Analisis Pengaruh Data *Preprocessing* terhadap *Imbalance*d dan *Outlier* Dataset pada Kasus Klasifikasi *Fraud Credit Card*

Syahmi Sajid[1]

Magister Kecerdasan Artifisial, Universitas Gadjah Mada
Bulaksumur, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281
[1]syahmisajid12@gmail.com

Abstrak— **Penelitian ini mengkaji pengaruh berbagai teknik *Preprocessing* data terhadap dataset yang tidak seimbang dalam kasus klasifikasi *fraud* kartu kredit. Penelitian ini penting dalam konteks industri keuangan, karena ketidakseimbangan antara kasus *fraud* dan non-*fraud* dapat mengarah pada bias dalam model klasifikasi. Studi ini mengevaluasi teknik *Preprocessing* seperti *Handling missing data, Data balancing, Outlier Handling,* dan kombinasi keduanya terhadap dataset transaksi kartu kredit yang mencakup fitur-fitur yang relevan untuk deteksi *fraud*. Hasil eksperimen menunjukkan bahwa penggunaan teknik *Data balancing* dan *Outlier Handling* secara signifikan meningkatkan perfoma model, memungkinkan deteksi *fraud* yang lebih baik. Namun, penelitian ini juga menekankan pentingnya memilih metrik evaluasi yang tepat sesuai dengan tujuan bisnis, karena hasil dapat bervariasi tergantung pada metrik yang digunakan. Temuan ini memberikan panduan berharga bagi praktisi dalam industri keuangan untuk meningkatkan efektivitas deteksi *fraud* kartu kredit dengan memanfaatkan teknik *Preprocessing* data yang sesuai. Hal ini dapat membantu melindungi pelanggan dan aset perusahaan dari risiko *fraud* yang potensial.**

Kata kunci*— Data *Preprocessing*, *Imbalance* Dataset, *Outlier Handling*, Logistic Regression, *Fraud Credit Card

I. PENDAHULUAN

Kasus penipuan kartu kredit merupakan salah satu tantangan serius dalam industri keuangan yang terus berlanjut hingga saat ini. Deteksi penipuan umumnya dipandang sebagai masalah klasifikasi data mining, dimana tujuannya adalah untuk mengklasifikasikan dengan benar transaksi kartu kredit sebagai sah atau curang [5]. Kejahatan ini tidak hanya merugikan pemegang kartu kredit dan penyedia layanan keuangan, tetapi juga mengganggu integritas sistem keuangan secara keseluruhan. Oleh karena itu, pengembangan metode yang efektif untuk mendeteksi penipuan kartu kredit sangat penting. Namun, pengklasifikasian transaksi kartu kredit sebagai sah atau penipuan seringkali melibatkan masalah klasifikasi data yang tidak seimbang (*imbalanced dataset*). Banyak teknik telah diterapkan untuk mendeteksi penipuan kartu kredit, jaringan saraf tiruan [1], algoritma genetika [2, 3], dukungan mesin vektor [4], penambangan item yang sering [5], pohon keputusan [6], algoritma optimasi burung yang bermigrasi [7], naïve bayes [8]. Analisis komparatif regresi logistik dan naif bayes dilakukan di [9].

Imbalanced dataset terjadi ketika jumlah sampel dalam kelas minoritas (dalam kasus ini, transaksi penipuan kartu kredit) sangat jauh lebih sedikit daripada jumlah sampel dalam kelas mayoritas (transaksi sah). Masalah ini dapat mengarah pada kinerja model yang buruk, di mana model cenderung cenderung memprediksi mayoritas kelas dan mengabaikan kelas minoritas yang sebenarnya lebih penting dalam konteks deteksi penipuan. Oleh karena itu, penanganan dataset yang tidak seimbang menjadi salah satu aspek kunci dalam mengatasi masalah ini.

