Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 4, No. 11, November 2024, Hal. 525-535
DOI: https://doi.org/10.52436/1.jpti.518
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Implementasi XGBoost dan SMOTE untuk Meningkatkan Deteksi Transaksi Fraud di Industri Jasa Keuangan

Rinno Yunanto[*1], Utomo Budiyanto[2]

[1,2]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia
Email: [1]rinnoyunanto@gmail.com, [2]utomo.budiyanto@budiluhur.ac.id

Abstrak

Tindakan fraud *(fraud)* di industri jasa keuangan, khususnya pada Bank Perekonomian Rakyat (BPR), menjadi tantangan serius yang memengaruhi stabilitas keuangan, kepercayaan nasabah, dan perekonomian nasional. Masalah ini semakin sulit terdeteksi dengan berkembangnya teknologi yang mempermudah pelaku melakukan manipulasi laporan keuangan dan penggelapan dana. Penelitian ini bertujuan untuk mengembangkan model deteksi fraud berbasis algoritma *XGBoost* yang dipadukan dengan teknik *SMOTE (Synthetic Minority Over-sampling Technique)* untuk mengatasi ketidakseimbangan data. Melalui analisis terhadap data transaksi keuangan BPR, model ini menunjukkan kemampuan yang tinggi dalam mendeteksi transaksi mencurigakan. Hasil penelitian menunjukkan bahwa model *XGBoost* dapat meningkatkan akurasi deteksi fraud hingga 96%, dengan nilai *Area Under the Curve (AUC)* yang signifikan. Pendekatan ini tidak hanya efektif dalam mendeteksi transaksi mencurigakan, tetapi juga memberikan kontribusi dalam memperkuat sistem pengendalian internal dan menerapkan prinsip tata kelola perusahaan yang baik *(good corporate governance)*. Penelitian ini memberikan solusi inovatif dalam pencegahan *fraud* di BPR dan berperan penting bagi pengembangan ilmu pengetahuan di bidang data sains serta praktik industri. Diharapkan, hasil penelitian ini dapat menjadi panduan strategis bagi manajemen BPR dalam merancang sistem deteksi *fraud* yang lebih efektif, sehingga meningkatkan kepercayaan masyarakat terhadap institusi keuangan.

Kata kunci: *bpr, deteksi fraud, machine learning, smote, transaksi keuangan, xgboost*

Implementing Xgboost Models For Enhanced Detection Of Fraud Transaction In Financial Services Industries

Abstract

Fraud in the financial services industry, particularly in Rural Banks (BPR), has become a serious challenge affecting financial stability, customer trust, and the national economy. This issue has become increasingly difficult to detect due to the advancement of technology, which facilitates perpetrators in manipulating financial reports and embezzling funds. This study aims to develop a fraud detection model based on the XGBoost algorithm combined with the SMOTE (Synthetic Minority Over-sampling Technique) to address data imbalance. By analyzing BPR financial transaction data, the model demonstrates high capability in detecting suspicious transactions. The results show that the XGBoost model can improve fraud detection accuracy to 96%, with a significant Area Under the Curve (AUC) value. This approach is not only effective in detecting suspicious transactions but also contributes to strengthening internal control systems and applying good corporate governance principles. This research provides innovative solutions for fraud prevention in financial institutions and plays a vital role in the advancement of knowledge in data science and industry practices. It is expected that the findings of this study will serve as a strategic guide for BPR management in designing more effective fraud detection systems, thereby increasing public trust in financial institutions.

Keywords: *financial transactions, fraud detection, machine learning, rural banks, smote, xgboost*

1. PENDAHULUAN

Tindakan fraud *(fraud)* dalam sektor jasa keuangan, khususnya pada Bank Perekonomian Rakyat (BPR), merupakan masalah yang semakin kompleks dan menjadi perhatian utama dalam menjaga stabilitas sistem keuangan. *Fraud* tidak hanya merugikan pihak perusahaan, tetapi juga menurunkan kepercayaan nasabah, yang pada gilirannya dapat berdampak negatif pada ekonomi secara keseluruhan. Seiring dengan kemajuan teknologi,

p-ISSN: 2775-4227
e-ISSN: 2775-4219

pelaku *fraud* kini dapat dengan mudah memanipulasi data dan laporan keuangan, sehingga tindakan fraud ini sering kali tidak terdeteksi dalam waktu yang lama. Sebagian besar kasus *fraud* melibatkan akses langsung ke sistem informasi internal, yang memungkinkan manipulasi transaksi atau penggelapan dana tanpa terdeteksi.

Salah satu tantangan utama dalam deteksi *fraud* adalah ketidakseimbangan data, di mana transaksi *fraud* jauh lebih sedikit dibandingkan transaksi sah. Ketidakseimbangan ini membuat banyak model deteksi tradisional yang bergantung pada analisis statistik atau aturan berbasis data historis menjadi kurang efektif. Metode tradisional sering kali tidak fleksibel terhadap perubahan pola transaksi yang cepat dan memerlukan keahlian domain yang mendalam. Selain itu, banyak model ini tidak mampu menangani *dataset* yang sangat besar dan kompleks.

Untuk mengatasi masalah ini, *machine learning (ML*) dan *data mining* muncul sebagai solusi yang menjanjikan. Teknik-teknik *machine learning* dapat memproses dataset besar, mengidentifikasi pola-pola tersembunyi dalam data, dan memperbaiki akurasi deteksi seiring dengan bertambahnya data. Salah satu algoritma yang paling efektif untuk deteksi fraud dalam konteks ini adalah algoritma berbasis pohon keputusan yang dapat menangani *dataset* besar dan tidak seimbang dengan efisien.

