JURNAL LARIK
LADANG ARTIKEL ILMU KOMPUTER
http://jurnal.bsi.ac.id/index.php/larik

KOMBINASI TOMEK-LINK DAN SMOTE UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA CREDIT CARD FRAUD

Wahyu Nugraha[1], Deni Risdiansyah[2], Deasy Purwaningtias[3], Taufik Hidayatulloh[4], Satia Suhada[5]

[1,2,3]Program Studi Sistem Informasi , Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika Kampus Pontianak
Jl. Abdurrahman Saleh No.18A, Bansir Darat, Pontianak Tenggara, Kota Pontianak, Kalimantan Barat, Telp. (0561) 583924
[4,5]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika Kampus Sukabumi
Jl. Cemerlang No.8, Sukakarya, Warudoyong, Kota Sukabumi, Jawa Barat, Telp. (0266) 6251992
wahyu.whn@bsi.ac.id[1], deni.drx@bsi.ac.id[2], deasy.dwg@bsi.ac.id[3], taufik.tho@bsi.ac.id[4], satia.shq@bsi.ac.id[5]

Abstrak-- Meningkatnya aktivitas perdagangan secara *online* atau *e-commerce* telah menjadi trend saat ini. Akibatnya kejahatan yang paling sering terjadi adalah penipuan kartu kredit (*credit card fraud*) atau *carding*. Kurang lebih terdapat 1.000 kasus penipuan dalam satu juta transaksi sehingga data tersebut dikumpulkan dalam bentuk *dataset credit card fraud risk*. Pada beberapa kasus, kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas seperti pada kasus transaksi *credit card*. Pada penelitian ini untuk menangani masalah ketidakseimbangan kelas pada *dataset credit card fraud risk* maka diusulkan metode *resampling* yaitu pendekatan level data Tomek-Link dan SMOTE dengan model klasifikasi C5.0. Penelitian ini dilakukan untuk meningkatkan nilai akurasi AUC pada model algoritma klasifikasi C5.0. Hasil penelitian menunjukkan bahwa metode usulan mampu meningkatkan nilai AUC sebesar 0,134 dibandingkan tanpa metode *resampling*.

Kata kunci: Ketidakseimbangan kelas , Oversampling, Tomek-Link, SMOTE, C5.0

***Abstrac**t - Increasing online trading activities or e-commerce has become a trend today. As a result the most common crime is credit card fraud or carding. There are approximately 1,000 cases of fraud in one million transactions so that data is collected in the form of datasets of credit card fraud risk. In some cases, minority classes are more important to identify than the majority class as in the case of credit card transactions. In this study to deal with the problem of class imbalances on credit card fraud risk datasets, the proposed resampling method is the Tomek-Link and SMOT data level with the C5.0 classification model. This research was conducted to improve the accuracy of AUC in the C5.0 classification algorithm model. The results showed that the proposed method was able to increase the AUC value of 0.134 compared to without the resampling method.*

Keywords: Imbalanced Class , Oversampling, Tomek-Link, SMOTE, C5.0

I. PENDAHULUAN

Meningkatnya aktivitas perdagangan secara online atau yang biasa disebut *e-commerce* ini begitu pesat dan menjadi trend saat ini. Hal ini disebabkan pertumbuhan perekonomian yang membaik dan tumbuhnya kelas menengah. Di Indonesia saja dilaporkan dari Bank Dunia bahwa 56,5 persen populasi Indonesia atau sekitar 134 juta jiwa masuk kategori kelas menengah dengan nilai belanja 2-20 dollar AS per hari. Kelompok kelas menengah ini berpenghasilan relatif tinggi, melek teknologi, dan selalu terhubung dengan internet. Seiring dengan berkembangnya *e-commerce*, juga terbuka peluang munculnya tindakan-tindakan anti-sosial dan perilaku kejahatan yang sebelumnya dianggap tidak mungkin terjadi. Kejahatan yang termasuk dalam *cyber crime* ini mencakup semua jenis kejahatan beserta modus operandinya yang dilakukan sebagai dampak negatif aplikasi internet. Dan terkait dengan transaksi *e-commerce*, kejahatan yang paling sering terjadi adalah penipuan kartu kredit atau biasa diistilahkan dengan *credit card fraud* atau *carding*[1]. Penipuan kartu kredit terus meningkat

*Korenpondensi.
Alamat E-mail : jurnal.larik@bsi.ac.id.
Diterima 30 Nopember 2022; Direvisi 13 Desember 2022; Diterima 22 Desember 2022

dikarenakan *e-commerce* yang lebih dikenal secara umum di kalangan masyarakat. Saat ini belum ada teknik deteksi penipuan yang akurat untuk menangani kasus *carding* sehingga dapat mengakibatkan kerugian keuangan yang signifikan baik bagi pedagang maupun penerbit kartu. Strategi yang digunakan pedagang untuk mengadopsi teknik deteksi penipuan (pencegahan) yang lebih akurat yaitu verifikasi sekunder. Penipuan kartu kredit menyebabkan kerugian keuangan yang signifikan bagi pedagang dan perusahaan jasa keuangan (bank) yang menerbitkan kartu kredit.

