ISSN (p) : 2407-1323 DOI : https://doi.org/10.25124/jett.v9i1.5019
ISSN (e) : 2442-4404 | Vol. 9 | No. 1 | Halaman : 1152 - 1161

JETT

Jurnal Elektro Telekomunikasi Terapan Juli 2022

ANALISIS PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA) PADA DETEKSI KECURANGAN KARTU KREDIT MENGGUNAKAN RANDOM FOREST

AN ANALYSIS OF PRINCIPAL COMPONENT ANALYSIS IMPLEMENTATION ON CREDIT CARD FRAUD DETECTION USING RANDOM FOREST

Khabib Astoni[1], Muhammad Haris[2]
[1,2] Universitas Nusa Mandiri
[1]14002414@nusamandiri.ac.id, [2]muhammad.uhs@nusamandiri.ac.id

Abstrak
Pandemi COVID-19 yang menjangkiti hampir seluruh penjuru dunia menyebabkan berbagai perubahan hampir pada semua bidang. Demi mencegah penyebarannya, banyak negara yang menerapkan protokol ketat yang membatasi mobilitas warganya, yang kemudian menyebabkan peralihan dalam dunia perdagangan dimana meningkatnya transaksi secara online. Seiring dengan meningkatnya transaksi secara online, diikuti pula dengan meningkatnya penggunaan kartu kredit yang memang memudahkan, tetapi hal tersebut juga diikuti oleh meningkatnya ancaman keamanan. Salah satu ancaman yang mengintai adalah penggunaan kartu kredit secara ilegal oleh orang lain yang sangat merugikan baik bagi penggunanya maupun penyedia layanan kartu kredit sehingga diperlukan langkah untuk mendeteksi transaksi yang dicurigai sebagai kecurangan. Penelitian ini bertujuan untuk membangun suatu model yang mampu melakukan deteksi potensi kecurangan penggunaan kartu kredit. Penelitian ini menggunakan algoritma Random Forest yang kemudian dikombinasikan dengan SMOTE dan PCA. Dari beberapa skenario yang dilakukan, kemudian dilakukan analisa performa model dari masing-masing metode kombinasi. Hasil penelitian ini menunjukkan bahwa model dengan algoritma dasar menghasilkan nilai recall yang lebih tinggi, sedangkan model dengan penerapan PCA menghasilkan nilai precision yang jauh lebih baik
Kata kunci: Deteksi kecurangan, PCA, Random Forest, SMOTE

Abstract
The COVID-19 pandemic that infected almost all country of the world caused changes in almost all fields. In order to prevent its spread, many countries are implementing strict protocols that limit the mobility of their citizens, which then leads to a shift in the world of trade where online transactions are increasing. Along with the increase in online transactions, followed by the increasing use of credit cards that are indeed easy, but it is also followed by an increase in security threats. One of the lurking threats is the illegal use of credit cards by others that are so detrimental to both users and credit card service providers, so that steps are needed to detect suspected fraudulent transactions. This research aims to build a model that is able to detect potential credit cards fraud using. The study used the Random Forest algorithm which was then combined with SMOTE and PCA. From several scenarios carried out, then conducted an analysis of the performance model of each combination method. The results of this study showed that models with basic algorithms resulted in higher recall values, whereas models with PCA applications produced much better precision values.
Keywords: Fraud Detection, PCA, Random Forest, SMOTE

1. PENDAHULUAN

Pandemi COVID-19 yang membatasi mobilitas masyarakat membuka peluang semakin diminatinya transaksi jual beli online yang mengandalkan pembayaran non tunai karena dinilai praktis dan cepat. Salah satu metode pembayaran non tunai adalah menggunakan kartu kredit. Berdasarkan laporan dari Bank Indonesia, nilai transaksi menggunakan kartu kredit sepanjang Bulan Desember 2021 mengalami kenaikan sebesar 10,39% dibandingkan dengan bulan sebelumnya, selain itu, volume transaksi menggunakan kartu kredit juga mengalami peningkatan sebesar 5,57% [1].

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

Senada dengan peningkatan penggunaannya, ancaman besar mengintai terhadap keamanan pengguna kartu kredit yang menimbulkan kerugian finansial yang jumlahnya tidak sedikit. Korban tindak kecurangan kartu kredit yang menimpa pengguna dari Amerika serikat dilaporkan semakin meningkat dari tahun ke tahun dimana jumlah kerugian pada tahun 2020 enam kali lipat dari tahun 2019 [2].