Algoritma berbasis pohon keputusan seperti *XGBoost* telah terbukti sangat efektif dalam berbagai masalah klasifikasi, termasuk dalam deteksi fraud di sektor keuangan. Algoritma ini menawarkan keunggulan dalam hal performa dan kemampuan untuk mengatasi masalah *overfitting* serta menangani data yang hilang atau tidak seimbang. Keunggulan lainnya adalah kemampuannya dalam menangani berbagai jenis fitur data, baik numerik maupun kategorikal, sehingga dapat diterapkan secara fleksibel pada beragam kasus deteksi *fraud*. Beberapa penelitian telah menunjukkan bahwa algoritma ini memiliki keunggulan dibandingkan dengan metode tradisional dalam mendeteksi transaksi mencurigakan atau anomali yang sulit dikenali [1], [2].

Di sektor perbankan, penggunaan teknik pembelajaran mesin telah banyak diterapkan untuk mendeteksi transaksi *fraud* eksternal, seperti penggunaan kartu kredit palsu atau penggelapan dana oleh pihak luar. Namun, *fraud* yang dilakukan oleh pihak internal, seperti manipulasi laporan keuangan, sering kali lebih sulit terdeteksi. Untuk itu, pendekatan berbasis pembelajaran mesin dapat dikombinasikan dengan teknik lain, seperti *data augmentation*, untuk meningkatkan sensitivitas terhadap pola *fraud* yang jarang terjadi [3].

Salah satu kendala utama dalam deteksi *fraud* adalah ketidakseimbangan data, di mana jumlah transaksi *fraud* sangat jauh lebih sedikit dibandingkan transaksi sah. Masalah ini sering kali menyebabkan model *machine learning* kurang sensitif terhadap transaksi *fraud*, sehingga banyak transaksi yang mencurigakan terlewatkan. Untuk mengatasi masalah ini, teknik pengayaan data seperti *Synthetic Minority Over-sampling Technique (SMOTE)* telah digunakan secara luas [4].

SMOTE adalah teknik yang memungkinkan penciptaan sampel baru untuk kelas minoritas *(fraud)* dengan mensintesis data berdasarkan karakteristik yang ada dalam *dataset* yang ada. Penggunaan *SMOTE* dalam kombinasi dengan algoritma *machine learning* dapat *meningkatkan* akurasi model dalam mendeteksi transaksi *fraud*, terutama pada *dataset* yang tidak seimbang. Dengan menggunakan *SMOTE*, model dapat mengidentifikasi pola transaksi *fraud* yang lebih luas, memperbaiki ketidakseimbangan kelas, dan meningkatkan akurasi keseluruhan deteksi [5].

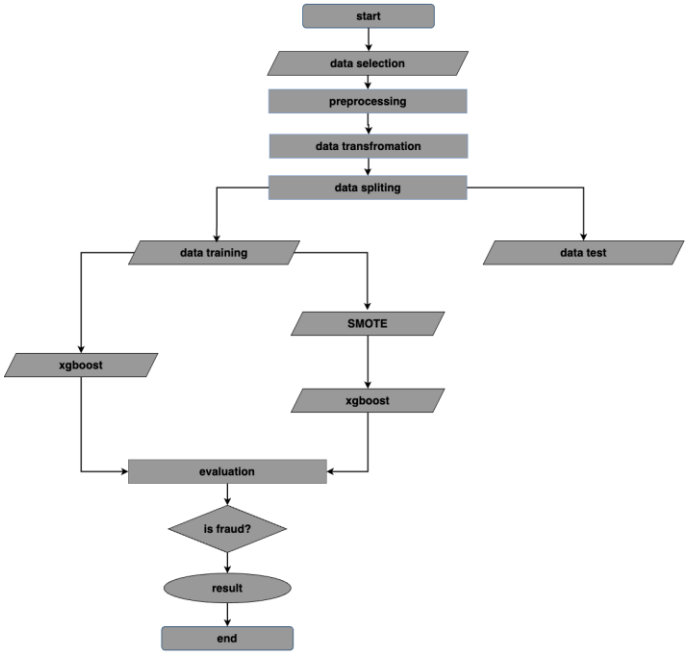
Meskipun berbagai penelitian telah mengaplikasikan algoritma pembelajaran mesin dalam deteksi *fraud*, sebagian besar penelitian lebih terfokus pada deteksi *fraud* eksternal, seperti penggelapan dana oleh pihak luar. Penelitian yang lebih mendalam mengenai deteksi *fraud* internal, yang melibatkan manipulasi laporan keuangan atau penggelapan oleh pihak internal, masih terbatas. BPR menghadapi tantangan khusus di mana *fraud* internal lebih sulit dideteksi karena pelaku memiliki akses langsung ke sistem dan data internal perusahaan.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model deteksi *fraud* yang dapat mengidentifikasi baik transaksi *fraud* eksternal maupun internal di BPR. Penelitian ini juga bertujuan untuk mengevaluasi penerapan sistem pengendalian internal dan prinsip *good corporate governance* yang ada dalam rangka membantu BPR dalam mengidentifikasi dan mencegah transaksi *fraud* dengan lebih efektif, serta memperkuat sistem pengendalian internal yang ada untuk melindungi integritas BPR.

2. METODE PENELITIAN

Penelitian ini menggunakan *dataset* transaksi keuangan yang tersedia secara publik di platform Kaggle. *Dataset* ini mencakup berbagai atribut, termasuk jumlah transaksi, waktu, jenis transaksi, dan informasi nasabah. Salah satu tantangan utama adalah ketidakseimbangan data , di mana transaksi sah mendominasi hingga 96%, sementara hanya 4% transaksi tergolong *fraud*. Metode yang digunakan untuk mencapai tujuan penelitian ini adalah metodologi *Knowledge Discovery* in *Database* (*KDD*), yang mencakup lima tahap, yaitu *Data selection, Preprocessing, Data Transformation, Data mining*[6].

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 1. Tahapan KDD

Data terdiri dari beberapa fitur numerik dan kategorikal. Fitur numerik mencakup nilai transaksi, durasi, dan saldo rekening, sedangkan fitur kategorikal mencakup jenis transaksi dan lokasi. Semua fitur numerik dinormalisasi menggunakan metode *min-max scaling* untuk memastikan keseragaman skala, sedangkan fitur kategorikal diubah menjadi bentuk numerik menggunakan teknik *one-hot encoding* [7].