Kerugian penipuan kartu di seluruh dunia meningkat dari $ 7,6 miliar pada tahun 2010 menjadi $ 21,81 miliar pada tahun 2015, atau 300% selama 5 tahun. Pada tahun 2020, kerugian penipuan kartu global diperkirakan akan mencapai $ 31,67 miliar [2]. Terdapat dua jenis utama penipuan kartu kredit: penipuan kartu hadir (CP), seperti kartu palsu yang digunakan di titik penjualan atau mesin teller otomatis, dan penipuan kartu tidak hadir (CNP), yang terjadi ketika transaksi dibuat online atau melalui surat, telepon, atau aplikasi seluler.

Dalam kasus-kasus seperti deteksi penipuan kartu kredit dimana hanya ada 1.000 kasus penipuan dalam lebih dari satu juta transaksi, mewakili 0,1% sedikit dari dataset. Hal yang umum sekarang ini adalah persoalan tingkat ketidakseimbangan kelas. Umumnya, hanya sebagian kecil dari jumlah total transaksi adalah penipuan yang sebenarnya. *Dataset* yang tidak seimbang adalah satu dimana jumlah pengamatan yang termasuk dalam satu kelompok atau kelas secara signifikan lebih tinggi daripada yang termasuk kelas-kelas lain. Ambil penipuan kartu kredit misalnya. Dari 1000 transaksi dari pengguna yang diberikan, hanya 1 dari mereka adalah penipuan yang sebenarnya. Ini terjadi karena informasi kartu kredit klien dicuri atau perangkat Vendor PoS disusupi. Pada saat yang sama, kita perlu menyadari kesalahan positif. Biasanya, pemilik kartu kredit tidak akan senang jika kartu kredit diblokir oleh bank ketika tidak ada penipuan yang sebenarnya terjadi. Masalah ketidakseimbangan kelas dimulai sebagai hasil dari *dataset* kehidupan nyata dengan distribusi data yang tidak merata, biasanya dikenal sebagai data yang tidak seimbang. Kelas yang kurang terwakili disebut sebagai kelas minoritas [3]. Ketidakseimbangan kelas ada dalam dataset dimana jumlah kelas tidak merata. Hal ini banyak ditemui dalam berbagai situasi dunia nyata seperti diagnosis medis, analisis data *micro array* atau evaluasi kualitas perangkat lunak.

Klasifikasi data yang tidak seimbang merupakan masalah yang krusial pada bidang *machine learning* dan *data mining*. Ketidakseimbangan data memberikan dampak yang buruk pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas. Pada penelitian ini, *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan untuk menyelesaikan masalah ketidak seimbangan kelas pada *dataset Credit Card Fraud* [4]. Dengan menerapkan skema evaluasi *10-cross fold validation*. Masalah ketidakseimbangan kelas mengacu pada fakta bahwa kinerja algoritma pembelajaran dapat sangat terhambat. Secara umum, kelas mayoritas dinotasikan sebagai negatif, sedangkan kelas minoritas disebut sebagai positif. Teknik klasifikasi standar mungkin tidak berjalan dengan baik dalam hal ini, karena secara internal menganggap distribusi kelas yang sama.

Preprocessing data dengan strategi sampling 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*) [5]. Dari dua strategi *resampling* ini, *undersampling* telah terbukti menjadi pilihan yang lebih baik daripada *oversampling*. Hal ini dikarenakan *oversampling* dapat meningkatkan kemungkinan *overfitting* data ketika proses pelatihan model [6]. Namun, strategi *undersampling* dapat menyebabkan beberapa data yang berguna (*useful data*) untuk proses pembelajaran model akan ikut tereliminasi [7].

