Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

Analisis Kinerja Algoritma Machine Learning Untuk Klasifikasi Potensi FRAUD Klaim Layanan Kesehatan Rumah Sakit

Imanullah Ali Ubed, Iwan Syarif, Ferry Astika Saputra*

Magister Terapan, Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia
Email: [1]imanullah@pasca.student.pens.ac.id, [2]iwanarif@pens.ac.id, [3,*]ferryas@pens.ac.id
Email Penulis Korespondensi: ferryas@pens.ac.id
Submitted: **14/05/2025**; Accepted: **04/06/2025**; Published**: 05/06/2025**

Abstrak−Fraud dalam klaim layanan kesehatan merupakan tantangan serius yang mengancam efisiensi dan keberlanjutan sistem Jaminan Kesehatan Nasional (JKN) di Indonesia. Penelitian ini berkontribusi dengan mengusulkan pendekatan komparatif berskala besar menggunakan lima algoritma machine learning untuk klasifikasi potensi fraud pada klaim BPJS Kesehatan, yaitu: Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), XGBoost + SMOTE, dan Logistic Regression (LR). Pendekatan ini juga mengeksplorasi penggunaan SMOTE dalam kombinasi dengan XGBoost untuk menangani ketidakseimbangan data yang umum terjadi pada dataset fraud. Dataset yang digunakan terdiri dari lebih dari 200.000 entri klaim yang telah melalui proses pembersihan data, normalisasi, dan seleksi fitur. Evaluasi dilakukan menggunakan metrik presisi, recall pada kelas fraud (positif), f1-score, akurasi, serta visualisasi confusion matrix untuk menilai distribusi kesalahan klasifikasi. Hasil menunjukkan bahwa ANN dan XGBoost + SMOTE memiliki keunggulan dalam mendeteksi klaim fraud dengan recall yang tinggi, sementara SVM menawarkan performa paling seimbang antara presisi dan sensitivitas. Random Forest dan Logistic Regression memberikan baseline performa yang moderat namun kurang optimal untuk mendeteksi pola fraud yang kompleks. Penelitian ini memberikan kontribusi pada pengembangan sistem deteksi fraud berbasis machine learning yang lebih adaptif dan efisien dalam konteks klaim JKN. Implikasi praktis dari studi ini diharapkan dapat memperkuat sistem verifikasi otomatis yang diterapkan oleh BPJS Kesehatan..

Kata Kunci: Fraud; Klasifikasi; Klaim Asuransi; Klaim BPJS Kesehatan; Machine Learning

Abstract-Fraud in healthcare claims represents a critical challenge that undermines the efficiency and sustainability of Indonesia's National Health Insurance (JKN) system. This study contributes a large-scale comparative evaluation of five machine learning algorithms for classifying potential fraud in BPJS Kesehatan claims, namely Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), XGBoost + SMOTE, and Logistic Regression (LR). A novelty of this study lies in applying the SMOTE technique in conjunction with XGBoost to address class imbalance in fraud datasets. The dataset consists of over 200,000 claim entries, which have undergone data cleaning, normalization, and feature selection. Performance was assessed using precision, recall on fraud class (positive), f1-score, accuracy, and confusion matrix visualizations to capture classification error distribution. Results demonstrate that ANN and XGBoost + SMOTE are superior in detecting fraudulent claims with high recall, while SVM achieves the most balanced performance in terms of precision and sensitivity. Random Forest and Logistic Regression serve as moderate baselines but are less effective in identifying complex fraud patterns. This study contributes to the development of a more adaptive and efficient fraud detection system based on machine learning, with practical implications for strengthening the automatic verification system used by BPJS Kesehatan.

Keywords: BPJS Kesehatan Claims; Classification; Claim Health Insurance; Fraud; Machine Learning

1. PENDAHULUAN

Program Jaminan Kesehatan Nasional (JKN) yang diselenggarakan oleh BPJS Kesehatan merupakan upaya pemerintah dalam memberikan perlindungan kesehatan menyeluruh bagi seluruh rakyat Indonesia. Sejak diluncurkan pada tahun 2014, JKN telah berkembang menjadi sistem jaminan sosial berskala nasional yang melibatkan berbagai pemangku kepentingan, termasuk fasilitas kesehatan tingkat pertama (FKTP), fasilitas kesehatan rujukan tingkat lanjutan (FKRTL), serta masyarakat umum sebagai peserta. Namun, seiring dengan meningkatnya cakupan layanan dan jumlah peserta, tantangan terhadap efektivitas dan efisiensi sistem JKN pun semakin kompleks. Salah satu tantangan utama yang menjadi perhatian adalah potensi kecurangan (fraud) dalam proses pengajuan klaim layanan kesehatan yang dapat membebani keuangan negara, mengancam keberlanjutan JKN dan merusak kredibilitas sistem jaminan sosial itu sendiri [1].

Fraud dalam layanan kesehatan didefinisikan sebagai tindakan manipulatif yang disengaja untuk memperoleh pembayaran atau penggantian biaya layanan secara tidak sah. Dalam konteks BPJS Kesehatan, praktik fraud sering kali mencakup phantom billing (penagihan layanan fiktif), pemberian obat yang tidak dibutuhkan, pengkodean diagnosis yang tidak sesuai, serta klaim ganda atas layanan yang sama. Studi oleh Prof. Laksono Trisnantoro dari Universitas Gadjah Mada menegaskan bahwa Tantangan dan Inovasi dalam Pembiayaan Kesehatan di Indonesia, tetapi juga membahayakan keberlanjutan program JKN secara keseluruhan karena menyebabkan ketidakseimbangan antara iuran yang diterima dan klaim yang dibayarkan [2].