Dari permasalahan di atas, sangat penting dilakukan langkah khusus untuk mendeteksi tindak kecurangan penggunaan kartu kredit untuk meminimalisir kerugian yang terjadi. Metode yang diusulkan dalam penelitian ini diharapkan mampu untuk membangun sebuah model untuk memprediksi kecurangan penggunaan kartu kredit berdasarkan variabel-variabel yang ada pada dataset dan menentukan variabel apa saja yang mempunyai pengaruh paling besar pada setiap model yang dibangun

Penelitian yang terkait dengan prediksi kecurangan kartu kredit sebelumnya pernah dilakukan oleh Ismini Psychoula, Andreas Gutmann, Pradip Mainali, S.H. Lee, Paul Dunphy, dan Fabien A.P. Petitcolas dengan judul *Explainable Machine Learning for Fraud Detection* [3]. Penelitian ini membahas tentang deteksi kecurangan pada transaksi kartu kredit menggunakan *Naive Bayes*, *Logistic regression*, *Decision trees*, *Gradient boosted trees*, *Random forests*, *Neural network*, *Autoencoder* dan *Isolation forest* dengan paling baik pada Algoritma *Autoencoder* dengan hasil performa terbaik pada Algoritma *Autoencoder* dimana nilai presisi sebesar 0.944, *recall* 0.767, F1 *Score* 0.839 dan skor AUC sebesar 0.617. Kemudian penelitian oleh Yazid [4] dengan judul Mendeteksi Kecurangan Pada Transaksi Kartu Kredit Untuk Verifikasi Transaksi Menggunakan Metode SVM yang mendeteksi data *outlier* sebagai indikator tindak kecurangan.

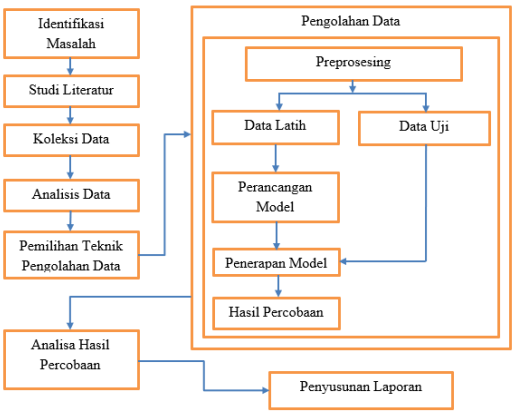
Dalam penelitian yang berkaitan dengan kecurangan kartu kredit ini, salah satu masalah yang dihadapi adalah dimensi fitur yang jumlahnya sangat besar sehingga diperlukan langkah-langkah tertentu untuk melakukan reduksi fitur. Proses reduksi fitur ini dilakukan dengan menerapkan *Principal component analysis* (PCA) yang mampu melakukan eliminasi korelasi antara satu variabel dengan variabel lain yang dijadikan sebagai masukannya.

Tujuan dari penelitian ini adalah melakukan analisa dan klasifikasi tindak kecurangan kartu kredit akan dilakukan dengan menerapkan algoritma *Random Forest* yang kemudian akan dikombinasikan dengan metode *Synthetic Minority Oversampling Technique* (SMOTE) dan reduksi fitur menggunakan PCA.

2. METODE PENELITIAN

2.1. Tahapan Penelitian

Gambar 1 di bawah menjelaskan tahapan-tahapan penelitian yang dilakukan secara lengkap.



Gambar 1. Tahapan Penelitian

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

Dalam penelitian ini terdapat tahap yang dilakukan, yang secara umum terbagi atas tiga tahapan. Tahap awal dimulai dengan identifikasi masalah hingga analisis dan persiapan pengolahan data, kemudian dilanjutkan dengan pengolahan data hingga hasil penerapan model, dan tahap akhir yaitu analisa hasil pengujian serta penyusunan laporan.

2.2. Dataset

Dataset yang digunakan dalam penelitian ini adalah data transaksi pelanggan yang dikumpulkan oleh peneliti dari IEEE Computational Intelligence Society (IEEE-CIS) dan dapat diakses di laman web Kaggle. Dataset awal yang didapatkan terpecah menjadi dua file, yaitu dataset yang memuat identitas dan dataset lainnya memuat data transaksi. Data identitas terdiri atas 41 fitur sedangkan data transaksi terdiri atas 393 fitur dan satu kelas yaitu 'isFraud' yang nilainya 0 dan 1. Jumlah baris data pada kedua file tersebut berbeda dimana data identitas terdiri atas 144233 baris data dan data transaksi yang berjumlah 590540 baris data.