Pengambilan sampel data dapat berupa *undersampling* (menghapus *instance* kelas utama) atau *Oversampling* (menambahkan *instance* kelas minoritas duplikat). Penghapusan acak dari *instance* kelas utama meningkatkan kerugian informasi yang berguna serta berdampak buruk bagi proses klasifikasi [8]. Pendekatan level data ditunjukkan untuk memperbaiki kualitas data dengan cara menyeimbangkan kelas menggunakan teknik *resampling* seperti *undersampling* dan *oversampling*, mensintesis data dengan *Synthetic minority Over-Sampling Technique* (SMOTE), atau memilih fitur yang tepat dengan teknik seleksi fitur (*feature selection*). Pengaruh penggunaan data tidak seimbang untuk membuat model sangat besar pada

hasil model yang diperoleh. Pengolahan algoritma yang tidak menghiraukan ketidakseimbangan data akan cenderung diliputi oleh kelas mayor dan mengacuhkan kelas minor. Metode *Oversampling* berprinsip memperbanyak pengamatan secara acak, sedangkan metode SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan [9]. Penyeimbangan kelas dapat juga dengan melakukan sampling pada kelas minoritas (*Oversampling*). *Oversampling* merupakan metode penyeimbangan distribusi kelas dengan mereplikasi *instance* pada kelas minoritas secara acak. Namun, *Oversampling* meningkatkan kemungkinan munculnya *overfitting* karena menduplikasi *instance* secara sama persis. SMOTE ini digunakan untuk pendekatan data bertipe numerik.

Mengatasi permasalahan ketidakseimbangan kelas tersebut, maka penelitian ini akan dilakukan pengukuran kinerja pada pendekatan level data dengan dilakukan teknik SMOTE dan dikombinasikan dengan Tomek-Link. Algoritma pengklasifikasi yang digunakan adalah C5.0. Model C5.0 merupakan model klasifikasi yang cocok untuk data dengan attribut yang bersifat numerik ataupun attribut yang bernilai nominal yaitu bersifat kategorik dimana tiap nilai tidak bisa dijumlahkan atau dikurangkan, sedangkan pengukuran performa model menggunakan *confusion matrix* agar didapat nilai akurasi secara umum, *sensitivity*, *specificity*, akurasi prediksi kelas minoritas, dan *Area Under the ROC Curve* (AUC). Metode SMOTE+Tomek-Link yang diharapkan lebih baik daripada metode lain. Metode kombinasi antara SMOTE dan Tomek Link sebagai metode pembersihan data. Cara kerja Tomek Link adalah dengan menghapus data minor ataupun mayor yang memiliki kesamaan karakteristik. Untuk setiap data, jika satu tetangga yang paling dekat memiliki kelas label yang berbeda dengan data tersebut maka kedua data akan dihapus karena dianggap sebagai *noise* atau *misclassify*.

Berdasarkan uraian yang telah dijabarkan dapat diidentifikasi permasalahan (*research problem*). *Credit Card Fraud Risk* merupakan *dataset* yang mengalami ketidakseimbangan kelas. Pendekatan level data seperti *oversampling* dapat mengangani ketidakseimbangan kelas. Namun cara ini memiliki kelemahan yaitu *overfitting* pada data sehingga menyebabkan performa klasifikasi menjadi berkurang.

II. METODE PENELITIAN

A. Desain Penelitian

Penelitian ini menggunakan metode penelitian eksperimen dengan menggunakan Rstudio dalam pengujian hasil eksperimen. Penelitian eksperimen mencakup investigasi hubungan sebab-akibat menggunakan pengujian yang dikontrol sendiri [10]. Langkah-langkah yang dilakukan pada proses penelitian adalah sebagai berikut:

1. Pengumpulan data
   Pada tahap ini ditentukan data yang akan diproses yaitu, mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam *dataset*, termasuk variabel yang diperlukan dalam proses
2. Pengolahan awal data
   Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.
3. Model yang diusulkan
   Pada tahap ini data dianalisis, kemudian dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.
4. Eksperimen dan pengujian model
   Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan.
5. Evaluasi dan validasi hasil
   Pada bagian ini dilakukan evaluasi dan validasi terhadap model penelitian yang dilakukan untuk mengetahui tingkat keakurasian model.

B. Pengumpulan Data

Pada penelitian ini digunakan pengumpulan data sekunder. Data sekunder adalah data yang diperoleh tidak secara langsung dari objek penelitian, namun berasal dari data yang telah dikumpulkan sebelumnya oleh pihak lain baik di terbitkan atau tidak [11]. Data yang digunakan dalam penelitian ini adalah data yang terkait dengan masalah ketidakseimbangan kelas. *Dataset* yang dipilih berasal dari *Kaggle Dataset Repository*.