Data *Preprocessing* adalah tahap kritis dalam pengembangan model klasifikasi, dan dalam kasus ini, penanganan data yang tidak seimbang menjadi fokus utama. Dalam penelitian ini, kami akan mengkaji pengaruh data *Preprocessing* terhadap kinerja model klasifikasi pada kasus deteksi penipuan kartu kredit. Data *Preprocessing* mencakup tiga aspek utama: penanganan data yang hilang (*handling missing data*), penyeimbangan dataset (*data balancing*), dan penanganan *Outlier* (*Outlier Handling*). Ketiga aspek ini akan dianalisis secara terperinci untuk memahami bagaimana mereka memengaruhi hasil klasifikasi pada dataset yang tidak seimbang.

A. Penanganan Data yang Hilang (*Handling missing data*)

Penanganan data yang hilang adalah langkah awal dalam *Preprocessing* data yang kritis. Dalam kasus dataset penipuan kartu kredit, data yang hilang bisa muncul karena berbagai alasan, seperti kesalahan teknis, gangguan sistem, atau bahkan usaha penipuan yang merusak integritas data. Penting untuk memahami bagaimana menangani data yang hilang, apakah dengan menghapusnya, mengisi nilainya, atau menggunakan teknik lain yang lebih kompleks. Penghapusan data yang hilang secara kasar dapat mengurangi jumlah sampel dalam dataset, sementara pengisian data yang hilang dengan nilai yang tidak tepat dapat mempengaruhi hasil klasifikasi. Oleh karena itu, dalam penelitian ini, kami akan menganalisis berbagai metode penanganan data yang hilang dan dampaknya pada klasifikasi penipuan kartu kredit.

B. Penyeimbangan Dataset (*Data balancing*)

Imbalanced dataset adalah masalah kritis dalam deteksi penipuan kartu kredit. Karena transaksi penipuan relatif jarang terjadi dibandingkan dengan transaksi sah, model klasifikasi cenderung menjadi bias terhadap mayoritas kelas. Oleh karena itu, perlu ada upaya untuk mengimbangi dataset agar model

dapat lebih akurat mendeteksi transaksi penipuan. Kami akan menganalisis berbagai teknik penyeimbangan dataset, seperti oversampling (penambahan sampel pada kelas minoritas), undersampling (pengurangan sampel pada kelas mayoritas), dan metode lainnya. Kami akan membandingkan kinerja model klasifikasi setelah menerapkan berbagai teknik ini untuk memahami pengaruhnya pada deteksi penipuan kartu kredit.

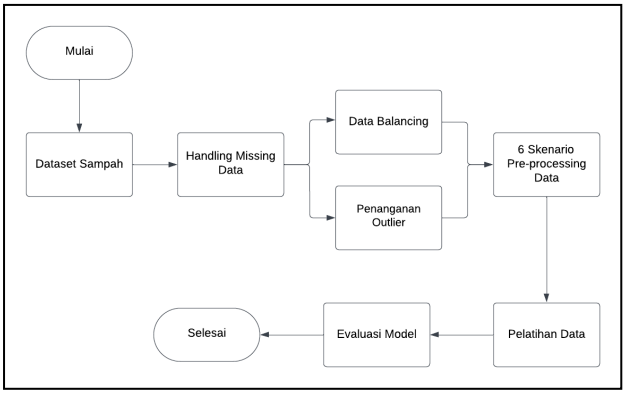
C. Penanganan *Outlier* (*Outlier Handling*)

Outlier adalah nilai yang jauh dari pola data yang umumnya. Dalam konteks deteksi penipuan kartu kredit, *Outlier* mungkin mengindikasikan aktivitas yang mencurigakan atau penipuan yang sebenarnya. Namun, *Outlier* juga dapat menjadi gangguan jika tidak ditangani dengan benar. Dalam penelitian ini, kami akan mengevaluasi berbagai metode penanganan *Outlier* dan dampaknya pada kinerja model klasifikasi. Kami akan mencari tahu apakah penanganan *Outlier* yang tepat dapat meningkatkan kemampuan model untuk mendeteksi transaksi penipuan dengan lebih baik.

