Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4.5, Random Forest, dan SVM

Resampling Technique for Handling Class Imbalance in the Classification of Diabetes using C4.5, Random Forest, and SVM

Wahyu Nugraha[1], Raja Sabaruddin[2]
[1,2,]Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika PSDKU Pontianak
E-mail: [1]wahyu.whn@bsi.ac.id, [2]raja.rjd@bsi.ac.id

Abstrak

Penderita diabetes di seluruh dunia terus mengalami peningkatan dengan angka kematian sebesar 4,6 juta pada tahun 2011 dan diperkirakan akan terus meningkat secara global menjadi 552 juta pada tahun 2030. Pencegahan Penyakit diabetes mungkin dapat dilakukan secara efektif dengan cara mendeteksinya sejak dini. Data mining dan machine learning terus dikembangkan agar menjadi alat yang handal dalam membangun model komputasi untuk mengidentifikasi penyakit diabetes pada tahap awal. Namun, masalah yang sering dihadapi dalam menganalisis penyakit diabetes ialah masalah ketidakseimbangan class. Kelas yang tidak seimbang membuat model pembelajaran akan sulit melakukan prediksi karena model pembelajaran didominasi oleh instance kelas mayoritas sehingga mengabaikan prediksi kelas minoritas. Pada penelitian ini kami mencoba menganalisa dan mencoba mengatasi masalah ketidakseimbangan kelas dengan menggunakan pendekatan level data yaitu teknik resampling data. Eksperimen ini menggunakan R language dengan library ROSE (version 0.0-4). Dataset Pima Indians dipilih pada penelitian ini karena merupakan salah satu dataset yang mengalami ketidakseimbangan kelas. Model pengklasifikasian pada penelitian ini menggunakan algoritma decision tree C4.5, RF (Random Forest), dan SVM (Support Vector Machines). Dari hasil eksperimen yang dilakukan model klasifikasi SVM dengan teknik resampling yang menggabungkan over dan under-sampling menjadi model yang memiliki performa terbaik dengan nilai AUC (Area Under Curve) sebesar 0.80

Kata kunci: Resampling, Ketidakseimbangan Kelas, Klasifikasi, Area Under Curve (AUC)

Abstract

People with diabetes worldwide continue to experience an increase with a death rate of 4.6 million in 2011 and is expected to continue to increase globally to 552 million by 2030. Prevention of diabetes may be done effectively by detecting it early. However, the problem that is often faced in analyzing diabetes is the class imbalance problem. An unbalanced class makes it difficult for the learning model to make predictions because the learning model is dominated by instances of the majority class, thus ignoring the predictions of the minority class. In this study, we try to analyze and try to overcome the problem of class imbalance by using a data level approach, namely data resampling techniques. This experiment uses the R language with the ROSE library (version 0.0-4). The Pima Indians dataset was chosen in this study because it is one of the datasets that experience class imbalance. The classification model in this study uses the decision tree algorithm C4.5, RF (Random Forest), and SVM (Support Vector Machines). From the results of experiments conducted on the SVM classification model with a resampling technique that combines over and under-sampling into a model that has the best performance with an AUC (Area Under Curve) value of 0.80

Keywords: Resampling, Class Imbalance, Classification, Area Under Curve (AUC)

1. PENDAHULUAN

Terdapat 347 juta penderita diabetes di seluruh dunia dengan angka kematian sebesar 4,6 juta pada tahun 2011 yang disebakan oleh penyakit diabetes. Jumlah penderita diabetes diperkirakan akan terus meningkat secara global menjadi 552 juta pada tahun 2030 dan diperkirakan akan menjadi peringkat ketujuh penyebab utama kematian [1]. Menurut *International Diabetes Federation (IDF)* jumlah penderita diabetes mendekati 463 juta. Angka tersebut diperoleh dari hasil surver pada tahun 2019, bahkan peneliti memperkirakan jumlah penderita diabetes bisa terus meningkat menjadi 642 juta [2]. Selain itu, Organisasi Kesehatan Dunia atau disingkat WHO menyatakan bahwa ada sekitar 1,6 juta orang meninggal diakibatkan oleh diabetes setiap tahunnya [3]. Diabetes merupakan salah satu penyakit yang disebabkan oleh kadar gula darah pada tubuh manusia sangat tinggi. Penyakit diabetes terjadi ketika penkreas tidak cukup untuk memproduksi insulin atau biasa disebut diabetes tipe 1. Akan tetapi, ketika tubuh tidak dapat menggunakan insulin dengan baik maka ini disebut sebagai diabetes tipe 2 [2][3]. Pencegahan Penyakit diabetes mungkin dapat dilakukan secara efektif dengan cara mendeteksinya sejak dini sehingga orang akan menjalankan pola hidup sehat untuk mencegah penyakit diabetes [4]. *Data mining* dan *machine learning* terus dikembangkan agar menjadi alat yang handal dalam membangun model komputasi untuk mengidentifikasi penyakit diabetes pada tahap awal [5].