Tabel 1. Informasi dataset yang digunakan

Datasets	*Data Sample*	*Feature*	*Ir*
Credit Card Fraud Risk	284,807	31	577,87

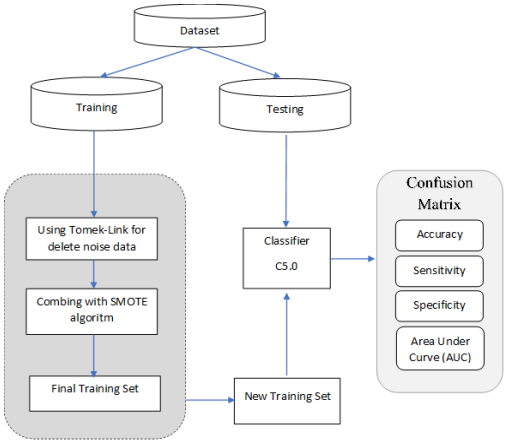
C. Pengolahan Data Awal

Dataset yang digunakan pada penelitian ini perlu dilakukan beberapa teknik pengolahan awal data hal ini bertujuan untuk mendapatkan data yang berkualitas baik. Beberapa teknik data *preparation* diantaranya [12]:

1. Data *validation*, untuk mengatasi masalah Incompleteness atau data yang mengalami missing value.
2. Data *transformation*, untuk meningkatkan akurasi dan efisiensi algoritma beberapa data yang digunakan dalam penelitian ini ada yang bernilai kategorikal.

D. Metode yang Diusulkan

Pada penelitian ini model yang diusulkan adalah kombinasi pendekatan level data Tomek-Link dan SMOTE dengan algoritma klasifikasi C5.0. Tahap pertama adalah membagi *dataset* menjadi *data training* dan *data testing*. Lalu *data training* akan diolah dengan model yang diusulkan yaitu pendekatan level data Tomek-Link yang berfungsi untuk menghilangkan *noise data*. Selanjutnya *data training* yang telah bersih dari *noise* akan diolah kembali dengan pendekatan level data SMOTE dan terbentuklah *dataset training* baru yang seimbang. Metode pengklasifikasian pada penelitian ini menggunakan algoritma klasifikasi C5,0. sedangkan untuk evaluasi kinerja dari menggunakan pengukuran *confusion matrix* dan nilai AUC. Gambar 1 merupakan desain dari algoritma yang diusukan yang dapat dilihat pada gambar berikut ini.



Gambar 1. Kombinasi level data Tomek-Link dan SMOTE dengan klasifikasi C5.0

E. Confusion Metric

Confusion matrix berisi informasi aktual (*actual*) dan prediksi (*predicted*) pada model klasifikasi. *Confusion matrix* memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah [11]. *Confusion matrix* merupakan matrik 2 dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kenyataan, ditunjukkan pada tabel dibawah ini.

Penelitian ini fokus pada mendeteksi kelas minoritas atau kelas positif sehingga *sensitivity* dan *specificity* dapat digunakan untuk menunjukkan performa/kinerja dari dua kelas. *Cutoff* dari sensitivitas dan spesifisitas bisa digunakan untuk membuat kurva ROC [13]. Untuk dua masalah kelas, sensitivitas, spesifisitas, nilai prediksi positif dan nilai prediktif negatif dihitung dengan menggunakan argumen positif. Pada penelitian ini *function confusion matrix* pada *package caret* digunakan untuk menentukan nilai performa akurasi model. Berikut adalah persamaan model *confusion matrix* yang diusulkan pada *function Confusion Matrix* di R.

Tabel 1. Model Confusion Matrix

Predicted	***Reference***	
	Event	***No Event***
Event	A	B
No Event	C	D

Rumus yang digunakan adalah:

Akurasi $= \frac{A+D}{A+B+C+D}$ (1)

Sensitivity $= \frac{A}{A+C}$ (2)

Specificity $= \frac{D}{D+B}$ (3)

Prevalence $= \frac{A+C}{A+B+C+D}$ (4)

PPV = $\frac{Sensitivity * Prevalence}{((Sensitivity * Prevalence)+((1-specificity)*(1-prevalence)))}$ (5)

NPV = $\frac{Specificity*(1-Prevalence)}{(((1-Sensitivity)*Prevalence)+((specificity)*(1-prevalence)))}$ (6)

Detection Rate $= \frac{A}{A+B+C+D}$ (7)

Detection Prevalence $= \frac{A+B}{A+B+C+D}$ (8)

Balanced Accuracy = AUC $= \frac{sensitivity+specificity}{2}$ (9)

Precision $= \frac{A}{A+B}$ (10)

Recall $= \frac{A}{A+C}$ (11)

III. HASIL DAN PEMBAHASAN

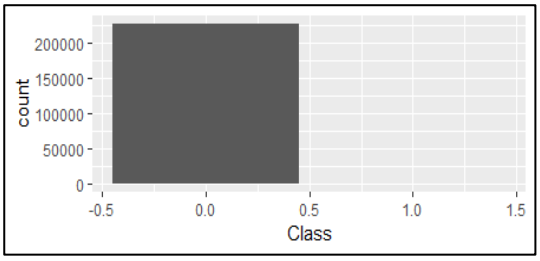
Pengujian model T-Link dan SMOTE dilakukan menggunakan *Kaggle dataset* (*credit card fraud risk*). Skenario pengujian untuk mengatasi ketidakseimbangan kelas dilakukan dengan cara menambah jumlah kelas minoritas (*oversampling*). Eksperimen yang dilakukan pada *Kaggle dataset* menggunakan persentase *imbalance* kisaran antara 51% sampai 52% untuk minoritas kelas dan 48% sampai 49% untuk mayoritas kelas.