2.3. Data *Preprocessing*

Sebelum dapat diolah lebih lanjut, data dilakukan *preprocessing* terlebih dahulu agar memudahkan percobaan dan menghasilkan performa model yang lebih baik. Pada Langkah ini dilakukan beberapa tahapan dimana tahap yang pertama adalah seleksi fitur. Fitur yang dieliminasi adalah fitur yang isinya 100% null dan fitur-fitur yang tidak mempunyai pengaruh signifikan.

Langkah *preprocessing* kedua adalah penanganan *missing value*. Pada Langkah ini dilakukan untuk mengatasi nilai *null* dengan nilai tertentu berdasarkan karakteristik dari fitur tersebut. Dalam penelitian ini, nilai *null* diisi dengan nilai rata-rata atau diisi dengan nilai tertentu yang disesuaikan berdasarkan isi data dari masing-masing fitur.

Langkah ketiga adalah melakukan proses *feature engineering* yang meliputi konversi data, reduksi fitur, *one hot encoding* dan langkah lainnya. Konversi data pada penelitian ini pada umunya dilakukan pada fitur-fitur yang berjenis teks ataupun kategorikal. Sebelum dilakukan reduksi dimensi dari suatu fitur, dilakukan pengelompokkan terlebih dahulu berdasarkan karakteristik dari data pada fitur tersebut, sebagai ilustrasi, pada fitur yang berisi tipe sebuah telepon genggam, kemudian disederhanakan hanya diambil mereknya saja, ilustrasi lainnya adalah pada fitur yang berisi kategori domain email, maka akan disederhanakan menjadi domain utama saja, contohnya adalah yahoo.com, yahoo.co.id, yahoo.co.uk dan seterusnya kemudian disederhanakan menjadi Yahoo saja, sehingga dapat mereduksi variansi data pada fitur tersebut.

Seleksi fitur pada *preprocessing* tahap awal ini hanya difokuskan pada fitur-fitur yang semua nilainya adalah *null*, sedangkan untuk seleksi fitur tingkat lanjut akan menerapkan metode Principal component analysis (PCA).

2.4. Skenario Percobaan

Dalam penelitian ini terdapat beberapa skenario yang dilakukan yaitu penerapan model RF saja tanpa melakukan *tuning* parameter, kemudian penerapan RF yang dikombinasikan dengan teknik SMOTE, kombinasi RF dengan PCA saja, dan yang terakhir adalah penerapan RF yang dikombinasikan dengan teknik PCA dan SMOTE. Hasil dari skenario - skenario pengujian tersebut nantinya akan dievaluasi dan dianalisis lebih lanjut.

3. DASAR TEORI

3.1. Random Forest

Random Forest (RF) merupakan metode ensemble yang menggabungkan beberapa pohon keputusan dimana setiap pohon menghasilkan aturan klasifikasi untuk satu kelas [5]. Secara

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

mendasar, RF mempunyai cara yang kerja yang sama dengan *Decision Tree* (DT) dalam menentukan node akar dan aturannya. Hanya saja, RF yang terdiri atas lebih dari satu DT secara teori akan menghasilkan performa yang lebih baik jika dibandingkan dengan DT [6]. Algoritma RF tergolong dalam metode klasifikasi yang *supervised* [7].

3.2. *Principal component analysis* (PCA)

Data dengan dimensi fitur yang sangat besar pada umumnya akan membutuhkan sumber daya yang lebih besar pula untuk diolah. Selain itu, dimensi data yang lebih sederhana relatif lebih mampu untuk menghasilkan model prediktif yang jauh lebih baik performanya [8]. *Principal component analysis* (PCA) adalah salah satu metode yang dapat dilakukan untuk mengurangi dimensi fitur [8] [9]. PCA digunakan untuk mereduksi variabel dengan korelasi yang besar menjadi data dengan variabel yang lebih sedikit [10]. Cara kerja PCA adalah dengan menciptakan garis-garis yang dinamai *principal components* melalui operasi matriks sederhana dari aljabar linear dan statistik untuk melakukan proyeksi dari dimensi data asli menjadi data dengan jumlah yang sama atau lebih sedikit dari dimensi data awal.