Algoritma *machine learning* dapat menemukan pola pembelajaran tertentu dari dataset sehingga dapat melakukan prediksi dengan cukup akurat terhadap penyakit diabetes [3]. Namun, masalah utama yang sering terjadi dalam menganalisis data medis adalah masalah ketidakseimbangan *class*. Sebuah dataset bisa dianggap tidak seimbang jika kategori klasifikasi sangat kurang terwakili atau distribusi kelas sangat tidak seimbang [6]. Beberapa penelitian sebelumnya telah menggunakan metode *machine learning* untuk memprediksi penyakit diabetes menggunakan dataset Pima Indian diabetes karena merupakan salah satu dataset yang mengalami ketidakseimbangan kelas [3]. Dataset Pima Indian diperoleh dari repositori UCI yang memiliki 9 attribut dengan total data sebanyak 768 data [7]. Ketidakseimbangan *class* pada dataset membuat model pembelajaran akan sulit melakukan prediksi karena model pembelajaran didominasi oleh *instance* kelas mayoritas sehingga mengabaikan prediksi kelas minoritas [8]. Metode untuk mengatasi ketidakseimbangan kelas dibagi menjadi empat pendekatan, yaitu: *algorithmic level, cost-sensitive, data level*, dan *ensembles of classifiers* [9]. Pendekatan level data banyak menjadi bahan pertimbangan untuk penelitian pada berbagai literatur [10]. Namun, akan sangat sulit untuk menentukan rasio *resampling* yang optimal secara otomatis [9].

Resampling merupakan teknik dengan pendekatan *level* data yang digunakan untuk mengatasi ketidakseimbangan kelas dengan mengeliminasi beberapa data dari kelas mayoritas (undersampling) atau menambahkan beberapa data menggunakan hasil dari proses *generated* atau duplikat data ke kelas minoritas *(oversampling)* [11]. Salah satu teknik oversampling yang cukup berhasil dalam menghasilkan data baru dari kelas minoritas adalah *SMOTE (Synthetic Minority Over-sampling Technique).* Metode sampling lain dapat digunakan untuk mengatasi masalah ketidakseimbangan kelas salah satunya dengan menggabungkan teknik over-sampling dan *under-sampling* secara bersamaan dengan cara menghapus kelebihan dari kelas mayoritas dan menambahkan jumlah kelas minoritas sehingga menghasilkan distribusi kelas yang seimbang [7].

Pada penelitian ini kami mencoba menganalisa dan mencoba mengatasi masalah ketidakseimbangan kelas dengan menggunakan pendekatan level data yaitu teknik *resampling data.* Eksperimen ini menggunakan *R language* dengan *library ROSE (version 0.0-4)* di mana pada library tersebut terdapat metode resampling seperti *over-sampling, under-sampling, combination of over- and under-sampling* serta *Generation of synthetic data by Randomly Over Sampling Examples (ROSE).* Keempat metode sampling ini akan digunakan untuk mengatasi masalah ketidak seimbangan kelas pada dataset Pima Indian. Model pengklasifikasian pada penelitian ini menggunakan algoritma *decision tree C4.5, RF (Random Forest), dan SVM (Support Vector Machines).* Dari hasil eksperimen yang dilakukan akan dilakukan komparasi

secara keseluruhan untuk mencari model terbaik berdasarkan hasil mengukuran. Pengukuran terhadap performa model menggunakan nilai *Area Under Curve (AUC)* yang diperoleh dari tabel *confusion matrix. Confusion matrix* dapat memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah agar didapat nilai akurasi, *sensitivity, specificity dan Area Under Curve (AUC)* [12].