A. Eksperimen C50

Dataset credit card fraud risk terdiri dari 31 variabel (*features*) dan 284.807 data sampel. *Dataset* akan dibagi menjadi 2, *data training* sebanyak 80% dan *data testing* 20%. *Data training* akan diolah menggunakan klasifikasi C5.0. Persentase ketidakseimbangan kelas *dataset* adalah 99,83% *negative class* dan 0,17% *positive class*.

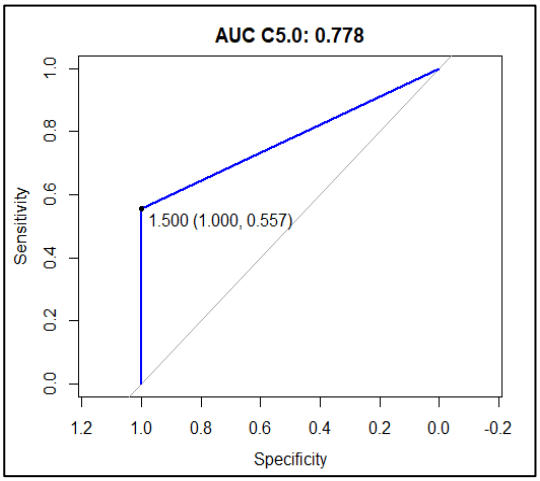
```
> prop.table(table(datatrain$Class))*100

        0         1
99.8266373  0.1733627
```

Gambar 2. *Training Set Credit Card Fraud Risk* setelah C5.0



Gambar 3. Grafik *Training Set Credit Card Fraud Risk* setelah C5.0

Setelah melakukan tahap level data SMOTE dan klasifikasi menggunakan C5.0 menghasilkan AUC sebesar 0,778.



Gambar 4. AUC *dataset Credit Fraud Risk* setelah C5.0

Berdasarkan *training set* diatas, model prediksi dibentuk kemudian diukur performa model menggunakan *confusion matrix*. Berikut adalah hasil perhitungan manual nilai evaluasi dari kinerja model. Selanjutnya akan dibandingkan hasilnya dengan perhitungan menggunakan *function confusion Matrix* pada R tools.

Tabel 2. *Confusion Matrix dataset Credit Fraund Risk* setelah C5.0

Class	*Actual*	
	Negative	*Positive*

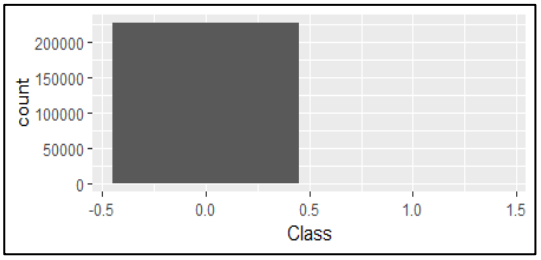
Prediction	*Negative*	56858	6
	Positive	43	54

B. Eksperimen SMOTE dan C5.0

Dengan *dataset* yang sama tahap ini juga akan melakukan pembagian data menjadi 2, *data training* sebanyak 80% dan *data testing* 20%. *Data training* akan diolah menggunakan level data SMOTE terlabih dahulu, tahap selanjutnya menggunakan klasifikasi C5.0. Persentase ketidakseimbangan kelas *dataset* adalah 99,83% *negative class* dan 0,17% *positive class*.

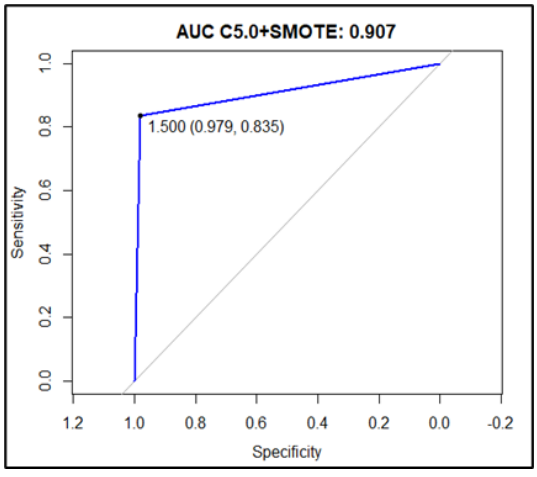
```
> prop.table(table(datatrain$Class))*100

        0         1
99.8266373  0.1733627
```

Gambar 5. *Training Set Credit Card Fraud Risk* setelah SMOTE dan C5.0



Gambar 6. Grafik *Training Set Credit Card Fraud Risk* setelah SMOTE dan C5.0

Setelah melakukan tahap klasifikasi menggunakan C5.0 menghasilkan AUC sebesar 0,907.