3.3. Synthetic Minority Over-sampling Technique

Synthetic Minority Oversampling Technique (SMOTE) merupakan suatu teknik statistik *oversampling* yaitu dengan melakukan duplikasi pada data minoritas agar didapatkan kelas data yang seimbang [11]. Pada SMOTE, proses duplikasi data minoritas tidak hanya sebatas proses menyalin data saja, akan tetapi dengan melakukan suatu proses yang melibatkan tetangga terdekat dengan kelas target yang kemudian menghasilkan data baru dari pendekatan tersebut [12] [13].

3.4. Performa Peramalan

Pengukuran performa model pada dataset yang *imbalanced* dimana satu kelas jauh lebih mendominasi dari kelas yang lain, penggunaan nilai akurasi klasifikasi saja tidak cukup kuat untuk menggambarkan kinerja keseluruhan sebuah model, dimana dalam kasus ini, akurasi pada prediksi kelas minoritas jauh lebih penting daripada akurasi secara keseluruhan sehingga dipakailah metrik pengukuran yang lain, contohnya *recall*, *precision*, F1 Score ataupun *Area Under the Curve* (AUC) Score [14].

Recall atau seringkali juga disebut dengan *sensitivity* adalah perbandingan antara data positif yang diprediksi benar dibandingkan dengan keseluruhan data positif. Nilai *recall* dapat dihitung dengan Rumus 1 di berikut ini.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (1)$$

Precision adalah perbandingan antara data data positif yang diprediksi benar dengan keseluruhan data yang diprediksi positif. Nilai dapat dihitung Rumus 2 berikut ini.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (2)$$

F1 Score atau *F-Measure* mencari nilai tengah antara nilai *recall* dan *precision* atau bisa dibilang merupakan nilai rerata yang harmonik dari kedua parameter tersebut karena keduanya mempunyai karakteristik yang saling bertolak belakang. Semakin tinggi nilai *recall*, maka akan semakin rendah nilai *precision*-nya. Nilai *F1 Score* ini merupakan solusi untuk mencari keseimbangan dalam model, dan rumusnya adalah sebagai berikut.

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

$$\text{F1 Score} = 2(\text{recall x precision}) / (\text{recall} + \text{precision}) \quad (3)$$

Sedangkan nilai akurasi dapat dihitung dengan rumus 4 berikut ini.

$$\text{Akurasi} = (\text{TP x TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (4)$$

Nilai AUC merupakan metrik pengukuran yang banyak digunakan pada data dengan kelas yang tidak seimbang dimana semakin tinggi nilainya, semakin baik pula model yang dibangun. AUC menggambarkan kemampuan suatu model untuk membedakan dua kelas yang berbeda. Nilai AUC berkisar antara 0,5 dan 1, yang secara analogi bisa digambarkan misalnya sebuah model mempunyai nilai AUC 0,75 atau 75%, bisa diartikan bahwa model tersebut mempunyai kemungkinan sebesar 75% untuk membedakan antara satu kelas dengan kelas yang lainnya.

4. HASIL PENELITIAN

4.1. *Preprocessing Data* dan *Feature Engineering*

Sebelum dapat dapat dilakukan penerapan pada model algoritma, data terlebih dahulu dianalisa untuk dilakukan pembersihan, transformasi data ataupun pemilihan fitur untuk memastikan bahwa tidak ada *missing values*, *outliers* dan kesalahan lain sehingga menghasilkan dataset yang handal untuk diproses lebih lanjut.

Langkah pertama yang dilakukan adalah melakukan pembersihan fitur yang semua nilainya *null*. Dari fitur-fitur tersebut, masih terdapat banyak fitur yang sebagian berisi data *null* sehingga diperlukan penanganan dalam pengisiannya berdasarkan karakteristik pada masing-masing fitur. Penanganan *missing values* pada semua fitur yang terdapat nilai *null* secara detail dijelaskan dalam Tabel 1.