Pada metode *machine learning*, terdapat beberapa nilai parameter yang diperkirakan dapat meningkatkan kinerja model yang disebut *hyperparameter*. *Hyperparameter* digunakan untuk meningkatkan hasil kinerja algoritma, yang mana cukup mempengaruhi berbagai uji model [8]. Beberapa hyperparameter utama yang disesuaikan dalam penelitian ini meliputi:

Table 1. *Hyperparameter tunning*

Hyperparameter	**Kegunaan *Hyperparameter***
n_estimaors	Banyaknya pohon yang digunakan untuk proses klasifikasi.
max_depth	Tingkat kedalaman maksimum pohon untuk menangkap kompleksitas pola data.
min_child_weight	Bobot minimal untuk memisahkan simpul, mengontrol overfitting dengan mengurangi kompleksitas.
eta (learning_rate)	Membantu mempersingkat langkah dalam pembaruan model, mengontrol kecepatan pembelajaran.
gamma	Meminimumkan pengurangan kerugian untuk menentukan pemisahan simpul pada pohon.
subsample	Rasio instance dari data latih yang digunakan untuk membangun setiap pohon.
colsample_bylevel	Rasio fitur dari data latih yang digunakan untuk membangun setiap level pohon.

2.1. *eXtreme Gradient Boosting (XGBoost)*

Metode *XGBoost* merupakan pengembangan dari *gradient boosting* yang diusulkan oleh Dr. Tianqi Chen dari *University of Washington* pada tahun 2014 [8]. *Gradient boosting* adalah algoritma yang dapat menemukan solusi optimal untuk berbagai masalah, terutama dalam regresi, klasifikasi, dan perankingan. Konsep dasar dari algoritma ini adalah menyesuaikan parameter pembelajaran secara berulang untuk meminimalkan fungsi *loss*, yang berfungsi sebagai mekanisme evaluasi model. *XGBoost* menggunakan model yang lebih teratur untuk membangun struktur pohon regresi, sehingga dapat memberikan kinerja yang lebih baik dan mengurangi kompleksitas model untuk menghindari *overfitting* [9]. Hasil prediksi akhir dari *XGBoost* adalah penjumlahan hasil prediksi dari setiap pohon regresi [10]. *XGBoost* dipilih sebagai algoritma utama karena performanya yang tinggi dalam berbagai tugas klasifikasi, termasuk deteksi fraud. Algoritma ini menggunakan pendekatan boosting untuk membangun model pohon keputusan yang kuat dengan menggabungkan pohon-pohon lemah secara iteratif [3].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Dalam metode ini, diperlukan fungsi objektif yang berguna untuk menilai seberapa baik model yang diperoleh sesuai dengan data latih [11]. Karakteristik terpenting dari fungsi objektif terdiri dari dua bagian, yaitu nilai pelatihan yang hilang dan nilai regularisasi, seperti yang terdapat dalam persamaan berikut ini.

$$obj(\theta) = L(\theta) + \Omega(\theta) \quad (1)$$

Di mana L adalah fungsi pelatihan yang hilang, dan Ω adalah fungsi regularisasi, serta θ adalah parameter model terkait. Fungsi pelatihan yang hilang secara umum dapat dituliskan seperti pada persamaan sebagai berikut.

$$L(\theta) = \sum_{i=1}^{n} l(y_i, x_i) \quad (2)$$

Dimana y_i adalah nilai data sebenarnya yang dianggap benar dan x_i adalah hasil nilai prediksi dari model, sedangkan *n* adalah jumlah iterasi nilai dari model.

2.2. SMOTE

Penerapan teknik *SMOTE (Synthetic Minority Over-sampling Technique)* untuk meningkatkan deteksi fraud dalam industry jasa keuangan, khususnya dalam transaksi perbankan digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset, di mana jumlah transaksi fraud jauh lebih sedikit dibandingkan transaksi sah. Dengan menghasilkan sampel sintetis dari data fraud yang ada, *SMOTE* membantu mendistribusikan kelas lebih seimbang, yang memungkinkan model pembelajaran mesin seperti *XGBoost* untuk lebih efektif mempelajari pola-pola fraud [12]. Ketidakseimbangan kelas diatasi menggunakan metode *Synthetic Minority Over-sampling Technique (SMOTE)*. Teknik ini menghasilkan data sintetis untuk kelas minoritas dengan membuat sampel baru berdasarkan tetangga terdekat dari data minoritas yang ada. *SMOTE* dipilih karena mampu meningkatkan representasi kelas minoritas tanpa menggandakan data yang sama, sehingga memperkaya informasi yang dapat dipelajari oleh model. Penelitian sebelumnya menunjukkan bahwa *SMOTE* efektif dalam meningkatkan performa algoritma pada dataset yang tidak seimbang [12].

Skenario pembagian *dataset* dilakukan untuk membangun dan mengevaluasi model deteksi *fraud* menggunakan algoritma *XGBoost*. Dataset dibagi menjadi beberapa rasio untuk menganalisis kinerja model dalam kondisi yang berbeda [13].

Table 2. Skenario Pembagian *Dataset*

Skenario	Rasio Pembagian Dataset (Training)	Model XGBoost	Model *XGBoost* + SMOTE
1	90:10	Model 1	Model 1 + SMOTE
2	80:20	Model 2	Model 2 + SMOTE
3	70:30	Model 3	Model 3 + SMOTE
4	60:40	Model 4	Model 4 + SMOTE
5	50:50	Model 5	Model 5 + SMOTE

Tahap terakhir pada metodologi *KDD* adalah evaluasi. Pada tahap ini dilakukan perubahan pola menjadi informasi yang mudah dipahami. Tahapan ini untuk mengetahui pola atau informasi yang diperoleh apakah sesuai dengan tujuan sebelumnya untuk mengevaluasi performa algoritma *XGBoost* dengan menggunakan *confusion matrix*, dan melihat nilai *Area Under the Curve* (*AUC)* dari kurva *Receiver Operating Characteristic* (*ROC) [1]*..