Gambar 7. AUC dataset Credit Fraud Risk setelah SMOTE dan C5.0

Berdasarkan *training set* diatas, model prediksi dibentuk kemudian diukur performa model menggunakan *confusion matrix*. Berikut adalah hasil perhitungan manual nilai evaluasi dari kinerja model. Selanjutnya akan dibandingkan hasilnya dengan perhitungan menggunakan *function confusion Matrix* pada R tool.

Tabel 3. *Confusion Matrix dataset Credit Fraund Risk* setelah SMOTE dan C5.0

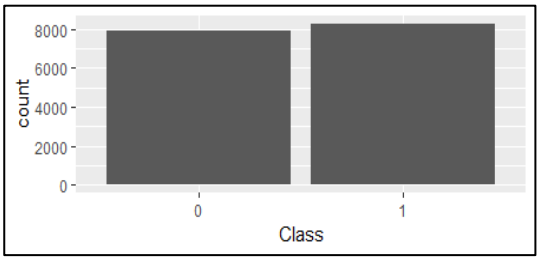
Class		Actual	
		Negative	Positive
Prediction	Negative	55698	1166
	Positive	16	81

C. Eksperimen Tomek-Link dan SMOTE serta C5.0

Dengan *dataset* yang sama tahap ini juga akan melakukan pembagian data menjadi 2, *data training* sebanyak 80% dan *data testing* 20%. *Data training* akan diolah menggunakan level data Tomek-Link dan SMOTE terlabih dahulu, tahap selanjutnya menggunakan klasifikasi C5.0. Persentase ketidakseimbangan kelas *dataset* adalah 48,78% *negative class* dan 51,22% *positive class*.

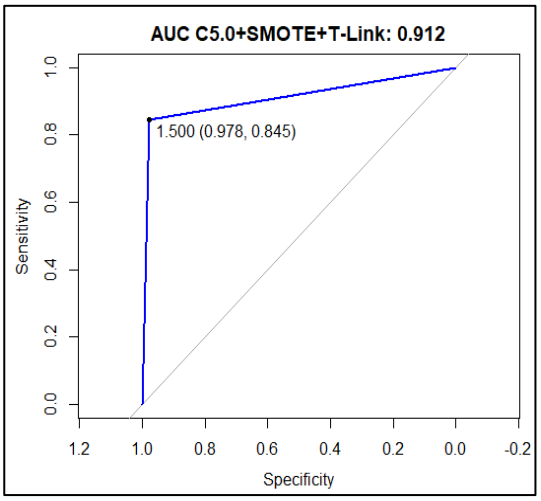
```
> prop.table(table(train_smote_tomek$Class))*100

       0        1
48.78049 51.21951
```

Gambar 8. *Training Set Credit Card Fraud Risk* setelah Tomek-Link dan SMOTE serta C5.0



Gambar 9. Grafik *Training Set Credit Card Fraud Risk* setelah Tomek-Link dan SMOTE serta C5.0

Setelah melakukan tahap klasifikasi menggunakan C5.0 menghasilkan AUC sebesar 0,912.



Gambar 10. AUC *dataset Credit Fraud Risk* setelah Tomek-Link dan SMOTE serta C5.0

Berdasarkan *training set* diatas, model prediksi dibentuk kemudian diukur performa model menggunakan *confusion matrix*. Berikut adalah hasil perhitungan manual nilai evaluasi dari kinerja model. Selanjutnya akan dibandingkan hasilnya dengan perhitungan menggunakan *function confusion Matrix* pada R tools.

