Performa Model A

Lihat pada Gambar 17, hasilnya terdapat 174 hasil prediksi keliru terhadap data uji pada model A di mana 173 keliru memprediksi mendonorkan darah (faktanya tidak mendonor) dan 1 tidak mendonorkan darah (faktanya mendonor).

Gambar 18. Grafik Plot Hasil Prediksi Model A

Visualisasi plot distribusi hasil prediksi diperlihatkan pada Gambar 18. Terdapat 1 *outlier* label 1 yang diprediksi sebagai label 2. Namun pada label 2, *outlier* didominasi prediksi label 1 sebagai label 2, artinya lebih banyak kekeliruan hasil prediksi pada label 2.

Tabel 6. Hasil Pengujian Model A

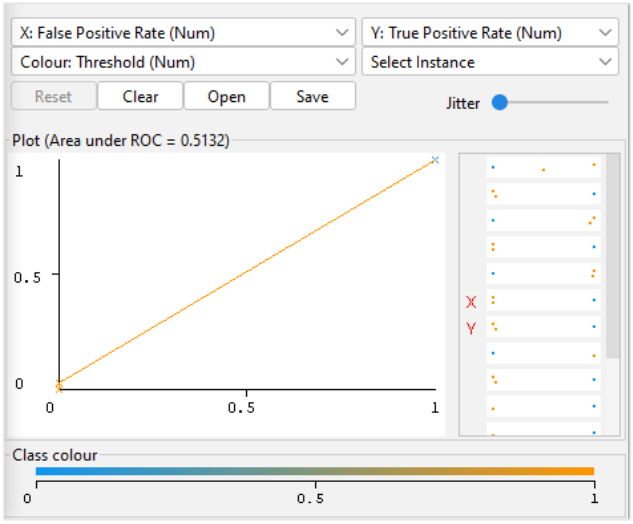
	Aktual 1	Aktual 2
Prediksi 1	569	173
Prediksi 2	1	5

Dengan menggunakan persamaan (2) maka didapatkan nilai akurasi untuk model A sebagai berikut,

$$Accuracy = \frac{569 + 5}{569 + 1 + 5 + 173}\ x100\%$$

$$Accuracy = \frac{574}{748} x100\% = 76.7\%$$

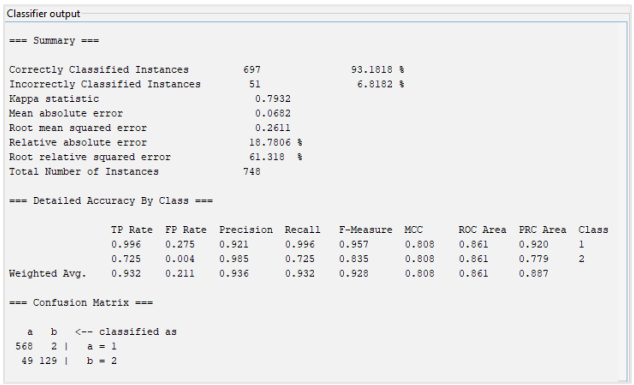
Terhadap data tes ini model A memiliki prediksi benar 76.7%, sesuai dengan *output* pengklasifikasi pada Gambar 17.



Gambar 19. Plot *ROC* Model A

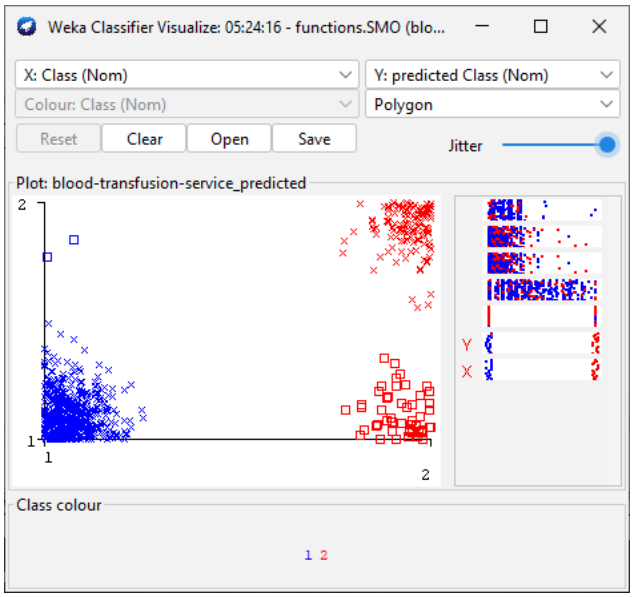
Diperlihatkan kurva *Receiver Operating Characteristic (ROC)* pada Gambar 19. Distribusi *ROC* cenderung mendekati garis *baseline*. Nilai *Area Under Curve (AUC)* = 0.5132, sesuai dengan perhitungan pada Gambar 17.