3. HASIL DAN PEMBAHASAN

3.1. *Data Selection*

Proses ini dirancang untuk memilih dan mempersiapkan data secara sistematis sebagai langkah awal dalam pengembangan model deteksi *fraud*. Salah satu elemen kunci dalam tahap ini adalah penyusunan deskripsi data, yang mencakup informasi mendetail mengenai jumlah data, jenis data, serta relevansi setiap sumber data yang digunakan. Penjelasan data yang rinci dan terstruktur sangat penting untuk memastikan pemahaman menyeluruh terhadap karakteristik *dataset*, yang nantinya akan mendukung akurasi dan efektivitas model dalam mendeteksi pola transaksi *fraud* [6].

Dalam konteks BPR seperti BPR, kasus transaksi *fraud* sering kali disebabkan oleh penyalahgunaan wewenang oleh karyawan internal atau adanya kerja sama tidak sah dengan pihak eksternal untuk memanipulasi transaksi. Oleh karena itu, pemahaman mendalam tentang data dan konteks kasus sangat diperlukan untuk

p-ISSN: 2775-4227
e-ISSN: 2775-4219

mengidentifikasi potensi kelemahan sistem dan meningkatkan kemampuan model dalam mendeteksi anomali secara efektif.

Deskripsi data yang dilakukan terdiri dari penjelasan jumlah data dan tipe data. *Dataset* berjumlah 6.353.307 *row data* yang memiliki 9 kolom fitur dan 1 kolom target. Fitur-fiturnya meliputi step (mewakili hari dalam simulasi), *Customer* (ID pelanggan), zipCodeOrigin (kode pos asal), *Merchant* (*ID merchant*), *zipMerchant* (kode pos *merchant*), Age (kategori usia), Gender (jenis kelamin pelanggan), *Category* (kategori pembelian), dan *Amount* (jumlah transaksi). Kolom target, *Fraud*, menunjukkan apakah transaksi tersebut bersifat *fraud* (1) atau tidak (0).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 11 columns):
 #   Column          Dtype
---  ------          -----
 0   step            int64
 1   type            object
 2   amount          float64
 3   nameOrig        object
 4   oldbalanceOrg   float64
 5   newbalanceOrig  float64
 6   nameDest        object
 7   oldbalanceDest  float64
 8   newbalanceDest  float64
 9   isFraud         int64
 10  isFlaggedFraud  int64
dtypes: float64(5), int64(3), object(3)
memory usage: 534.0+ MB
```

Gambar 2. Tipe Data Objek

Ada 5 kolom dengan tipe data float64 (numerik dengan desimal), termasuk kolom seperti *amount*, *oldbalanceOrg*, *newbalanceOrig*, *oldbalanceDest*, dan *newbalanceDest*. Kolom-kolom ini kemungkinan berhubungan dengan saldo dan jumlah transaksi.

Ada 3 kolom dengan tipe data *object* (*string* atau teks), yaitu *type*, *nameOrig*, dan *nameDest*. Kolom *type* mungkin berisi jenis transaksi (misalnya transfer, debit, kredit), sedangkan *nameOrig* dan *nameDest* mungkin berisi informasi identitas atau kode akun dari pengirim dan penerima.

Ada 3 kolom dengan tipe data int64 (numerik tanpa desimal), yaitu step, *isFraud*, dan *isFlaggedFraud*. Kolom step mungkin mengindikasikan urutan atau waktu dalam dataset, sedangkan *isFraud* dan *isFlaggedFraud* adalah variabel target biner (0 atau 1) untuk mendeteksi apakah transaksi fraud atau tidak.

Table 3.Transformasi Dataset

step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0
1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0
1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1
1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1
1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0

Setelah dilakukan proses transformasi terhadap *dataset* didapatkan hasil sesuai Tabel 3. Proses transformasi selanjutnya normalisasi data dengan menggunakan metode *RobustScaler*. Normalisasi adalah proses mengubah skala data sehingga semua variabel berada dalam rentang tertentu [14], terlihat bahwa atribut yang bertipe kategorikal mengalami perubahan yang dapat dilihat di Tabel 3.

3.2. *Data Preprocessing*

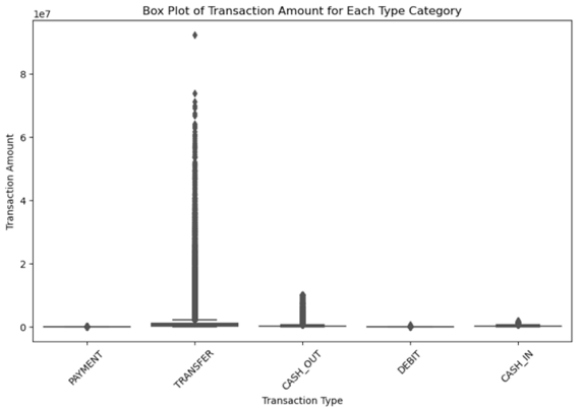
Data yang dikumpulkan kemudian melalui tahap *preprocessing* untuk memastikan kualitas data yang optimal sebelum model dilatih. Proses ini meliputi penanganan *missing values*, normalisasi data, serta transformasi variabel kategorikal. Mengingat *dataset* sering kali tidak seimbang (jumlah kasus *fraud* lebih sedikit dibandingkan transaksi normal), teknik *oversampling* digunakan untuk menangani masalah ini [15] Hal ini bertujuan agar model *XGBoost* dapat mendeteksi transaksi *fraud* yang tersembunyi secara lebih akurat .

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Mengingat *dataset* sering kali tidak seimbang di mana jumlah kasus *fraud* jauh lebih sedikit dibandingkan transaksi normal teknik *oversampling* diterapkan untuk mengatasi masalah ini. Dengan menggunakan metode seperti *SMOTE*, jumlah sampel pada kelas minoritas dapat ditingkatkan, sehingga model *XGBoost* dapat mendeteksi transaksi *fraud* yang tersembunyi dengan lebih akurat. Pendekatan ini tidak hanya meningkatkan akurasi tetapi juga memberikan pemahaman yang lebih baik tentang pola transaksi *fraud* dalam konteks yang lebih luas.