2. METODE PENELITIAN

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah data yang terkait penyakit diabetes dan juga terkait dengan masalah ketidakseimbangan kelas. Dataset yang digunakan berasal dari *Kaggle dataset repository* (UCI Pima Indians Diabetes Database) yang diunduh dari https://www.kaggle.com/uciml/pima-indians-diabetes-database. Dataset ini memiliki 9 attribut dengan total sampel valid sebanyak 768 sample. Nilai attribut dataset ini berasal dari semua wanita berusia minimal 21 tahun [1]. Dataset terdiri dari 8 variabel prediktor medis dan satu variabel target yaitu *Outcome.* Penelitian lain yang telah menggunakan dataset Pima Indians ini diantaranya dilakukan oleh Kumar dengan penelitian mengenai prediksi penyakit diabetes melitus menggunakan *Deep Neural Networks classifier* [2], Khanam dengan penelitian mengenai perbandingan algoritma *machine learning* untuk memprediksi penyakit diabetes [3], dan Hayashi dengan penelitian mengenai *Rule extraction* menggunakan algoritma *Recursive-Rule extraction* dengan J48graft yang dikombinasikan teknik pemilihan sampel untuk diagnosis penyakit diabetes melitus tipe 2 pada dataset Pima Indians [1].

2.2 Data Preparation

Data preparation bertujuan untuk mendapatkan data yang siap diolah dan berkualitas baik [13]. Beberapa teknik *data preparation* diantaranya:

1. *Data validation* digunakan untuk untuk mengatasi masalah *Incompleteness* atau data yang mengalami *missing value.* Gambar 1 menunjukkan hasil eksekusi bahwa dataset ini sudah tidak ditemukan adanya *missing value.*

```
10 ▾ ```{r}
11   data.frame(apply(data, 2, function(x){sum(is.na(x))}))
12 ▴ ```
```

	apply.data..2..function.x... <int>
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