Skenario yang sama disimulasikan pada model B.



Gambar 20. Performa Model B

Dan hasil uji pada model B pada Gambar 20 didapatkan 51 prediksi keliru, yakni sebanyak 49 kejadian mendonorkan darah (seharusnya tidak mendonorkan) dan 2 kejadian tidak mendonorkan darah (seharusnya mendonorkan).



Gambar 21. Grafik Plot Hasil Prediksi Model B

Hasil visualisasi plot distribusi prediksi diperlihatkan pada Gambar 21. Terdapat 2 *outlier* label 1 yang diprediksi sebagai label 2. Pada label 2, prediksi pada label 2 sebagai label 1 jauh lebih sedikit (49) dibandingkan dengan hasil pada simulasi model A. Lihat pada Gambar 20 Nilai *recall* pada kelas 2 ini nilainya 75%. Jauh lebih baik dibanding dengan model A , lihat Gambar 17 dengan nilai *recall* 2.8%.

Tabel 7. Hasil Pengujian Model B

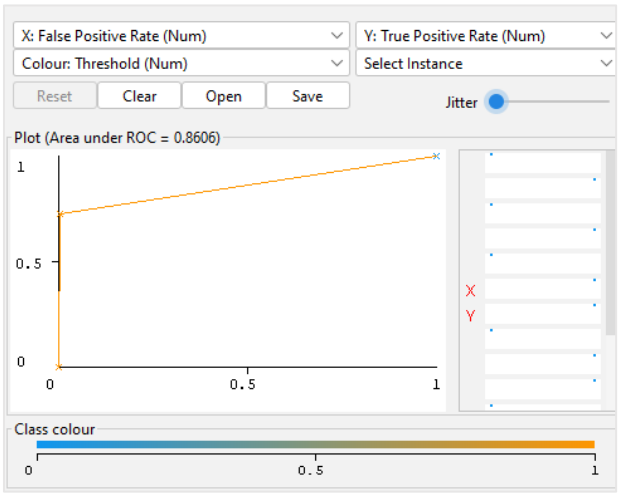
	Aktual 1	Aktual 2
Prediksi 1	568	49
Prediksi 2	2	129

Dengan menggunakan persamaan (2) maka didapatkan nilai akurasi untuk model B sebagai berikut,

$$Accuracy = \frac{568 + 129}{568 + 2 + 129 + 49}\ x100\%$$

$$Accuracy = \frac{697}{748} x100\% = 93.2\%$$

Terhadap data uji ini model B memiliki prediksi benar 93.2% sesuai dengan *output* pengklasifikasi model B pada Gambar 20.



Gambar 22. Plot *ROC* Model B

Diperlihatkan kurva *Receiver Operating Characteristic (ROC)* pada Gambar 22. Distribusi *ROC* cenderung menjauh di atas dari garis *baseline*. Nilai *Area Under Curve (AUC)* = 0.8606, sesuai dengan perhitungan pada Gambar 20.

4. KESIMPULAN

Terhadap tujuan penelitian, telah didapatkan *data set* yang seimbang dengan rasio ketimpangan 1:1. Himpunan data sebanyak 1.440 sampel, lihat Gambar 4 dan Gambar 10. Namun ini hanya digunakan untuk kebutuhan analisa, tidak untuk menggantikan *data set* donor darah sebelumnya pada repositori *UCI*.

Perbandingan evaluasi masing-masing model klasifikasi *SVM* telah dilakukan, lihat Gambar 13 dan Gambar 15. Mendasari atas hasil dari percobaan berulang, performa hasil latih yang didapatkan dengan *data set* yang tidak seimbang memang lebih tinggi dibandingkan dengan performa hasil latih setelah diterapkan *data set* yang seimbang. Performa tersebut lebih tinggi 6 hingga 8 persen pada metrik akurasi, presisi, dan *recall*.