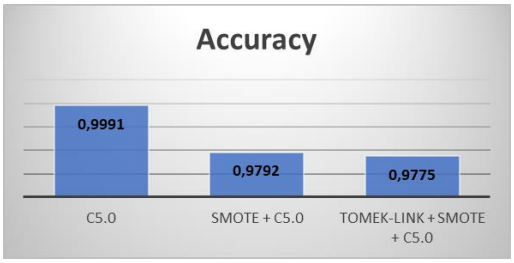
Tabel 4. Confusion Matrix dataset Credit Fraund Risk setelah Tomek-Link dan SMOTE serta C5.0

Class		Actual	
		Negative	Positive
Prediction	Negative	55599	1265
	Positive	15	82

D. Rekap Hasil Pengujian

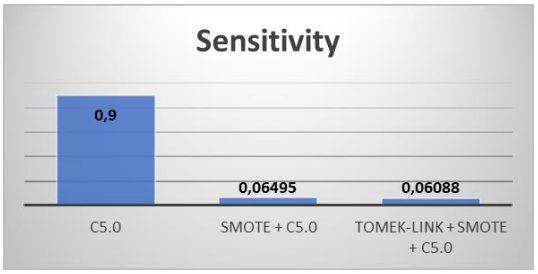
Pada bagian ini akan ditampilkan rekap hasil pengukuran kinerja model pada masing-masing dataset menggunakan *confusion matrix*. Rekap pengujian ini akan menampilkan hasil kinerja model pada *dataset*. Nilai yang diukur adalah *True positif* (TP), *False Positif* (FP), *False Negatif* (FN), *True Negatif* (TN), *Accuracy*, *Sensitivity*, *Specificity* dan *Area Under the ROC Curve* (AUC).

Pada Grafik Gambar 11 menunjukkan hasil pengukuran nilai *Accuracy* dengan metode level data usulan Tomek-Link dan SMOTE serta klasifikasi C5.0 terlihat menurun dibandingkan model SMOTE dan C5.0 atau saat menggunakan model klasifikasi C5.0 saja. Berdasarkan dari perhitungan *Accuracy* menghasilkan nilai klasifikasi menggunakan metode C5.0 sebesar 0,9991, saat menggunakan metode pendekatan level data SMOTE dan klasifikasi C5.0 memperoleh nilai 0,9792 sedangkan menggunakan metode pendekatan level data Tomek-Link dan SMOTE dan klasifikasi C5.0 memperoleh nilai 0,9775.



Gambar 11. Grafik Perbandingan Nilai Accuracy

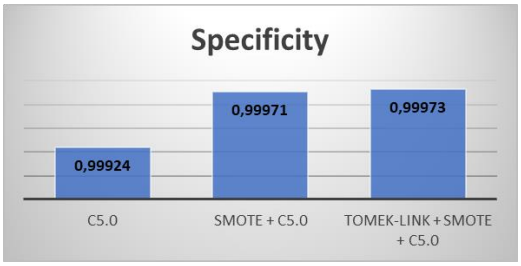
Pada Grafik Gambar 12 menunjukkan hasil pengukuran nilai *Sensitivity* dengan metode level data usulan Tomek-Link dan SMOTE serta klasifikasi C5.0 terlihat menurun dibandingkan saat menggunakan metode klasifikasi C5.0 saja. Berdasarkan dari perhitungan *Sensitivity* menghasilkan nilai klasifikasi menggunakan metode C5.0 sebesar 0,9, menggunakan metode pendekatan level data SMOTE dan klasifikasi C5.0 memperoleh nilai 0,06495 sedangkan menggunakan metode pendekatan level data Tomek-Link dan SMOTE dan klasifikasi C5.0 memperoleh nilai 0,06088.



Gambar 12. Grafik Perbandingan Nilai Sensitivity

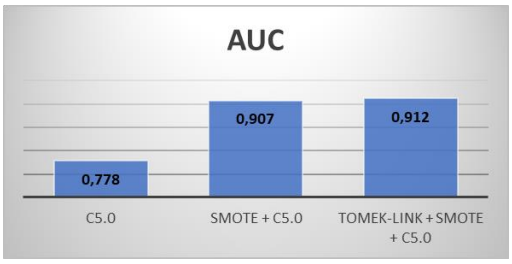
Pada Grafik Gambar 13 menunjukkan hasil pengukuran nilai *Specificity* dengan metode level data usulan Tomek-Link dan SMOTE serta klasifikasi C5.0 terlihat mengalami kenaikan yang cukup signifikan. Selanjutnya model SMOTE dan C5.0 juga mengalami kenaikan yang tidak jauh beda dengan metode usulan. Berdasarkan dari perhitungan *specificity* menghasilkan nilai klasifikasi menggunakan metode C5.0 sebesar

0,99924, menggunakan metode pendekatan level data SMOTE dan klasifikasi C5.0 memperoleh nilai 0,99971 sedangkan menggunakan metode pendekatan level data Tomek-Link dan SMOTE dan klasifikasi C5.0 memperoleh nilai 0,99973.



