

**ANALISIS PENGGUNAAN *CORRELATION-BASED FEATURE
SELECTION* (CFS) PADA DETEKSI PENIPUAN *CREDIT CARD*
MENGUNAKAN *RANDOM FOREST***

SKRIPSI



Disusun Oleh :

AHMAD RIZAL FATONI

09020620019

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL
SURABAYA**

2026

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama : Ahmad Rizal Fatoni

NIM : 09020620019

Program Studi : Sistem Informasi

Angkatan : 2020

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul: “ANALISIS PENGGUNAAN *CORRELATION-BASED FEATURE SELECTION* (CFS) PADA DETEKSI PENIPUAN *CREDIT CARD* MENGGUNAKAN *RANDOM FOREST*”. Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan. Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 8 Januari 2026

Yang menyatakan,

(Ahmad Rizal Fatoni)



NIM/09020620019

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh:

NAMA : AHMAD RIZAL FATONI

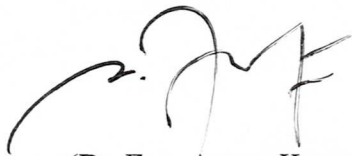
NIM : 09020620019

JUDUL : ANALISIS PENGGUNAAN *CORRELATION-BASED FEATURE SELECTION* (CFS) PADA DETEKSI PENIPUAN *CREDIT CARD* MENGGUNAKAN *RANDOM FOREST*

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 17 Juni 2025

Dosen Pembimbing 1



(Dr. Eng. Anang Kunaefi, M. Kom)

NIP. 197911132014031001

Dosen Pembimbing 2



(Dwi Rolliawati, MT)

NIP.197909272014032001

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Ahmad Rizal Fatoni ini telah dipertahankan
di depan tim penguji skripsi
di Surabaya, 8 Januari 2026

Dosen Penguji 1

Khalid, M.Kom

NIP. 197906092014031002

Dosen Penguji 2

Mujib Ridwan, S.Kom., M.T

NIP. 198604272014031004

Dosen Penguji 3

Dr. Eng. Anang Kunaefi, M. Kom

NIP. 197911132014031001

Dosen Penguji 4

Dwi Rolliawati, MT

NIP. 197909272014032001

Mengesahkan,

Dewan penguji

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Uin Sunan Ampel Surabaya



Dr. A. Saepul Hamdani, M.Pd

NIP. 196507312000031002



UIN SUNAN AMPEL
SURABAYA

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN

Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : MOHAMMAD FATHUR ROZAQ
NIM : 05020221058
Fakultas/Jurusan : Syariah dan Hukum / Hukum Ekonomi Syariah
E-mail address : mfathurrozaq99@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :

☒ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)

yang berjudul :

TINJAUAN HUKUM ISLAM TERHADAP PRAKTIK JUAL BELI GELONDONGAN
UDANG DENGAN PERHITUNGAN *SAMPLING* DAN SATUAN *REAN* DI USAHA
PENDEDERAN UDANG VANAME DESA CERME LOR

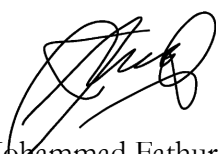
beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Eksklusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara **fulltext** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 2 Februari 2026

Penulis


(Mohammad Fathur Rozaq)