9 rows

Gambar 1 Contoh Data Tanpa Missing Value

2. *Data integration* dan *transformation*, Data yang bernilai kategorikal ditransformasikan ke dalam angka misalkan seperti attribut *Outcome (positive, negative)* diubah kedalam bentuk angka (0,1). Selain itu konversi tipe data ke bentuk format numerik juga diperlukan agar dataset dapat diuji coba menggunakan model klasifikasi. Gambar 2 menunjukkan bahwa data telah siap digunakan.

```
13 ```{r}
14 str(data)
15 ```
```

```
'data.frame':   768 obs. of  9 variables:
 $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

Gambar 2 Contoh Data yang Sudah Siap Diuji

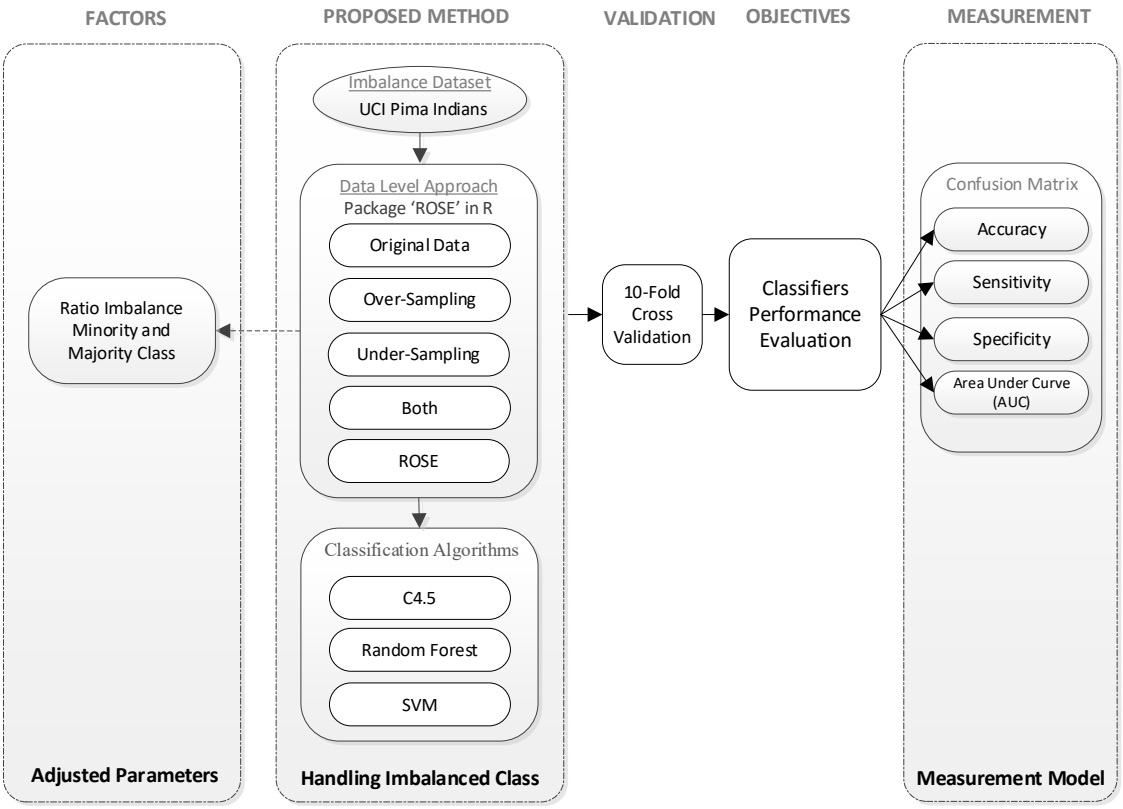
2.3 Ketidakseimbangan Kelas (Class Imbalance)

Ketidakseimbangan kelas maksudnya adalah dataset yang menunjukkan perbedaan yang signifikan antara kelas yang satu dengan yang lainnya. Secara umum pemahaman tentang ketidakseimbangan kelas berkaitan dengan situasi dimana beberapa *instance* kelas dari data sangat tidak terwakili jika dibandingkan dengan kelas yang lain [14]. Penelitian mengenai *imbalanced class* sering menganggap bahwa rata-rata jumlah minoritas class 10% sampai 20%. Kenyataannya, dataset biasanya jauh lebih tidak seimbang.

2.4 Pemodelan

Pada penelitian ini untuk mengatasi masalah ketidakseimbangan kelas diusulkan model dengan pendekatan *level data* yaitu menggunakan 4 teknik resampling yang terdapat pada *package ROSE: A Package for Binary Imbalanced Learning pada R language.* Gambaran dari kerangka pemikiran atau model tahapan penelitian dapat dilihat pada Gambar 3.

Eksperimen yang dilakukan pada dataset diabetes harus dilakukan secara adil terutama dalam menentukan rasio *resampling* yang optimal pada semua pengujian. Jadi semua *data training* pada dataset persentase *imbalance* datanya harus cukup imbang. Perbandingan jumlah antara kelas mayoritas dan minoritas data training pada dataset diabetes Pima Indian ditunjukkan pada Tabel 1. Hasil pengujian akan diukur menggunakan *confusion matrix* atau matrik konfusi ysng dapat dilihat pada Tabel 2 untuk mendapatkan nilai akurasi, *sensitivity, specificity dan Area Under Curve (AUC)* yang dinyatakan dalam pernyataan 1, 2, 3, dan 4. *Confusion matrix* memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah [12].



Gambar 3 Model Tahapan Penelitian

Tabel 1 Rasio *Imbalance Class Training* Set Pima Indian

Metode Sampling	Jumlah instance	
	Positif	Negatif
Original	206	409
Over-sampling	401	409
Under-sampling	206	204
Both	296	319
ROSE	296	319

Tabel 1 *Model Confusion Matrix*

Classification	*Predicted Class*		
		Class = YES	*Class = NO*
Observed Class	*Class = YES*	**A** (*true positive*-TP)	**B** (*false negative*-FN)
	Class = NO	**C** (*false positive*-FP)	**D** (*true negative*-TN)

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

$$Sensitivity = \frac{TP}{TP + FP} \quad (2)$$

$$Specificity = \frac{TN}{FN + TN} \quad (3)$$

$$Balanced\ Accuracy/AUC = \frac{Sensitivity + Specificity}{2} \qquad (4)$$

2.5 Validation and Evaluation Model

Model validasi untuk *learning* dan *testing* menggunakan *10 fold cross-validation. Data training* dibagi menjadi 10 bagian yang sama kemudian proses pembelajaran *(learning)* dilakukan sebanyak jumlah k yaitu 10 kali. Dari Tabel 3 dapat dilihat bahwa setiap kali dipilih satu bagian sebagai *data testing.* Maka, data yang lain sebanyak 9 bagian digunakan sebagai data pembelajaran *(learning).* Setelah itu, dihitung nilai rata-rata dan nilai penyimpangan *(deviation value)* dari 10 kali jumlah pengujian yang berbeda [15]. Penelitian ini menggunakan *k-fold cross validation* karena metode ini telah menjadi metode validasi standar dan *state-of-the-*art [16].