Dalam konteks data sains, *box plot* sering digunakan untuk menganalisis dan memahami karakteristik data sebelum melangkah ke tahap modeling lebih lanjut. *Box plot* dapat memberikan informasi yang berguna mengenai median, kuartil, serta data pencilan, yang penting dalam proses *exploratory data analysis (EDA)* untuk pemahaman awal terhadap dataset.

Outliers dapat mengganggu performa model dalam analisis data, terutama pada algoritma yang sensitif terhadap nilai ekstrem. *Box plot* memudahkan identifikasi *outliers*, yang bisa menjadi indikasi kesalahan data atau justru data penting, seperti dalam kasus deteksi *fraud* [16].



Gambar 3. Pengecekan *Outlayer*

Seperti yang tergambar pada *box plot* menunjukkan distribusi jumlah transaksi untuk setiap kategori tipe transaksi. Dari visualisasi ini, dapat diamati bahwa kategori *transfer* dan *cash_out* memiliki beberapa nilai ekstrim (*outliers*) yang jauh lebih besar dibandingkan kategori lainnya. Hal ini menandakan bahwa untuk beberapa jenis transaksi, terutama *transfer*, terdapat variabilitas yang tinggi dalam jumlah transaksi, yang bisa mengindikasikan adanya potensi *fraud*. Sebaliknya, kategori lain seperti *payment*, *cash_in*, dan *debit* memiliki distribusi yang lebih sempit dengan nilai transaksi yang relatif kecil.

3.3. *Data Transfomation*

Proses transformasi data bertujuan untuk memudahkan proses pemodelan dengan data yang digunakan. Jika dilihat pada Gambar 3 tertera tipe data dari kolom yang berupa kategorikal, sedangkan pemodelan hanya bisa jika data bertipe numerik, maka perlu di transformasi lagi agar data diubah menjadi numerik. Berdasarkan sumber *dataset* yang diperoleh juga menunjukan data belum seimbang untuk dilakukan pemodelan. Tahap transformasi ini juga dilakukan *handling imbalance data [17]*. Langkah pertama yang dilakukan data yang bertipe kategori harus diubah terlebih dahulu. Perubahan tipe menjadi numeric dapat dilihat pada table berikut.

Table 4. Hasil Konversi Numerik

	count	mean	std	min	25%	50%	75%	max	range
step	6362620.0	2.433972e+02	1.423320e+02	1.0	156.00	239.000	3.350000e+02	7.430000e+02	7.420000e+02
amount	6362620.0	1.798619e+05	6.038582e+05	0.0	13389.57	74871.940	2.087215e+05	9.244552e+07	9.244552e+07
oldbalanceOrg	6362620.0	8.338831e+05	2.888243e+06	0.0	0.00	14208.000	1.073152e+05	5.958504e+07	5.958504e+07
newbalanceOrig	6362620.0	8.551137e+05	2.924049e+06	0.0	0.00	0.000	1.442584e+05	4.958504e+07	4.958504e+07
oldbalanceDest	6362620.0	1.100702e+06	3.399180e+06	0.0	0.00	132705.665	9.430367e+05	3.560159e+08	3.560159e+08
newbalanceDest	6362620.0	1.224996e+06	3.674129e+06	0.0	0.00	214661.440	1.111909e+06	3.561793e+08	3.561793e+08
isFraud	6362620.0	1.290820e-03	3.590480e-02	0.0	0.00	0.000	0.000000e+00	1.000000e+00	1.000000e+00
isFlaggedFraud	6362620.0	2.514687e-06	1.585775e-03	0.0	0.00	0.000	0.000000e+00	1.000000e+00	1.000000e+00

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Setelah melalui *preprocessing* data, termasuk penanganan nilai hilang dan transformasi variabel kategorikal, dapat mengelompokkan berbagai jenis transaksi ini secara lebih akurat dalam model *XGBoost*. Kategori yang paling dominan adalah *cashout* dan *transfer*, yang mengindikasikan bahwa dua jenis transaksi ini memiliki frekuensi tertinggi dalam *dataset*. Hal ini penting untuk model deteksi *fraud* karena kategori transaksi tertentu mungkin lebih rentan terhadap penyalahgunaan.

3.4. *Data Mining*

Pada tahap *data mining*, pengujian dilakukan menggunakan *dataset* yang memiliki 9 kolom fitur dan 1 kolom target. *Dataset* ini dibagi menjadi data *training* dan data testing dengan rasio yang bervariasi dalam dua skenario. Skenario 1 menggunakan metode *SMOTE* untuk menangani ketidakseimbangan data, sementara skenario 2 tidak menggunakan teknik tersebut. Pembagian data dilakukan dalam lima skenario:

a. Skenario 1: 90% data training, 10% data testing
b. Skenario 2: 80% data training, 20% data testing
c. Skenario 3: 70% data training, 30% data testing
d. Skenario 4: 60% data training, 40% data testing
e. Skenario 5: 50% data training, 50% data testing

Melalui penerapan lima skenario pengujian yang terencana, proses evaluasi model dapat dioptimalkan untuk mengurangi waktu komputasi secara signifikan. Pendekatan ini memungkinkan fokus pada skenario yang paling relevan, sehingga tetap menjaga efisiensi tanpa mengorbankan kualitas hasil analisis. Namun, perlu dicatat bahwa semakin banyak skenario yang diuji, semakin besar pula kompleksitas proses, yang berujung pada meningkatnya durasi pelatihan dan pengujian model [13]. Dalam pengelolaan *dataset*, pembagian data menjadi *subset* pelatihan dan pengujian dilakukan menggunakan fungsi *train test split* dari library *scikit-learn*. Fungsi ini merupakan alat penting dalam proses analisis, karena memastikan data terdistribusi secara acak dan proporsional sesuai dengan parameter yang telah ditentukan, seperti rasio data pelatihan dan pengujian.