ABSTRAK

Deteksi transaksi *fraud* pada sistem pembayaran elektronik merupakan permasalahan penting karena karakteristik data yang bersifat tidak seimbang (*imbalanced*) dan kompleks. Penelitian ini bertujuan untuk menganalisis kinerja algoritma *Random Forest* dalam mendeteksi transaksi *fraud* dengan membandingkan tiga skenario pengolahan data, yaitu *Random Forest* tanpa *preprocessing* khusus, *Random Forest* dengan *Correlation-Based Feature Selection* (CFS), serta *Random Forest* dengan kombinasi CFS dan *Synthetic Minority Over-sampling Technique* (SMOTE). Dataset yang digunakan merupakan dataset transaksi kartu kredit yang memiliki proporsi kelas *fraud* yang jauh lebih sedikit dibandingkan transaksi normal. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa *Random Forest* tanpa *preprocessing* tambahan menghasilkan performa terbaik dengan nilai *F1-score* dan ROC-AUC yang tinggi. Penerapan CFS tidak secara signifikan meningkatkan kinerja model, sedangkan kombinasi CFS dan SMOTE mampu meningkatkan *recall* kelas *fraud* namun menyebabkan penurunan *precision* secara signifikan. Hal ini mengindikasikan bahwa teknik *oversampling* tidak selalu memberikan peningkatan kinerja pada kasus deteksi *fraud*. Kesimpulan dari penelitian ini adalah bahwa *Random Forest* mampu menangani data *fraud* yang *imbalance* dengan baik tanpa memerlukan teknik *preprocessing* yang kompleks, serta pemilihan metode *preprocessing* harus disesuaikan dengan karakteristik data dan tujuan deteksi.

Kata kunci: deteksi *fraud*, *Random Forest*, CFS, SMOTE, data *imbalance*

UIN SUNAN AMPEL
S U R A B A Y A

ABSTRACT

Fraud detection in electronic payment systems is a critical problem due to the highly imbalanced and complex nature of transaction data. This study aims to analyze the performance of the Random Forest algorithm in detecting fraudulent transactions by comparing three data processing scenarios: Random Forest without specific preprocessing, Random Forest with Correlation-Based Feature Selection (CFS), and Random Forest with a combination of CFS and Synthetic Minority Over-sampling Technique (SMOTE). The dataset used consists of credit card transactions with a significantly smaller proportion of fraudulent transactions compared to normal ones. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The experimental results indicate that Random Forest without additional preprocessing achieved the best overall performance, particularly in terms of F1-score and ROC-AUC. The application of CFS did not significantly improve model performance, while the combination of CFS and SMOTE increased the recall of the fraud class but led to a substantial decrease in precision. These findings suggest that oversampling techniques do not always improve fraud detection performance. In conclusion, Random Forest is effective in handling imbalanced fraud datasets without requiring complex preprocessing techniques, and the selection of preprocessing methods should be carefully aligned with data characteristics and detection objectives.

Keywords: fraud detection, Random Forest, CFS, SMOTE, imbalanced data



UIN SUNAN AMPEL
SURABAYA

DAFTAR ISI

SKRIPSI.....	i
LEMBAR PERSETUJUAN PEMBIMBING	i
PENGESAHAN TIM PENGUJI SKRIPSI	ii
PERNYATAAN KEASLIAN.....	iii
MOTTO	iv
HALAMAN PERSEMBAHAN	v
ABSTRAK	vi
ABSTRACT	vii
UCAPAN TERIMA KASIH.....	viii
KATA PENGANTAR.....	ix
DAFTAR ISI.....	x
DAFTAR GAMBAR	xiii
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah	4
1.3. Batasan Penelitian	4
1.4. Tujuan Penelitian.....	4
1.5. Manfaat Penelitian.....	5
BAB II TINJAUAN PUSTAKA.....	6
2.1. Penelitian Terdahulu	6
2.2. <i>Credit Card Fraud Detection</i>	7
2.3. <i>Machine learning – Fraud Detection</i>	8
2.4. <i>Random forest</i>	10
2.5. <i>Correlation-Based Feature Selection (CFS)</i>	11
2.6. <i>Synthetic Minority Oversampling Technique (SMOTE)</i>	12
2.7. <i>Confusion matrix</i>	14
2.8. Integrasi Keilmuan	17
BAB III METODOLOGI PENELITIAN	19
3.1. Jenis Penelitian	19
3.2. Alur Penelitian.....	20
3.2.1. Pengumpulan <i>Dataset</i>	21
3.2.2. Eksplorasi <i>Dataset</i>	21