Evaluasi terhadap kinerja model pada eksperimen ini menggunakan *area under curve (AUC)* sebagai indikator penentu performa dari model klasifikasi. Data yang diuji baik yang benar maunpun yang salah dihitung menggunakan *confusion matrix.* Dari tabel konfusi ini nantinya diperoleh nilai penunjang lain diantaranya akurasi, *sensitivity, specificity,* dan *Area Under Curve (AUC).* Lessmann menganjurkan penggunaan AUC untuk meningkatkan *crossstudy comparability* [17]. Nilai AUC pada masing-masing dataset digunakan sebagai penentuan metode yang menghasilkan akurasi yang paling tinggi. Penggunaan nilai AUC sebagai bahan evaluasi kinerja model disajikan oleh Gorunescu [15]. Interpretasi dari klasifikasi nilai AUC terhadap pengujian dapat dilihat pada Tabel 4.

Tabel 2 *Stratified 10 Fold Cross Validation*

n-validation	**Dataset's Partition**									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Tabel 3 Klasifikasi Nilai AUC

Nilai AUC	Klasifikasi
0.9 - 1	Excellent classification
0.8 - 0.9	Good classification
0.7 - 0.8	Fair classification
0.6 - 0.7	Poor classification
< 0.6	Failure

3. HASIL DAN PEMBAHASAN

Penerapan metode klasifikasi C4.5, *Random Forest dan SVM* pada dataset dilakukan apa adanya tanpa proses *sampling*. Kemudian akan dibandingkan dengan model yang menggunakan teknik *sampling* yang ada pada *library rose* pada *R language* seperti *over-sampling, under-sampling*, kombinasi dari *over-sampling* dan *under-sampling* serta *Generation of synthetic data by Randomly Over Sampling Examples.* Dari hasil pengujian yang telah dilakukan diperoleh nilai *confusion matrix* dari masing-masing algoritma klasifikasi yang dapat dilihat pada tabel 5, tabel 6, dan tabel 7. Nilai yang diperoleh dari matriks konfusi adalah akurasi, *sensitivity, specificity* dan *Area Under Curve.* Penelitian ini fokus pada deteksi kelas

minoritas sehingga *sensitivity* dan *specificity* dapat digunakan untuk menunjukkan performa dari dua kelas. *Cut off* dari sensitivitas dan spesifisitas bisa digunakan untuk membuat kurva ROC untuk mendapatkan nilai dari area under curve AUC [18]. Eksperimen dilakukan menggunakan perangkat komputasi Intel core i5 gen 6 dengan operasi sistem windows 10. *Development Environment* menggunakan IDE R Studio 1.4 dan bahasa pemograman *R Language versi 4.0.5* serta menggunakan beberapa *packages* atau *library* yang tersedia di https://cran.r-project.org/.

Metode C4.5 pada Tabel 5 menunjukkan bahwa teknik oversampling memiliki performa terbaik dibandingkan dengan teknik *sampling* yang lain maupun tanpa *resampling* dengan nilai AUC 0.75. Disisi lain jika melihat hasil prediksi kebenaran kelas minoritas, *oversampling* berhasil mencapai 47 aktual positif namun mengorbankan nilai aktual negatif. Hal ini dikarenakan sangat sulit untuk membuat model pembelajaran yang efektif jika distribusi kelas dalam kumpulan data training yang digunakan tidak seimbang hingga berdampak pada menurunnya kinerja algoritma [10]. Metode *RF (Random Forest)* pada Tabel 6 menunjukkan bahwa teknik *over-sampling* juga memiliki performa terbaik dengan nilai AUC 0.73. Jika dibandingkan dengan C4.5 metode *Random forest* mengalami penurunan pada kebenaran aktual positif menjadi 39 namun, kebenaran terhadap aktual negatif meningkat menjadi 77. Namun, jika dilihat secara umum metode RF memiliki performa sedikit lebih baik dari C4.5 dalam menangani masalah ketidakseimbangan kelas berdasarkan nilai AUC yang diperoleh.