Penelitian ini memiliki tujuan ganda. Pertama, kami akan mengidentifikasi metode *Preprocessing* data yang paling efektif dalam meningkatkan kinerja model klasifikasi pada dataset penipuan kartu kredit yang tidak seimbang. Kedua, kami akan memberikan panduan praktis kepada pemegang kartu kredit dan penyedia layanan keuangan tentang bagaimana mengoptimalkan deteksi penipuan kartu kredit dengan mempertimbangkan teknik *Preprocessing* data yang sesuai. Dengan memahami pentingnya penanganan data yang hilang, penyeimbangan dataset, dan penanganan *Outlier* dalam konteks ini, diharapkan dapat menghasilkan model klasifikasi yang lebih handal dalam mendeteksi penipuan kartu kredit, yang pada akhirnya akan mengurangi kerugian finansial dan menjaga keamanan transaksi keuangan.

II. METODOLOGI

A. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Tahapan penelitian seperti yang ditunjukkan pada Gambar 1 dimulai dari proses pengumpulan dataset *Fraud Credit Card* dengan jumlah 2 kelas yaitu 0 dan 1. Tahap berikutnya yaitu dilakukan *handling missing data* pada data NaN pada dataset. Tahap selanjutnya dilakukan pembagian 2 skenario yaitu teknik untuk mengatasi dataset yang tidak seimbang (*imbalanced dataset*) dan menghapus *Outlier*.

Masing-masing hasil dari proses sebelumnya dilakukan kombinasi percobaan terhadap 6 skenario pada proses pre-prosesing data. Setiap rangkaian kombinasi skenario dilakukan proses pelatihan (*training*) menggunakan model *Machine learning* logistik regresi. Semua hasil model yang terbentuk dilakukan proses pengujian model menggunakan beberapa matriks evaluasi dan semua skenario dibandingkan untuk menentukan skenario terbaik.

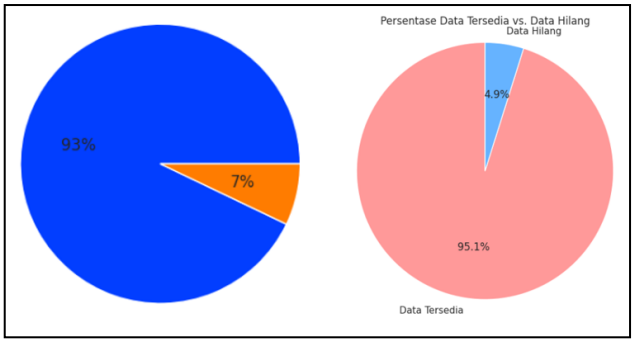
B. Dataset

Tabel 1. Dataset *Fraud Credit Card*

	distance_from_home	distance_from_last_transaction	ratio_to_median_purchase_price	repeat_retailer	used_chip	used_pin_number	online_order	fraud
763547	8.073065	0.614990	0.077806	1.0	0.0	NaN	1.0	0.0
23634	31.994433	5.808574	1.186485	1.0	0.0	1.0	0.0	0.0
157152	8.015637	NaN	0.906554	1.0	0.0	0.0	1.0	0.0
238071	6.230256	0.796154	0.206875	1.0	1.0	0.0	NaN	0.0
846711	44.997072	10.943190	0.856966	1.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...
593623	29.588257	66.519297	0.072054	1.0	0.0	NaN	1.0	1.0
70103	128.707098	0.308811	1.304773	1.0	0.0	NaN	1.0	1.0
668717	6.920051	0.402661	4.681969	1.0	1.0	0.0	1.0	1.0
28379	24.501468	12.169583	4.306104	1.0	0.0	NaN	1.0	1.0
988548	1.226412	0.058816	5.705950	0.0	0.0	0.0	0.0	1.0