Tabel 1. Penanganan *Missing Values*

Fitur	Tipe Data	*Value* Baru
id_02, id_03, id_04, id_05, id_06, id_07, id_08, id_09, id_10, id_11, id_13, id_14, id_17, id_18, id_19, id_20, id_21, id_22, id_24, id_25, id_26, id_32	Numerik	Nilai rerata dari masing-masing fitur
id_30, id_31, id_33, id_34, card4, card6	Objek	'Other'
id_12, id_16, id_27, id_28, id_29	Objek	'Not Found'
id_15, id_35, id_36, id_37, id_38, DeviceType, DeviceInfo, P_emaildomain, R_emaildomain	Objek	'Unknown'
id_23	Objek	'IP_PROXY: HIDDEN'
D1 - D15	Numerik	'-1'
M4	Objek	'Unknown'
V1 - V339	Numerik	'-1'

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

Tabel 2. Penyederhanaan Fitur

Fitur	Jumlah kategori semula	Jumlah kategori setelah disederhanakan
id_30	75	7
id_31	130	11
id_33	260	11
DeviceInfo	1786	47
P_emaildomain	59	26
R_emaildomain	60	27

Setelah semua fitur mempunyai *value*, terdapat permasalahan lebih lanjut dimana *instances* pada fitur-fitur tersebut sangat variatif sehingga perlu dilakukan penyederhanaan untuk meningkatkan performa model yang dibangun. Contohnya adalah melakukan penyederhanaan pada data yang berjenis kategori ataupun penyederhanaan lain seperti pengelompokan email berdasarkan domain atau penyedia layanannya, penyederhanaan tipe perangkat menjadi merek, sistem operasi, resolusi layar dan lain-lain.

4.2. Percobaan dan Pengujian Model

Pada penelitian ini dilakukan setidaknya dua skenario utama dimana skenario pertama dilakukan pada dataset tanpa dilakukan reduksi fitur, sedangkan pada skenario kedua dilakukan reduksi fitur menggunakan metode PCA yang kemudian hasil dari masing-masing skenario percobaan tersebut kemudian dibandingkan performanya.

Skenario yang pertama adalah model dasar *Random forest* tanpa reduksi fitur dimana jumlah fitur yang awalnya 411 kemudian bertambah menjadi 529 fitur karena telah melalui tahap *one hot encoding* pada fitur yang bersifat kategorikal. Model ini adalah model dasar atau model *baseline*. Model *baseline* ini juga kemudian dikombinasikan lagi dengan metode SMOTE untuk mengetahui sejauh mana pengaruh penerapannya pada dataset ini dengan pertimbangan pokok dimana dataset yang digunakan bersifat *imbalanced*. Hasil pengujian pada skenario pertama terdapat pada Tabel berikut ini.

Tabel 3. Performa Model Baseline

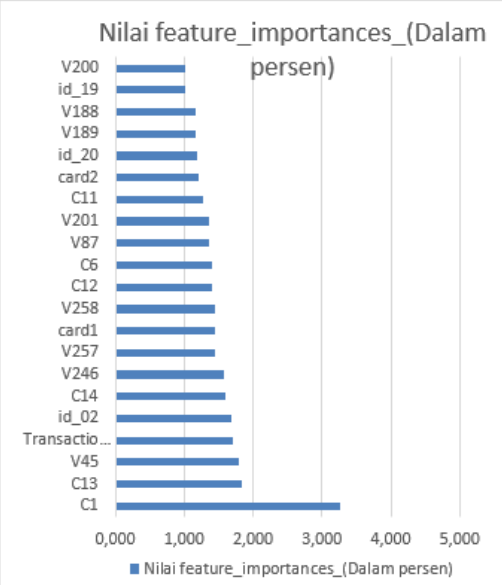
Metrik	Baseline	Baseline - SMOTE
Akurasi	0.968	0.968
Recall	0.645	**0.647**
Precision	**0.927**	0.926
F1 Score	0.761	**0.762**
Skor AUC	0.820	**0.821**

Dari hasil pengujian yang telah dilakukan, penerapan SMOTE ternyata tidak terlalu menghasilkan dampak yang signifikan, bahkan bisa dikatakan sama saja dimana model menghasilkan tingkat akurasi dan *precision* yang tinggi yaitu masing-masing 0.96 dan 0.92. Nilai *F1-Score* dan AUC *Score* yang dihasilkan bisa dikatakan baik, yaitu 0.761 dan 0.820. Namun demikian, nilai *recall* yang dihasilkan pada model ini tergolong rendah, yaitu hanya di angka 0.64.

Pada skenario pertama ini, terdapat 21 fitur yang mempunyai nilai *importance* lebih dari satu persen, dimana fitur C1 mempunyai pengaruh yang paling besar dalam menetukan apakah data transaksi terindikasi *fraud* atau tidak. Selain itu, dapat dilihat bahwa fitur-fitur pada sub dataset

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

transaksi lebih mendominasi dalam menentukan kelas data dari pada data identitas dimana dari 21 satu fitur, hanya terdapat dua fitur identitas yaitu id_02 dan id_19.