Metode *SVM (Support Vector Machines)* pada Tabel 7 menunjukkan bahwa teknik yang menggabungkan antara *under* dan *over-sampling* memiliki performa terbaik dengan nilai AUC 0.80. Disisi lain jika melihat hasil prediksi kebenaran kelas minoritas atau kelas positif, teknik both ini berhasil mencapai angka kebenanran aktual positif sebesar 49 dan angka kebenaran aktual negatif 74. Jika diambil nilai rata-rata maka metode SVM memiliki performa lebih baik dibanding C4.5 dan RF dalam mendeteksi penyakit diabetes pada *dataset Pima Indians* baik sebelum menerapkan teknik resampling maupun setelah menggunakan teknik *sampling.*

Tabel 4 *Confusion Matrix* Algoritma C4.5

Metode Sampling	Prediksi	Aktual		Akurasi	Sensitivity	Specificity	AUC
		Positif	Negatif				
Original	**Positif**	33	9	0.7516	0.5323	0.9011	0.7167
	Negatif	29	82				
Over-Sampling	**Positif**	47	23	0.7516	0.7581	0.7473	0.7527
	Negatif	15	68				
Under-sampling	**Positif**	36	21	0.6928	0.5806	0.7692	0.6749
	Negatif	26	70				
Both	**Positif**	36	20	0.6993	0.5806	0.7802	0.6804
	Negatif	26	71				
ROSE	**Positif**	34	11	0.7451	0.8791	0.5484	0.7138
	Negatif	28	80				

Tabel 5 *Confusion Matrix* Algoritma *Random Forest*

Metode Sampling	Prediksi	Aktual		Akurasi	Sensitivity	Specificity	AUC
		Positif	Negatif				
Original	**Positif**	34	12	0.7386	0.5484	0.8681	0.7083
	Negatif	28	79				
Over-Sampling	**Positif**	39	14	0.7582	0.6290	0.8462	0.7376
	Negatif	23	77				
Under-sampling	**Positif**	43	25	0.7124	0.6935	0.7253	0.7094
	Negatif	19	66				

Both	**Positif**	41	19	0.7386	0.7912	0.6613	0.7262
	Negatif	21	72				
ROSE	**Positif**	41	20	0.732	0.6613	0.7802	0.7208
	Negatif	21	71				

Tabel 6 *Confusion Matrix* Algoritma SVM

Metode Sampling	**Prediksi**	**Aktual**		**Akurasi**	**Sensitivity**	**Specificity**	**AUC**
		Positif	**Negatif**				
Original	**Positif**	34	7	0.7712	0.5484	0.9231	0.7357
	Negatif	28	84				
Over-Sampling	**Positif**	42	12	0.7908	0.6774	0.8681	0.7728
	Negatif	20	79				
Under-sampling	**Positif**	44	16	0.7778	0.7097	0.8242	0.7669
	Negatif	18	75				
Both	**Positif**	49	17	0.8039	0.7903	0.8132	0.8018
	Negatif	13	74				
ROSE	**Positif**	46	15	0.7974	0.7419	0.8352	0.7886
	Negatif	16	76				

Tabel 7 Nilai *AUC (area under curve)* Metode Klasifikasi

Algoritma Klasifikasi	**Metode Resampling**				
	Original	**Over-sampling**	**Under-sampling**	**Both**	**ROSE**
C4.5	0.7167	0.7527	0.6749	0.6804	0.7138
Random Forest	0.7083	0.7376	0.7094	0.7262	0.7208
SVM	0.7357	0.7728	0.7669	0.8018	0.7886

4. KESIMPULAN DAN SARAN

Sebagian studi atau penelitian terkait masalah ketidakseimbangan kelas mencoba untuk mengatasinya menggunakan pendekatan level data yaitu *random undersampling* dan *oversampling.* Pengembangan metode dalam hal untuk melakukan teknik *resampling* juga semakin berkembang.

Pada penelitian ini peneliti menggunakan teknik *resampling* dasar menggunakan salah satu *library* yang sering digunakan pada *R Language* yaitu *Package 'ROSE' (Random Over-Sampling Examples)* v 0.0-4 yang di dalamnya terdapat fungsi dari teknik *sampling* lain yang telah kami gunakan. Performa yang dihasilkan oleh *function* dari paket ini memang terlihat lebih baik dari sampling standar yang ada di R. Kesimpulan dari penelitian ini dapat dilihat pada Tabel 8 dimana performa tertinggi untuk prediksi penyakit diabetes ini diperoleh dengan menggunakan teknik kombinasi over dan under-sampling (Both) dengan metode klasifikasi SVM dengan nilai AUC 0.80.