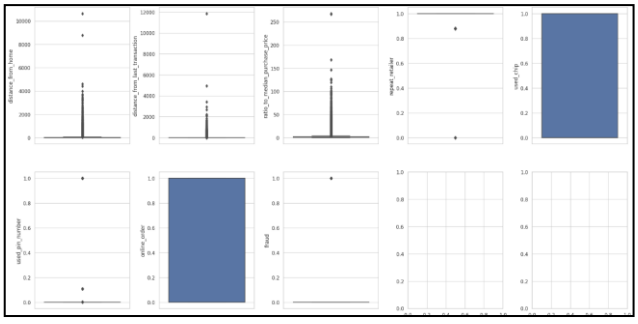
982519 rows × 8 columns

Dataset yang digunakan dapat dilihat dari tabel 1. Dataset pada penelitian ini bersumber dari Kaggle yang berjudul "*Credit Card Fraud Analysis*" dan dipublish oleh Dhanush Narayanan R. Total dataset yang diperoleh yaitu berjumlah 982519 data dengan 7 variabel independen (*distance from home, distance from last transaction, ratio to median purchase price, repeat retailer, used chip, used pin number, dan online order*) dan 1 variabel dependen (*fraud*),



Gambar 2. *Imbalance* dataset dan Missing Data

Dapat dilihat dari gambar 2. Dataset terdiri dari 2 kelas yaitu 0 (bukan *fraud*) berjumlah 912597 dan 1 (*fraud*) berjumlah 69922. Perbandingan kelas 0 dan kelas 1 sekitar 93% : 7% yang berarti dataset tidak seimbang dan missing data sebanyak 5% yang harus dilakukan data *Preprocessing*.



Gambar 3. *Outlier* dalam Dataset

Pada gambar 3 dapat dilihat adanya *Outlier* pada beberapa variable seperti variable *distance from home, distance from last transaction, ratio to median purchase price, repeat retailer dan used pin number*.

C. Skenario

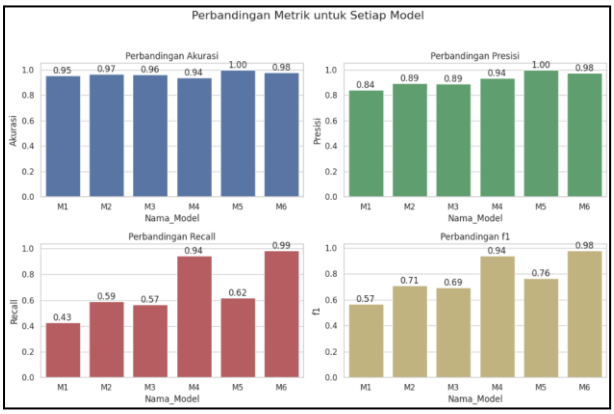
Untuk mendapatkan hasil yang terbaik dalam penelitian ini menggunakan 6 skenario pada proses *Preprocessing*. 6 skenario yang digunakan dalam penelitian ini dapat dilihat pada Tabel 2

Tabel 2. Skenario *Preprocessing*

No	Kode	Skenario
1	M1	*Handling missing data* (NaN = 0)
2	M2	*Handling missing data* (NaN = *Mean* per label)
3	M3	*Handling missing data* (NaN = delete)
4	M4	*Handling missing data* (M2) + *Data balancing*
5	M5	*Handling missing data* (M2) + *Outlier Handling*
6	M6	*Handling missing data* (M2) + *Data balancing + Outlier Handling*

Tabel 2 berisi 6 Skenario yang akan dilakukan pada penelitan ini, masing masing akan di modelkan dan dievaluasi untuk mencari mana skenario terbaik untuk kasus ini dan melihat pengaruh data *Preprocessing* pada performa model *machine learning*.