Pendekatan ini dirancang untuk memastikan representasi yang seimbang antara kedua subset, yang sangat penting dalam mencegah bias pada model. Selain itu, *train test split* mendukung fleksibilitas dalam mengatur seed acak (*random_state*), yang memungkinkan reproduksi eksperimen secara konsisten. Dengan demikian, penggunaan fungsi ini tidak hanya berkontribusi pada validitas hasil, tetapi juga menjadi landasan penting dalam membangun model yang dapat diandalkan.

Proses ini merupakan bagian integral dari strategi analisis, karena pembagian *dataset* yang tepat berdampak langsung pada akurasi evaluasi model serta efisiensi waktu komputasi secara keseluruhan.

	Balanced Accuracy	Precision	Recall	F1	Kappa
XGBoost	0.925	0.975	0.851	0.909	0.908

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1016706
1	0.97	0.85	0.91	1314
accuracy			1.00	1018020
macro avg	0.99	0.93	0.95	1018020
weighted avg	1.00	1.00	1.00	1018020

Gambar 4. *Confusion Matrix*

Pada gambar 4 menunjukkan performa yang sangat baik dalam mendeteksi transaksi *fraud* dengan *balanced accuracy* 0.925, *precision* 0.975, *recall* 0.851, dan *F1-score* 0.909. Model ini mampu menangani ketidakseimbangan kelas dan secara akurat mengidentifikasi transaksi *fraud*, terbukti dengan nilai kappa yang tinggi (0.908) serta akurasi hampir sempurna (1.00)..

3.5. *Evaluation*

Pada tahap ini, setiap model yang telah dibangun dievaluasi untuk menentukan model yang paling optimal dan sesuai dengan tujuan penelitian. Evaluasi dilakukan menggunakan metrik yang relevan untuk mengukur performa model dalam memprediksi atau mengklasifikasikan data baru. Tujuan evaluasi ini adalah untuk mengukur seberapa baik setiap model berfungsi, dan melalui perbandingan hasil dari berbagai model, dapat diidentifikasi model dengan performa terbaik. Penilaian dilakukan dengan membandingkan prediksi model

p-ISSN: 2775-4227
e-ISSN: 2775-4219

dengan data sebenarnya dari set uji atau validasi, dan hasil performa tersebut ditampilkan dalam bentuk tabel, seperti yang terlihat pada Tabel 5 dan Tabel 6.

Table 5. Skenario Akurasi SMOTE

Skenario	Perbandingan Data	Akurasi Tanpa *SMOTE*	Akurasi dengan *SMOTE*
1	90:10	92,0%	96,0%
2	80:20	85,0%	93,0%
3	70:30	72,3%	78,5%
4	60:40	68,1%	76,9%
5	50:50	55,4%	63,2%

Tabel 5 menyajikan akurasi model dalam berbagai kondisi dan skenario yang berbeda. Dari data yang ditampilkan, dapat disimpulkan bahwa model yang dibangun tanpa penerapan metode *SMOTE* menunjukkan hasil akurasi yang cukup rendah. Di sisi lain, ketika metode *SMOTE* diterapkan, terlihat adanya peningkatan signifikan dalam nilai akurasi yang diperoleh. Hal ini menunjukkan bahwa penggunaan *SMOTE* dapat membantu meningkatkan performa model dalam mendeteksi pola yang lebih kompleks dalam data.

Table 6. Skenario Nilai *AUC*

Skenario	Perbandingan Data	AUC Tanpa *SMOTE*	AUC dengan *SMOTE*
1	90:10	0,92	0,96
2	80:20	0,85	0,93
3	70:30	0,73	0,78
4	60:40	0,68	0,76
5	50:50	0,55	0,63

Tabel 6 memperlihatkan nilai *AUC* untuk setiap kondisi dan skenario. *AUC* berfungsi untuk menilai kinerja model dalam memisahkan kelas-kelas. Model yang dikembangkan tanpa menggunakan metode *SMOTE* menunjukkan nilai *AUC* yang lebih rendah dibandingkan dengan model yang menerapkan metode *SMOTE*.

Table 7. Kategori Nilai *AUC*

No	Rentang Nilai AUC	Kategori
1	0,90 – 1.00	*Very Good*
2	0.80 – 0,89	*Good*
3	0,70 – 0,79	*Fair*
4	0,60 – 0,69	*Poor*
5	0,00 – 0,59	*Fail*

Berdasarkan evaluasi yang dilakukan, Skenario 1 dan 2 terbukti sangat efektif untuk mendeteksi potensi *fraud* dan disarankan untuk digunakan dalam pengujian lebih lanjut. Skenario 3 juga menunjukkan kinerja yang memadai dan dapat diterapkan, meskipun tidak seefisien skenario sebelumnya. Sementara itu, Skenario 4 dan 5 memiliki keterbatasan dalam kemampuan deteksi, sehingga perlu diperhatikan dan mungkin memerlukan perbaikan sebelum diimplementasikan untuk tujuan deteksi *fraud* yang lebih akurat.

Penelitian ini berhasil menunjukkan bahwa kombinasi model *XGBoost* dengan teknik oversampling *SMOTE* menghasilkan performa yang signifikan dalam mendeteksi transaksi *fraud* pada dataset dengan ketidakseimbangan kelas. Berdasarkan hasil evaluasi, model mencapai nilai *AUC* sebesar 0.96, yang menunjukkan kemampuan tinggi dalam membedakan kelas *fraud* dan *non-fraud.*