Nilai terendah didapat oleh teknik *under-sampling* dengan metode klasifikasi C4.5 dengan nilai AUC 0.67. Selain itu, secara keseluruhan metode klasifikasi SVM menunjukkan performa sedikit lebih baik dibandingkan dengan metode C4.5 dan RF dalam mengatasi ketidakseimbangan kelas baik itu setelah menerapkan teknik *resampling* maunpun tanpa *resampling.* Penelitian selanjutnya sebaiknya lebih menekankan untuk melakukan pendekatan komprehensif dalam mempelajari teknik *resampling* baik *under* dan *over-sampling.*

Teknik pengembangan dari *under* dan *over-sampling* bisa dilakukan seperti menggunakan teknik *clustering base undersampling* atau perbaikan dari metode *SMOTE (Synthetic Minority Over-sampling Technique)* seperti *A self-adaptive robust SMOTE* yang dikembangkan oleh Chen [19].

DAFTAR PUSTAKA

[1] Y. Hayashi and S. Yukita, "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset," *Informatics Med. Unlocked*, vol. 2, pp. 92–104, 2016, doi: 10.1016/j.imu.2016.02.001.

[2] B. P. Manoj Kumar, S. R. Perumal, and N. R. K, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier," *Int. J. Cogn. Comput. Eng.*, vol. 1, pp. 55–61, 2020, doi: 10.1016/j.ijcce.2020.10.002.

[3] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, no. xxxx, 2021, doi: 10.1016/j.icte.2021.02.004.

[4] S. A. Kaveeshwar and J. Cornwall, "The current state of diabetes mellitus in India," *Australas. Med. J.*, vol. 7, no. 1, pp. 45–48, 2014, doi: 10.4066/AMJ.2014.1979.

[5] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.icte.2018.10.005.

[6] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[7] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[8] W. Nugraha, M. S. Maulana, and A. Sasongko, "Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 1–6, 2020, doi: 10.1088/1742-6596/1641/1/012014.

[9] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, pp. 1653–1672, 2015, doi: 10.1016/j.patcog.2014.10.032.

[10] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Inf. Sci. (Ny).*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.

[11] M. M. Rahman and D. N. Davis, "Cluster Based Under-Sampling for Unbalanced Cardiovascular Data," *Proc. World Congr. Eng. 2013*, vol. 3, pp. 1–6, 2013.

[12] F. Gorunescu, *Data mining: concepts and techniques*. Berlin, 2011.

[13] C. Vercellis, *Business Intelligence : Data Mining and Optimization for Decision Making*. John Wiley & Sons, Ltd, 2009.

[14] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2019, doi: 10.1109/TKDE.2008.239.

[15] R. S. Wahono, N. S. Herman, and S. Ahmad, "A Comparison Framework of Classification Models for Software Defect Prediction," vol. 20, no. 10, pp. 1945–1950, 2014, doi: 10.1166/asl.2014.5640.

[16] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. Elsevier Inc, 2011.

[17] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings," *IEEE Trans. Softw. Eng.*, vol. 34, no. 4, pp. 485–496, 2008, doi: 10.1109/TSE.2008.35.

[18] M. Kuhn, "Building Predictive Models in R Using the caret Package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, 2008, [Online]. Available: http://www.jstatsoft.org/v28/i05/paper.

[19] B. Chen, S. Xia, Z. Chen, B. Wang, and G. Wang, "RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise," *Inf. Sci. (Ny).*, vol. 553, pp. 397–428, 2021, doi: 10.1016/j.ins.2020.10.013.