Gambar 2. *Feature Importance Score* Paling tinggi pada model dasar

Sebelum masuk ke skenario kedua yang menerapkan PCA, terlebih dahulu dilakukan *splitting* dataset menjadi beberapa bagian dengan tujuan untuk mempermudah proses penerapan PCA berdasarkan kelompok fitur tertentu. Setiap bagian tersebut kemudian ditambahkan indeks agar memudahkan dalam penggabungan kembali.

Tabel 4. *Split* dataset menjadi beberapa bagian

Bagian	Keterangan
Bagian 1	Fitur TransactionID sampai dengan R_emaildomain
Bagian 2	Fitur C1 sampai dengan C14
Bagian 3	Fitur D1 sampai dengan D15
Bagian 4	Fitur M4
Bagian 5	Fitur V1 sampai dengan V339

Pada penelitian ini, penerapan PCA dilakukan pada satu fitur berjenis kategorikal dimana jumlah kategorinya masih sangat banyak yang kemudian dilakukan proses *one hot encoding*, maupun sekumpulan fitur yang mempunyai karakteristik hampir sama.

Tabel 5. Penerapan PCA pada fitur

Fitur	Jumlah fitur semula	Jumlah fitur setelah disederhanakan
DeviceInfo	47	5
P_emaildomain	26	5
R_emaildomain	27	5
C1 – C14	14	3
D1 – D15	14	3
V12 – V339	328	4

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v9i1.5019
| Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

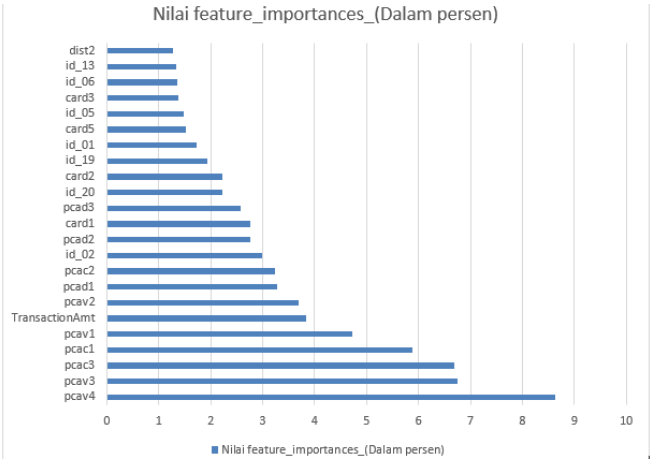
Sub dataset yang masing-masing telah di-PCA-kan tersebut kemudian digabungkan kembali dan kemudian dilanjutkan pada skenario pengujian yang kedua yaitu melakukan pengujian model menggunakan algoritma *Random Forest* yang kemudian dilanjutkan dengan penerapan SMOTE pada dataset yang telah di-PCA kan, sehingga seperti pada skenario sebelumnya, pada skenario ini juga terdapat dua sub-model. Berbeda dari skenario yang pertama, fitur dengan tipe data kategorikal diubah menggunakan metode *label encoder* untuk meminimalisir jumlah fitur walaupun masih tetap menggunakan metode *one hot encoding*. Dari jumlah fitur yang awalnya 77, fitur kemudian bertambah menjadi 98 setelah dilakukan *one hot encoding*.

Dari pengujian yang telah dilakukan, penerapan PCA menghasilkan performa model dengan karakteristik yang agak berbeda dan jelas nampak pada *recall* yang nilainya mengalami penurunan dibandingkan dengan skenario pengujian pertama dimana didapatkan nilai 0.616 dan 0.617, sedangkan nilai *precision* mengalami kenaikan dibandingkan dengan skenario yang pertama dimana pada skenario ini didapatkan nilai 0.961 dan 0.953. Performa dari model yang telah dikembangkan menggunakan skenario kedua ini dapat dilihat pada Tabel 4.5.