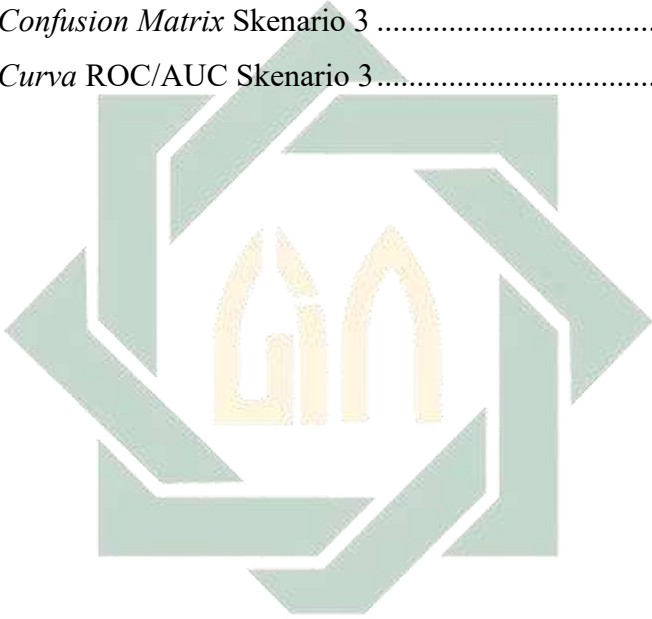
3.2.3.	<i>Preprocessing Dataset</i>	24
3.2.4.	<i>Split Dataset (Data Train dan Test)</i>	26
3.2.5.	Implementasi <i>Random forest</i> dan Skenario Pengujian.....	28
3.2.6.	Evaluasi Hasil.....	30
3.3.	Perangkat Penelitian	31
BAB IV	HASIL DAN PEMBAHASAN	32
4.1	Hasil Eksplorasi Data Analisis.....	32
4.1.1.	Hasil Analisis Kualitas Data.....	33
4.1.2.	Hasil Analisis Distribusi Kelas Target.....	33
4.1.3.	Hasil Analisis Fitur Numerik.....	34
4.1.4.	Hasil Analisis Fitur Kategorikal.....	35
4.1.5.	Hasil Analisis Pola Waktu Transaksi.....	39
4.1.6.	Hasil Analisis Hubungan Fitur Dengan Target.....	42
4.1.7.	Hasil Analisis Korelasi Dan Redundansi Fitur.....	44
4.2	Hasil Preprocessing Data	46
4.2.1.	Penghapusan Fitur Tidak Relevan.....	46
4.2.2.	Transformasi Nilai Transaksi	47
4.2.3.	Rekayasa Fitur (<i>Feature Engineering</i>).....	47
4.2.4.	<i>Encoding</i> Fitur Kategorikal.....	48
4.2.5.	Pembagian Data Latih dan Data Uji.....	49
4.3	Hasil Seleksi Fitur CFS.....	49
4.4	Hasil Pemodelan Skenario	51
4.3.1.	Hasil Skenario 1	51
4.3.2.	Hasil Skenario 2	53
4.3.3.	Hasil Skenario 3	56
4.5	Perbandingan Antar Skenario.....	58
4.6	Pembahasan.....	59
4.6.1.	Pembahasan Hasil Eksplorasi dan <i>Preprocessing Data</i>	59
4.6.2.	Pembahasan Seleksi Fitur Menggunakan CFS.....	60
4.6.3.	Pembahasan Hasil Pemodelan Antar Skenario.....	61
BAB V	PENUTUP	62
5.1.	Kesimpulan.....	62
5.2.	Saran.....	63
DAFTAR PUSTAKA		65

DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu.....	6
Tabel 2. 2 <i>Confusion matrix</i> 2 kelas.....	15
Tabel 3. 1 Implementasi <i>Random forest</i> dan Skenario pengujian	28
Tabel 4. 1 Daftar Fitur Per Kategori.....	32
Tabel 4. 2 hasil cek <i>missing values</i>	33
Tabel 4. 3 Statistik Deskriptif Fitur Numerik Nilai Transaksi	34
Tabel 4. 4 Korelasi Fitur Numerik terhadap <i>is_fraud</i>	35
Tabel 4. 5 Daftar merchant yang paling banyak melakukan transaksi.....	36
Tabel 4. 6 Distribusi Transaksi Berdasarkan Kategori.....	37
Tabel 4. 7 Proporsi <i>Fraud</i> Berdasarkan Kategori Transaksi	38
Tabel 4. 8 Distribusi dan Proporsi <i>Fraud</i> Berdasarkan <i>Gender</i>	38
Tabel 4. 9 Negara Bagian dengan Jumlah Transaksi Terbanyak.....	39
Tabel 4. 10 Distribusi dan Proporsi <i>Fraud</i> Berdasarkan Jam Transaksi	40
Tabel 4. 11 Distribusi dan Proporsi <i>Fraud</i> Berdasarkan Hari dalam Minggu	40
Tabel 4. 12 Distribusi dan Proporsi <i>Fraud</i> Berdasarkan Bulan	41
Tabel 4. 13 Korelasi Fitur Waktu terhadap <i>is_fraud</i>	42
Tabel 4. 14 Korelasi Fitur Numerik terhadap <i>is_fraud</i>	42
Tabel 4. 15 Fitur yang di hapus	46
Tabel 4. 16 Pembagian Data latih dan Data uji	49
Tabel 4. 17 Daftar Fitur Yang Terpilih oleh CFS	49
Tabel 4. 18 Perbandingan Antar Skenario.....	58

DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian.....	21
Gambar 4. 1 Hasil Distribusi kelas Target	34
Gambar 4. 2 <i>Correlation Matrik</i>	44
Gambar 4. 3 <i>Confusion Matrix</i> Skenario 1	52
Gambar 4. 4 <i>Curva</i> ROC/AUC Skenario 1	53
Gambar 4. 5 <i>Confusion Matrix</i> Skenario 2	54
Gambar 4. 6 <i>Curva</i> ROC/AUC Skenario 2.....	55
Gambar 4. 7 <i>Confusion Matrix</i> Skenario 3	56
Gambar 4. 8 <i>Curva</i> ROC/AUC Skenario 3.....	57