Penggunaan teknik *SMOTE* secara konsisten terbukti meningkatkan performa dalam mendeteksi kelas minoritas dengan menyeimbangkan distribusi kelas pada dataset yang tidak seimbang. *SMOTE* membantu model untuk mengenali pola transaksi fraud yang sebelumnya mungkin terabaikan. Kelemahan utama dalam dataset dengan ketidakseimbangan kelas adalah model lebih cenderung mengklasifikasikan sebagian besar transaksi sebagai *non-fraud*, sementara *SMOTE* memanfaatkan teknik oversampling untuk menambah jumlah data pada kelas minoritas, yang memungkinkan model mempelajari pola-pola fraud dengan lebih baik. Penerapan hyperparameter tuning pada *XGBoost*, seperti pada parameter *max_depth*, *learning_rate*, dan *min_child_weight*, juga terbukti krusial dalam meningkatkan kemampuan generalisasi model tanpa mengorbankan akurasi. Parameter-parameter ini, jika diatur dengan tepat, membantu model dalam memaksimalkan prediksi tanpa mengarah pada overfitting, yang berisiko menurunkan kinerja pada data yang tidak terlihat.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Hasil penelitian ini sejalan dengan temuan Zhang et al. [12], yang menunjukkan bahwa *XGBoost* efektif dalam mengatasi ketidakseimbangan kelas pada deteksi fraud. Namun, penelitian ini berinovasi dengan mengintegrasikan *SMOTE* yang jarang diterapkan bersama *XGBoost* dalam konteks dataset keuangan. Penelitian Li et al. [18] juga menunjukkan bahwa *SMOTE* dapat meningkatkan kemampuan prediksi pada kelas minoritas, meskipun mereka tidak mengkombinasikan teknik ini dengan model pohon seperti *XGBoost*. Dengan demikian, penelitian ini memberikan kontribusi baru dalam penerapan kedua teknik secara bersamaan untuk meningkatkan kinerja deteksi *fraud*. Hasil penelitian ini memiliki implikasi signifikan pada dunia nyata, khususnya dalam sektor jasa keuangan.

4. DISKUSI

Penggunaan algoritma *XGBoost* untuk mendeteksi *fraud* menunjukkan variasi kinerja berdasarkan rasio pembagian data dan penggunaan metode *SMOTE*. Metode ini diterapkan untuk menangani ketidakseimbangan antara kelas *fraud* dan *non-fraud*, di mana transaksi *fraud* biasanya jauh lebih sedikit dibandingkan transaksi normal.

Hasil penelitian memperlihatkan bahwa penerapan *SMOTE* secara konsisten meningkatkan nilai akurasi dan *AUC* semua skenario. Dari tabel evaluasi, terlihat bahwa akurasi tanpa *SMOTE* menurun drastis pada rasio data yang lebih seimbang (50:50), dengan akurasi sebesar 55,4%. Namun, setelah *SMOTE* diterapkan, akurasi meningkat hingga 63,2%. Peningkatan ini menunjukkan bahwa *SMOTE* membantu model lebih baik dalam mendeteksi *fraud* ketika data kelas minoritas ditingkatkan.

Dalam hal *AUC*, yang digunakan untuk mengevaluasi kemampuan model membedakan antara kelas *fraud* dan *non-fraud*, penggunaan *SMOTE* juga menghasilkan peningkatan yang signifikan. Pada skenario pembagian data 50:50 tanpa SMOTE, *AUC* hanya mencapai 0,55 (kategori *fail*), namun setelah *SMOTE* diterapkan, nilainya meningkat menjadi 0,63 (kategori *poor*). Hasil ini menggarisbawahi pentingnya teknik *oversampling* untuk memperbaiki kemampuan deteksi model pada data yang tidak seimbang.

Pada rasio data yang lebih tidak seimbang (90:10 dan 80:20), model tanpa *SMOTE* sudah menunjukkan performa yang cukup baik, dengan *AUC* masing-masing 0,92 dan 0,85. Namun, penggunaan *SMOTE* tetap meningkatkan performa model dengan *AUC* yang lebih tinggi, yaitu 0,96 dan 0,93. Ini membuktikan bahwa meskipun *XGBoost* cukup handal dalam menangani ketidakseimbangan data, kombinasi dengan *SMOTE* memberikan hasil yang lebih optimal.

5. KESIMPULAN

Penelitian ini telah menunjukkan bahwa kombinasi algoritma *XGBoost* dan teknik *SMOTE* memberikan solusi yang efektif dalam mendeteksi transaksi *fraud* pada BPR, khususnya pada *dataset* yang tidak seimbang. Dengan mengintegrasikan *SMOTE*, model berhasil menyeimbangkan distribusi kelas yang tidak merata, meningkatkan sensitivitas dalam mendeteksi transaksi *fraud*, yang sering terabaikan pada *dataset* dengan ketidakseimbangan ekstrem. Hasil penelitian ini menekankan pentingnya penerapan metode ini dalam sistem deteksi *fraud* yang lebih akurat dan efisien, memungkinkan BPR untuk lebih responsif dalam mengidentifikasi dan mencegah transaksi *fraud*.

Penelitian ini memberikan implikasi yang signifikan dalam konteks industri keuangan, terutama dalam pencegahan *fraud*, termasuk BPR yang sering kali menghadapi tantangan dalam mengawasi transaksi dengan volume besar dan ketidakseimbangan data, dapat memanfaatkan pendekatan ini untuk meningkatkan efektivitas sistem deteksi *fraud* mereka. Dengan XGBoost dan SMOTE, model dapat mengidentifikasi pola fraud yang lebih kompleks meskipun dengan data yang tidak seimbang, yang sering menjadi kendala dalam penerapan model deteksi *fraud* tradisional. Untuk memastikan keberhasilan penerapan model ini, BPR dapat mengikuti langkah-langkah berikut:

1. BPR sebaiknya mengintegrasikan model *XGBoost* dengan *SMOTE* dalam sistem deteksi transaksi untuk memperbaiki akurasi dan sensitivitas dalam mendeteksi transaksi fraud.
2. Pastikan data yang digunakan untuk pelatihan model bersih, bebas dari duplikasi, serta terstruktur dengan baik. Penggunaan teknik seperti normalisasi dan *encoding* variabel kategorikal akan meningkatkan kinerja model.
3. Evaluasi performa model secara rutin menggunakan metrik yang relevan, seperti *AUC*, untuk memastikan model tetap efektif dalam mendeteksi transaksi *fraud*, khususnya di tengah variasi pola fraud yang berkembang.