Tabel 6. Performa Model dengan PCA

Metrik	PCA	PCA - SMOTE
Akurasi	0.968	0.968
Recall	0.616	**0.617**
Precision	**0.961**	0.953
F1 Score	**0.751**	0.749
AUC Score	0.807	0.807

Pada skenario kedua ini, setelah dilakukan reduksi fitur menggunakan metode PCA terdapat 23 dari 98 fitur yang mempunyai nilai *importance* lebih dari satu persen, dimana fitur pcav4 mempunyai pengaruh yang paling besar dalam menetukan apakah data transaksi terindikasi *fraud* atau tidak. Dari 23 fitur tersebut, 10 fitur diantaranya merupakan hasil reduksi fitur sub dataset transaksi menggunakan metode PCA, 6 fitur identitas dan 7 fitur sisanya merupakan fitur yang terdapat pada sub dataset transaksi. Fitur-fitur pada sub dataset transaksi masih mendominasi *feature importance* walaupun tidak sedominan pada model dasar.



Gambar 3. *Feature Importance Score* Paling tinggi pada model dasar

ISSN (p) : 2407-1323 DOI : https://doi.org/10.25124/jett.v9i1.5019
ISSN (e) : 2442-4404 | Vol. 9 | No. 1 | Halaman : 1152 - 1161
JETT
Jurnal Elektro Telekomunikasi Terapan Juli 2022

4.3. Evaluasi Model

Dari skenario pengujian dengan menggunakan model dasar, model dasar dengan SMOTE serta penggunaan metode PCA untuk reduksi fitur, terdapat persamaan dan perbedaan mendasar pada setiap model. Aplikasi SMOTE pada skenario pertama dan kedua tidak banyak memberikan pengaruh terhadap model yang dibangun.

Jika dilihat performa model pada Tabel 3, nilai akurasi dari setiap model berkisar di angka 0.968. Namun demikian, bisa dikatakan bahwa akurasi kurang cocok untuk diaplikasikan pada data *imbalanced*, sehingga pada penelitian ini, fokus utama terletak pada *Recall*, *Precision* dan skor AUC. Hal ini disebabkan karena karakteristik dari dataset dimana sebisa mungkin tingkat prediksi *False Negative* ataupun *False Positive* ditekan serendah mungkin.

Tabel 7. Perbandingan Performa Model Keseluruhan

Metrik	Baseline	Baseline - SMOTE	PCA	PCA – SMOTE
Akurasi	0.968	0.968	0.968	0.968
Recall	0.645	**0.647**	0.616	0.617
Precision	0.927	0.926	**0.961**	0.953
F1 Score	0.761	**0.762**	0.751	0.749
AUC Score	0.820	**0.821**	0.807	0.807

Dari sisi bisnis, *False Negative* pada deteksi kecurangan penggunaan kartu kredit akan menyebabkan kerugian baik itu bagi pengguna kartu kredit maupun penyedia jasanya karena transaksi yang seharusnya tidak sah ternyata bisa lolos dari pengamatan sistem. Sedangkan *False Positive*, dalam konteks kecurangan penggunaan kartu kredit bisa diilustrasikan sebagai transaksi yang sebetulnya tidak bermasalah, tetapi oleh sistem dideteksi sebagai kecurangan. Bagi pengguna, tentu saja hal ini akan sedikit merepotkan karena dipastikan akan mengurangi kemudahan dan menjadi hambatan dalam transaksi itu sendiri.

Model yang dibangun pada penelitian ini menghasilkan nilai *False Positive* yang relatif rendah. Artinya, model yang dibangun menghasilkan performa yang baik dalam memprediksi kelas yang dideteksi sebagai kelas *Non Fraud*. Sedangkan rerata *False Negative* pada model ini tidak sebaik rerata *False Positive* yang artinya hasil prediksi untuk kelas data positif tidak sebaik dalam memprediksi kelas data negatif. Skor AUC yang mencapai 0,821 dapat diartikan bahwa model mempunyai kapabilitas untuk memisahkan kelas data secara benar dengan prosentase 82,1%.

5. KESIMPULAN

Dari hasil penelitian ini telah dapat dihasilkan model untuk memprediksi kecurangan penggunaan kartu kredit menggunakan algoritma *Random Forest* dengan empat skenario dimana nilai akurasi pada semua skenario pengujian menghasilkan angka yang seragam yaitu mencapai 96,8%. Penerapan SMOTE terlihat tidak berpengaruh secara signifikan. Selain itu, terdapat perbedaan hasil antara model yang menerapkan PCA dengan model tanpa PCA dimana model tanpa PCA menghasilkan nilai *recall*, F1 *Score* dan AUC yang lebih baik dibandingkan dengan model yang menggunakan PCA, yaitu 64,7% berbanding 61,7% pada nilai *recall*, sedangkan nilai F1 *Score* 76,2% berbanding 75,1% dan untuk nilai AUC 82,1% berbanding 80,7%. Namun sebaliknya, nilai *precision* pada model dengan PCA jauh lebih tinggi dari pada model yang tidak menggunakan teknik PCA, yaitu 96,1% berbanding 92,7%. Selain itu itu, waktu yang dibutuhkan untuk melakukan *running*, model dengan penerapan PCA jauh lebih cepat dibandingkan dengan model dasar karena penerapan PCA berhasil mereduksi jumlah fitur secara signifikan sehingga pemrosesan data menjadi lebih cepat.