Gambar 13. Grafik Perbandingan Nilai Specificity

Pada Grafik Gambar 14 menunjukkan hasil pengukuran nilai *Specificity* dengan metode level data usulan Tomek-Link dan SMOTE serta klasifikasi C5.0 terlihat mengalami kenaikan yang cukup signifikan. Berdasarkan dari perhitungan AUC menghasilkan nilai klasifikasi menggunakan metode C5.0 sebesar 0,778, menggunakan metode pendekatan level data SMOTE dan klasifikasi C5.0 memperoleh nilai 0,907 sedangkan menggunakan metode pendekatan level data Tomek-Link dan SMOTE dan klasifikasi C5.0 memperoleh nilai 0,912.



Gambar 14. Grafik Perbandingan Nilai AUC

IV. KESIMPULAN DAN SARAN

Pada penelitian ini mencoba teknik *oversampling* menggunakan pendekatan level data Tomek-Link dan SMOTE serta algoritma klasifikasi C5.0. Tujuan dari penitian ini adalah mengembangkan kombinasi pendekatan level data untuk mengatasi ketidakseimbangan kelas sehingga dapat meningkatkan performa model klasifikasi C5.0. Eksperimen dilakukan menggunakan *dataset* tidak seimbang yang diambil dari *Kaggle dataset*.

Dari hasil penelitian yang sudah dilakukan untuk menjawab rumusan masalah (*research quetions*) dapat dilihat dari Gambar 13 dan Gambar 14. Hasil tersebut menunjukkan peningkatan yang signifikan pada nilai *specificity* dan nilai AUC pada model usulan dibandingkan model tanpa *resampling* atau saat menggunakan SMOTE saja. Ini menunjukkan teknik kombinasi pendekatan level data berupa Tomek-Link dan SMOTE serta algoritma klasifikasi C5.0 dapat menangani ketidakseimbangan kelas pada *dataset Credit Card Fraud Risk* dari *Kaggle-Repository*.

V. REFERENSI

[1] S. N. Prasetyo, "Rumusan Pengaturan Credit Card Fraud Dalam Hukum Pidana Indonesia Ditinjau Dari Asas Legalitas," *J. Ilm. Huk. Leg.*, vol. 24, no. 1, p. 101, 2017, doi: 10.22219/jihl.v24i1.4260.

[2] P. K. Robertson, "Cone penetration test (CPT)-based soil behaviour type (SBT) classification system — An update," *Can. Geotech. J.*, vol. 53, no. 12, pp. 1910–1927, 2016, doi: 10.1139/cgj-2016-0044.

[3] S. Vluymans, D. S. Tarrago, Y. Saeys, C. Cornelis, and F. Herrera, "Fuzzy multi-instance classifiers," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 6, pp. 1395–1409, 2016, doi: 10.1109/TFUZZ.2016.2516582.

[4] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[5] A. R. Ismail, N. Z. Abidin, and M. K. Maen, "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare," *J. Robot. Control*, vol. 3, no. 2, pp. 143–152, 2022, doi: 10.18196/jrc.v3i2.13133.

[6] 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.

[7] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015, doi: 10.1016/j.patcog.2014.11.014.

[8] E. Irawan and R. S. Wahono, "Penggunaan Random Under Sampling untuk Penanganan

Ketidakseimbangan Kelas pada Prediksi Cacat Software Berbasis Neural Network," *J. Softw. Eng.*, vol. 1, no. 2, pp. 92–100, 2015.

[9] R. Azmatul Barro, I. D. Sulvianti, and M. Afendi, "Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu," *Xplore J. Stat.*, vol. 1, no. 1, pp. 1–6, 2013.

[10] A. Asrin, "Metode Penelitian Eksperimen," *J. Maqasiduna Ilmu Humaniora, Pendidik. Ilmu Sos.*, vol. 2, no. 1, pp. 1–9, 2022, [Online]. Available: https://journal.mukhlisina.id/index.php/maqasiduna/article/view/24/15

[11] A. Nurwanda and E. Badriah, "Analisis Program Inovasi Desa Dalam Mendorong Pengembangan Ekonomi Lokal Oleh Tim Pelaksana Inovasi Desa (PID) Di Desa Bangunharja Kabupaten Ciamis," *J. Ilm. Ilmu Adm. Negara*, vol. 7, no. 1, pp. 68–75, 2020, [Online]. Available: https://jurnal.unigal.ac.id/index.php/dinamika/article/download/3313/pdf

[12] U. Enri, "PENERAPAN ALGORITMA C4.5 DALAM PEMILIHAN PROGRAM STUDI FAKULTAS ILMU KOMPUTER (Studi Kasus Sekolah Menengah Atas Negeri 1 Tambun Utara)," *J. Rekayasa Inf.*, vol. 7, no. 1, pp. 1–7, 2018.

[13] M. Kuhn and K. Johnson, *Applied Predictive Modeling [Hardcover]*. 2013. doi: 10.1007/978-1-4614-6849-3.

[14] M. Kuhn, "caret Package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, 2008.