4. Melakukan pelatihan bagi staf keuangan untuk memahami cara kerja model deteksi *fraud*, serta bagaimana mereka dapat mengoptimalkan sistem ini dalam pengawasan transaksi dan upaya pencegahan fraud.

Dengan mengimplementasikan pendekatan berbasis *XGBoost* dan *SMOTE*, BPR dapat memperkuat sistem deteksi *fraud* mereka, meningkatkan keakuratan dalam identifikasi transaksi mencurigakan, dan meminimalkan

p-ISSN: 2775-4227
e-ISSN: 2775-4219

potensi kerugian finansial akibat *fraud*. Penelitian ini juga membuka peluang untuk eksplorasi lebih lanjut dalam pengembangan strategi lain yang lebih responsif terhadap tantangan *fraud* yang berkembang pesat di dunia keuangan. Sebagai langkah lanjutan, penelitian ini dapat dijadikan dasar untuk pengembangan model yang lebih canggih, seperti penggunaan teknik *ensemble* atau algoritma lain yang lebih adaptif terhadap perubahan dalam pola *fraud*.

DAFTAR PUSTAKA

[1] A. Q. Abdulghani, O. N. UCAN, and K. M. A. Alheeti, "Credit Card Fraud Detection Using XGBoost Algorithm," in *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, Dec. 2021, pp. 487–492. doi: 10.1109/DeSE54285.2021.9719580.

[2] U. Detthamrong, W. Chansanam, T. Boongoen, and N. Iam-On, "Enhancing Fraud Detection in Banking using Advanced Machine Learning Techniques," *International Journal of Economics and Financial Issues*, vol. 14, no. 5, pp. 177–184, Sep. 2024, doi: 10.32479/ijefi.16613.

[3] T. R. Noviandy, G. M. Idroes, A. Maulana, I. Hardi, E. S. Ringga, and R. Idroes, "Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques," *Indatu Journal of Management and Accounting*, vol. 1, no. 1, pp. 29–35, Sep. 2023, doi: 10.60084/ijma.v1i1.78.

[4] P. Gupta, S. Shukla, V. Kikan, and A. Kumar, "Enhancing Fraud Detection in Credit Card Transactions using XGBoost and SMOTE: A Comparative Study," in *2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)*, 2024, pp. 1–6. doi: 10.1109/ICSPCRE62303.2024.10675080.

[5] K. Garg, K. S. Gill, S. Malhotra, S. Devliyal, and G. Sunil, "Implementing the XGBOOST Classifier for Bankruptcy Detection and Smote Analysis for Balancing Its Data," in *2024 2nd International Conference on Computer, Communication and Control (IC4)*, 2024, pp. 1–5. doi: 10.1109/IC457434.2024.10486274.

[6] S. MP, M. Pandey, Garima, J. V, V. B. Maral, and S. Das, "Investigating the Use of Data Mining for Knowledge Discovery," in *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 2023, pp. 1–7. doi: 10.1109/SMARTGENCON60755.2023.10442608.

[7] Y. Sei, J. A. Onesimu, and A. Ohsuga, "Machine Learning Model Generation With Copula-Based Synthetic Dataset for Local Differentially Private Numerical Data," *IEEE Access*, vol. 10, pp. 101656–101671, 2022, doi: 10.1109/ACCESS.2022.3208715.

[8] E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *JOMTA Journal of Mathematics: Theory and Applications*, vol. 4, no. 1, 2022.

[9] A. K. Saw, P. Luthra, and D. Thakur, "Optimizing Credit Card Fraud Detection: Random Forest and XGBoost Ensemble," in *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, 2024, pp. 1–6. doi: 10.1109/ACROSET62108.2024.10743811.

[10] S. Li and X. Zhang, "Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm," *Neural Comput Appl*, vol. 32, Apr. 2020, doi: 10.1007/s00521-019-04378-4.

[11] I. Hanif, "Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction," European Alliance for Innovation n.o., Jan. 2020. doi: 10.4108/eai.2-8-2019.2290338.

[12] H. Xia, W. An, and Zuopeng (Justin) Zhang, "Credit Risk Models for Financial Fraud Detection," *Journal of Database Management*, vol. 34, no. 1, pp. 1–20, Apr. 2023, doi: 10.4018/jdm.321739.

[13] A. A. Saputra, B. N. Sari, C. Rozikin, U. Singaperbangsa, and K. Abstrak, "Penerapan Algoritma Extreme Gradient Boosting (Xgboost) Untuk Analisis Risiko Kredit," *Jurnal Ilmiah Wahana Pendidikan*, vol. 10, no. 7, pp. 27–36, 2024, doi: 10.5281/zenodo.10960080.

[14] S. T. Mhlanga and M. Lall, "Influence of Normalization Techniques on Multi-criteria Decision-making Methods," *J Phys Conf Ser*, vol. 2224, no. 1, p. 12076, Apr. 2022, doi: 10.1088/1742-6596/2224/1/012076.

[15] H. G. W. Ansari and Dr. M. D. Patil, "Enhancing Credit Card Fraud Detection Using P-XGBoost: A Comparative Study Classical Machine Learning Techniques," *Int J Res Appl Sci Eng Technol*, vol. 11,

p-ISSN: 2775-4227
e-ISSN: 2775-4219

no. 9, pp. 618–623, Sep. 2023, doi: 10.22214/ijraset.2023.55698.

[16] Q. Zhang, "Financial Data Anomaly Detection Method Based on Decision Tree and Random Forest Algorithm," *Journal of Mathematics*, vol. 2022, pp. 1–10, Apr. 2022, doi: 10.1155/2022/9135117.

[17] A. Al Ali, A. M. Khedr, M. El-Bannany, and S. Kanakkayil, "A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique," *Applied Sciences (Switzerland)*, vol. 13, no. 4, Feb. 2023, doi: 10.3390/app13042272.

[18] X. Li, "Financial Fraud: Identifying Corporate Tax Report Fraud Under the Xgboost Algorithm," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 4, pp. 1–7, 2023, doi: 10.4108/eetsis.v10i3.3033.