III. HASIL DAN PEMBAHASAN



Gambar 4. Perbandingan Performa Model

M1, M2, dan M3 adalah *pre-processing* untuk missing value. M2 merupakan model dengan keseluruhan matriks evaluasi tertinggi dibandingkan model dengan *handling missing value* lain dengan nilai akurasi sebesar 0.97, Presisi 0.89, Recall 0.59 dan F1-score 0.71. M1 merupakan model dengan keseluruhan matriks evaluasi terendah dibandingkan model dengan *handling missing value* lain dengan nilai akurasi sebesar 0.95, Presisi 0.84, Recall 0.43 dan F1-score 0.57. Ketiga model ini memiliki recall dan F1-score yang berbeda jauh dengan akurasi dan presisi, hal ini disebabkan data yang tidak seimbang. Proses pertama untuk *handling missing value* didapat model M2 adalah model tertinggi dan akan digunakan pada proses selanjutnya.

M4 adalah lanjutan model M2 dengan menambahkan proses *balancing* data mendapat nilai akurasi sebesar 0.94, presisi 0.94, recall 0.94, dan f1 0.94. Nilai akurasi dari M4 memang berkurang sedikit tetapi nilai presisi, recall dan f1, meningkat secara signifikan yang menandakan performa model tidak lagi bias kepada satu label. M5 adalah lanjutan model M2 dengan menambahkan proses *handling Outlier* mendapat nilai akurasi sebesar 1.00, presisi 1.00, recall 0.62, dan f1 0.76. Akurasi dan presisi nya memang lebih baik dari M4 tetapi recall dan f1 masih rendah yang menandakan model *Machine learning* masih bias ke salah satu label.

M6 adalah lanjutan M2 dengan menambahkan proses *balancing* data dan *handling Outlier*. Model ini dapat dikatakan mempunyai performa terbaik dengan nilai akurasi 0.98 dan presisi 0.98 yang sangat mendekati model M5 tetapi memiliki recall 0.99 dan f1 0.98 yang tertinggi.

IV. KESIMPULAN

M2 adalah model terbaik pada kasus ini dalam *handling missing data* dengan mengganti nilai NaN dengan nilai *Mean* per label dibanding M1, model yang mengganti nilai Nan dengan nilai 0 dan M3, model yang menghapus baris yang berisi nilai Nan. M6 adalah model yang mendapat nilai keseluruhan performa tertinggi dibanding keenam model yang diujikan pada kasus ini dengan cara menggabungkan 3 proses *pre-processing* yaitu *handling missing data* dengan nilai NaN diganti dengan nilai *Mean* per label, *balancing* data, dan *handling Outlier*. Matriks akurasi tidak cocok menjadi matriks evaluasi untuk kasus *imbalance* data berbanding terbalik dengan f1-score.

REFERENSI

[1] [Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology, Vol. 6, No. 3, pp. 311 – 322

[2] [RamaKalyani, K. and UmaDevi, D., (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm, International Journal of Scientific & Engineering Research, Vol. 3, Issue 7, pp. 1 – 6, ISSN 2229-5518

[3] [Meshram, P. L., and Bhanarkar, P., (2012). Credit and ATM Card Fraud Detection Using Genetic Approach, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, pp. 1 – 5, ISSN: 2278-0181

[4] Singh, G., Gupta, R., Rastogi, A., Chandel, M. D. S., and Riyaz, A., (2012). A Machine Learning Approach for Detection of Fraud based on SVM, International Journal of Scientific Engineering and Technology, Volume No.1, Issue No.3, pp. 194-198, ISSN : 2277-1581

[5] Seeja, K. R., and Zareapoor, M., (2014). FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, The Scientific World Journal, Hindawi Publishing Corporation, Volume 2014, Article ID 252797, pp. 1 – 10, http://dx.doi.org/10.1155/2014/252797

[6] Patil, S., Somavanshi, H., Gaikwad, J., Deshmane, A., and Badgujar, R., (2015). Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.4, Issue 4, pp. 92-95, ISSN: 2320-088X

[7] Duman, E., Buyukkaya, A., & Elikucuk, I. (2013). A novel and successful credit card fraud detection system implemented in a turkish bank. In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on (pp. 162-171). IEEE.

[8] Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. In Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 677-685). Society for Industrial and Applied Mathematics.

[9] Ng, A. Y., and Jordan, M. I., (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 2, 841-848.