DAFTAR PUSTAKA

[1] Cindy Mutia Annur, "BI: Nilai Transaksi Kartu Kredit RI Tumbuh 10,39% pada Desember 2021 | Databoks." http://databoks.katadata.co.id/datapublish/2022/01/31/bi-nilai-transaksi-kartu-kredit-ri-tumbuh-1039-pada-desember-2021 (accessed Mar. 03, 2022).

[2] "Credit Card Fraud 2021 Annual Report: Prevalence, Awareness, and Prevention - Security.org." https://www.security.org/digital-safety/credit-card-fraud-report/ (accessed Mar. 03, 2022).

[3] I. PSYCHOULA, A. Gutmann, P. Mainali, S. H. Lee, P. Dunphy, and F. Petitcolas, "Explainable Machine Learning for Fraud Detection," *Computer (Long. Beach. Calif).*, vol. 54, no. 10, pp. 49–59, Oct. 2021, doi: 10.1109/MC.2021.3081249.

[4] Y. Yazid and A. Fiananta, "MENDETEKSI KECURANGAN PADA TRANSAKSI KARTU KREDIT UNTUK VERIFIKASI TRANSAKSI MENGGUNAKAN METODE SVM," *Indones. J. Appl. Informatics*, vol. 1, no. 2, pp. 61–66, May 2017, doi: 10.20961/IJAI.V1I2.14378.

[5] L. Breiman, "Random forests," *Mach. Learn.*, 2001, doi: 10.1023/A:1010933404324.

[6] Y. Azhar, G. A. Mahesa, and M. C. Mustaqim, "Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 1, pp. 15–21, Jan. 2021, doi: 10.14710/jtsiskom.2020.13790.

[7] R. A. Haristu, "Penerapan metode Random Forest untuk prediksi win ratio pemain player Unknown Battleground," 2019.

[8] Muhtadi, "Penerapan Principal Component Analysis (PCA) dalam Algoritma K-Means untuk Menentukan Centroid pada Clustering," *Konstanta*, vol. 1, no. 1. pp. 122–142, Jul. 17, 2017, Accessed: Dec. 25, 2021. [Online]. Available: https://journal.iainkudus.ac.id/index.php/Konstanta/article/view/3543.

[9] K. Chen, "Indirect PCA Dimensionality Reduction Based Machine Learning Algorithms for Power System Transient Stability Assessment," *2019 IEEE PES Innov. Smart Grid Technol. Asia, ISGT 2019*, pp. 4175–4179, May 2019, doi: 10.1109/ISGT-ASIA.2019.8881370.

[10] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker, and Shantanu, "Data analysis using principal component analysis," *2014 Int. Conf. Med. Imaging, m-Health Emerg. Commun. Syst. MedCom 2014*, pp. 45–48, 2014, doi: 10.1109/MEDCOM.2014.7005973.

[11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: 10.14710/JTSISKOM.8.2.2020.89-93.

[12] M. Mustaqim, B. Warsito, and B. Surarso, "Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 5, no. 2, pp. 116–127, Jul. 2019, doi: 10.26594/register.v5i2.1705.

[13] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," *J. Inf. Syst. Dev.*, vol. 3, no. 1, pp. 2528–5114, Feb. 2018, Accessed: Dec. 25, 2021. [Online]. Available: https://ejournal.medan.uph.edu/index.php/isd/article/view/177.

[14] 14611240 Julia Widiastuti, "KLASIFIKASI PEMBIAYAAN WARUNG MIKRO MENGGUNAKAN METODE RANDOM FOREST DENGAN TEKNIK SAMPLING KELAS IMBALANCED (Studi Kasus: Data Nasabah Pembiayaan Warung Mikro Bank Syariah Mandiri KC Jambi)," May 2018, Accessed: Jan. 04, 2022. [Online]. Available: https://dspace.uii.ac.id/handle/123456789/7690.