Namun simulasi prediksi dengan 748 sampel pada model yang telah dihasilkan, lihat Tabel 6 dan Tabel 7, menunjukkan hal yang berbeda dalam performa akurasi prediksinya . Setelah dilakukan simulasi prediksi dengan populasi 748 sampel, terlihat akurasi dari model yang dilatih dengan data seimbang faktanya lebih tinggi 15.8%. Model hasil pembelajaran mesin yang dihasilkan melalui pelatihan dengan rasio ketimpangan 1:1 memiliki akurasi 93.2%. Dan model yang dilatih dengan data tidak seimbang memiliki akurasi 76.7%. *Area Under Curve (AUC)* pada model B (lihat Gambar 22) dengan *ROC*=0.8606 lebih baik dibandingkan dengan model A dengan *ROC*=0.5132 (lihat Gambar 19).

Hal ini menjelaskan bahwa keseimbangan data memiliki pengaruh terhadap prediksi yang dihasilkan dari suatu model. Pengaruh ini hendaknya diberikan perhatian di tahapan pra-pemrosesan data guna melatih mesin sehingga menghasilkan prediksi dengan akurasi yang baik di masing-masing kelas. Teknik *SMOTE dan* normalisasi *Min-Max Scaling* turut berkontribusi positif dalam menunjang data latih yang baik dalam pembelajaran mesin.

Untuk mendukung lebih dalam optimasi keseimbangan data, variasi pelatihan dan pengujian dapat dilakukan dengan rasio ketimpangan yang lebih variatif. Kombinasi *SMOTE* dengan varian *ROS* lainnya berpotensi mendapatkan data sintesis yang lebih baik pada pembelajaran mesin.

DAFTAR PUSTAKA

[1] I. C. Yeh, K. J. Yang, and T. M. Ting, “Knowledge discovery on RFM model using Bernoulli sequence,” *Expert Syst Appl*, vol. 36, no. 3, pp. 5866–5871, Apr. 2009, doi: 10.1016/J.ESWA.2008.07.018.

[2] C. Kurniawan Putra Rukma, “Increase Accuracy of Naïve Bayes Classifier Algorithm with K-Means Clustering for Prediction of Potential Blood Donors,” *Journal of Advances in Information Systems and Technology*, vol. 4, no. 1, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jaist

[3] K. Akbar and M. Hayaty, “Data Balancing untuk Mengatasi Imbalance Data set pada Prediksi Produksi Padi Balancing Data to Overcome Imbalance Data set on Rice Production Prediction,” *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 2, no. 02, pp. 1–14, 2020.

[4] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, “Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance,” *Applied Acoustics*, vol. 196, p. 108882, Jul. 2022, doi: 10.1016/J.APACOUST.2022.108882.

[5] S. Mutmainah, “PENANGANAN IMBALANCE DATA PADA KLASIFIKASI KEMUNGKINAN PENYAKIT STROKE,” 2021. [Online]. Available: https://library.uii.ac.id/osr

[6] F. Yulian Pamuji, “Pengujian Metode SMOTE Untuk Penanganan Data Tidak Seimbang Pada Data set Binary,” *Seminar Nasional Sistem Informasi*, vol. 2022, 2022.

[7] R. Kembang Hapsari and T. Surabaya, “SNESTIK Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika Implementasi Algoritma SMOTE Sebagai Penyelesaian Imbalance Hight Dimensional Data sets,” p. 427, doi: 10.31284/p.snestik.2022.2868.

[8] E. Sutoyo, M. Asri Fadlurrahman, J. Telekomunikasi Jl Terusan Buah Batu, K.

Dayeuhkolot, K. Bandung, and J. Barat, “JEPIN (Jurnal Edukasi dan Penelitian Informatika) Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network”.

[9] R. Siringoringo, “KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR,” 2018.

[10] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/matrik.v21i3.1726.

[11] A. S. Alkahtani and M. Jilani, “Predicting return donor and analyzing blood donation time series using data mining techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019.

[12] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00516-9.

[13] Y. A. Sir and A. H. H. Soepranoto, “Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas,” *Jurnal Komputer dan Informatika*, vol. 10, no. 1, pp. 31–38, Mar. 2022, doi: 10.35508/jicon.v10i1.6554.

[14] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, “RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5059–5074, Sep. 2022, doi: 10.1016/j.jksuci.2022.06.005.

[15] N. Sekar Ramadhanti and W. Ananta Kusuma, “OPTIMASI DATA TIDAK SEIMBANG PADA INTERAKSI DRUG TARGET DENGAN SAMPLING DAN ENSEMBLE SUPPORT VECTOR MACHINE,” vol. 7, no. 6, 2020, doi: 10.25126/jtiik.202072857.

[16] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” 1998.

[17] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. S. Philip Kegelmeyer, “synthetic minority over-sampling Technique,” *J Artif Intell Res*, p. 16, 2018.