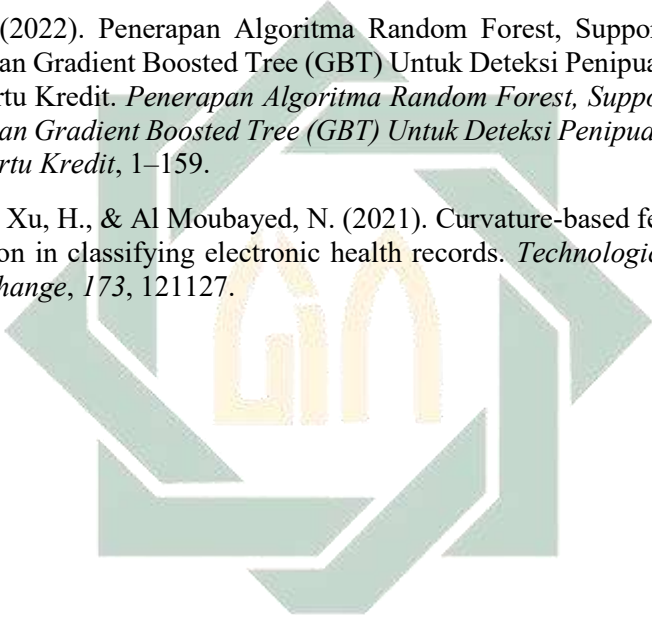


UIN SUNAN AMPEL
S U R A B A Y A

DAFTAR PUSTAKA

- Abdurrochman, M., Wibowo, A., Rosal, D., & Setiadi, I. M. (2024). Optimized Machine Learning Model for Credit Card Fraud Detection Using Smote-Tomek and Feature Engineering. *Journal of Applied Informatics and Computing (JAIC)*, 8(2), 580.
- Amiarrahman, M. R., & Handhika, T. (2018). Analisis dan implementasi algoritma klasifikasi Random Forest dalam pengenalan Bahasa Isyarat Indonesia (BISINDO). *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, 2(1), 83–88.
- Arsanto, A. T., Faizin, A., Lutfi, M., & Saadah, Z. N. (2024). Optimization of the Naive Bayes Algorithm with SMOTETomek Combination for Imbalance Class Fraud Detection. *Sistemasi*, 13(6), 2709. <https://doi.org/10.32520/stmsi.v13i6.4719>
- Azizah, A. N., Ritonga, A. S., & Atmojo, S. (2025). *Graph-Based Fraud Detection with Optimized Features and Class Balance*. 6(3), 227–239.
- Chen, J., Huang, H., Cohn, A. G., Zhang, D., & Zhou, M. (2022). Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning. *International Journal of Mining Science and Technology*, 32(2), 309–322.
- Dutta, P., Paul, S., Cengiz, K., Anand, R., & Kumar, A. (2023). A predictive method for emotional sentiment analysis by deep learning from EEG of brainwave dataset. In *Artificial Intelligence for Neurological Disorders* (pp. 25–48). Elsevier.
- Farida, & Mustopa, A. (2023). Perbandingan Logistic Regression dan Random Forest menggunakan Correlation-based Feature Selection untuk Deteksi Website Phishing. *SISTEMASI: Jurnal Sistem Informasi*, 12(1), 13–20.
- Fawcett, T. (2006). *An introduction to ROC analysis*. 27, 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning*, 113(7), 4845–4901.
- Gopika, N., & M.E., A. M. K. (2018). Correlation Based Feature Selection Algorithm for Machine Learning. *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, 692–695. <https://doi.org/10.1109/CESYS.2018.8723980>
- Hairani, H., Anggrawan, A., & Priyanto, D. (2023). Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link. *JOIV: International Journal on Informatics Visualization*, 7(1), 258–264.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239.
- Liang, J. (n.d.). *Confusion matrix: machine learning*. *POGIL Activity Clearinghouse*; 2022.
- Nadira, Z. (n.d.). *CYBERCRIME COUNSELING BY MAXIMIZING 2-FACTOR AUTHENTICATION (2-FA) FEATURES IN SEBATANG , KULON PROGO*. 60–70.

- Pyar, K. (2024). Improvement The Accuracy of Convolutional Neural Network with Using Undersampling Method on Unbalanced Credit Card Dataset. *International Journal of Advances in Data and Information Systems*, 5(2). <https://doi.org/10.59395/ijadis.v5i2.1333>
- ROMADI, S. (2025). *PERBANDINGAN KINERJA ALGORITMA GRADIENT BOOSTING DAN MULTILAYER PERCEPTRON DALAM KLASIFIKASI WEBSITE PHISHING*. Universitas Mercu Buana Jakarta.
- Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, November, 4023–4031. <https://doi.org/10.53555/ajbr.v27i4s.4345>
- Subramaniam, D. N., Jeyanathan, P., & Sathiparan, N. (2024). Soft computing techniques to predict the electrical resistivity of pervious concrete. *Asian Journal of Civil Engineering*, 25(1), 711–722.
- Zuhairah, A. (2022). Penerapan Algoritma Random Forest, Support Vector Machines (SVM) dan Gradient Boosted Tree (GBT) Untuk Deteksi Penipuan (Fraud Detection) Pada Kartu Kredit. *Penerapan Algoritma Random Forest, Support Vector Machines (SVM) Dan Gradient Boosted Tree (GBT) Untuk Deteksi Penipuan (Fraud Detection) Pada Kartu Kredit*, 1–159.
- Zuo, Z., Li, J., Xu, H., & Al Moubayed, N. (2021). Curvature-based feature selection with application in classifying electronic health records. *Technological Forecasting and Social Change*, 173, 121127.



UIN SUNAN AMPEL
SURABAYA