Berdasarkan data yang dirilis oleh Komisi Pemberantasan Korupsi (KPK), terdapat lebih dari 175.000 kasus indikasi fraud yang terjadi dalam sistem klaim BPJS Kesehatan sejak program ini berjalan. Estimasi kerugian negara akibat fraud ini mencapai ratusan miliar rupiah, pada tahun 2018 defisit mencapai Rp 11,69 trilliun [3]. Sayangnya, deteksi fraud selama ini masih bergantung pada audit manual bebasis pada rekam medis rumah sakit yang berpotensi *upcoding*, dan pemeriksaan administratif yang memiliki keterbatasan baik dalam hal sumber daya manusia maupun

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

efektivitas proses [4]. Di tengah beban kerja yang tinggi, verifikator manual sering kali luput mengidentifikasi pola-pola kecurangan yang terselubung dalam data klaim yang sangat besar dan kompleks .

Sebagai respons terhadap tantangan tersebut, pendekatan berbasis teknologi informasi dan kecerdasan buatan (Artificial Intelligence) mulai dilirik sebagai solusi yang dapat meningkatkan efektivitas sistem deteksi fraud. Salah satu pendekatan yang menjanjikan adalah penerapan algoritma machine learning, yang dirancang untuk mengenali pola data yang kompleks dan non-linear. Di antara algoritma yang banyak digunakan adalah Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), dan Logistic Regression (LR).

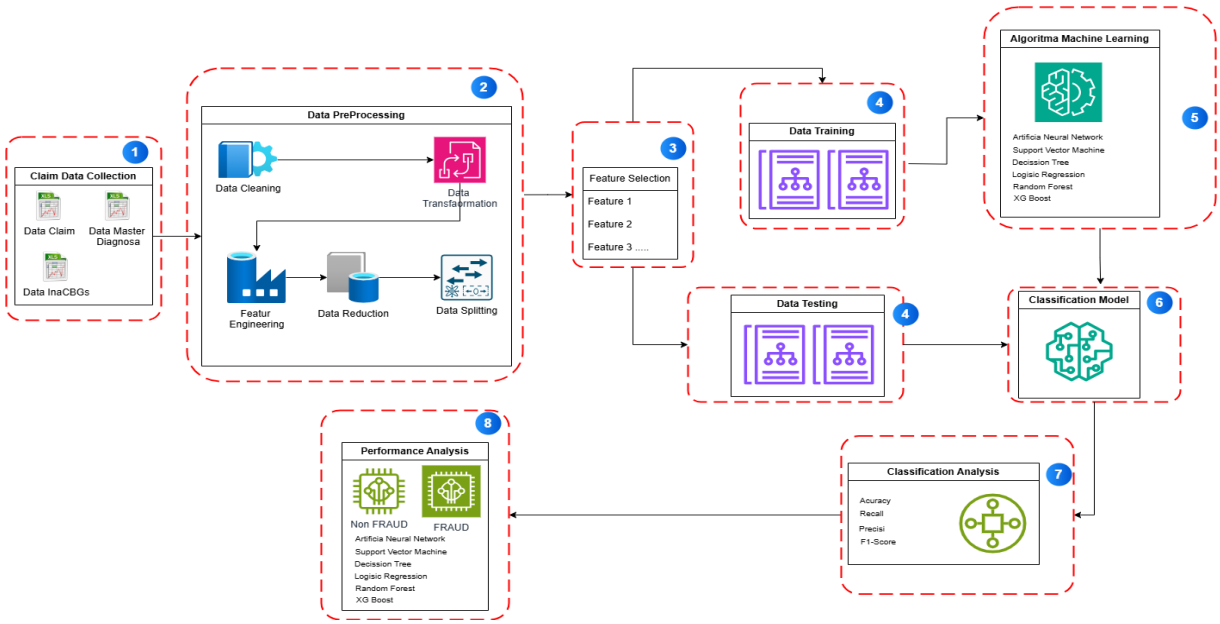
Sejumlah penelitian terdahulu telah mengkaji penerapan algoritma machine learning untuk mendeteksi fraud pada klaim layanan kesehatan. Simanjuntak (2021) menunjukkan bahwa model ANN mampu mendeteksi fraud dengan akurasi tinggi pada data klaim rumah sakit [5]. Nugraha dan Irawan (2023) melakukan komparasi antara SVM dan XGBoost dan menemukan bahwa XGBoost unggul dalam beberapa skenario dengan data yang tidak seimbang [6]. Selain itu, Nabrawi dan Alanazi (2023) menunjukkan efektivitas penggunaan machine learning dalam mendeteksi fraud asuransi kesehatan, termasuk dengan pendekatan penyeimbangan data menggunakan SMOTE [7]. Meskipun demikian, di Indonesia, pemanfaatan algoritma seperti XGBoost + SMOTE dan evaluasi komparatif berskala besar masih sangat terbatas.

Penelitian ini menjembatani kesenjangan tersebut dengan mengembangkan dan menguji model machine learning untuk klasifikasi potensi fraud dalam klaim layanan kesehatan menggunakan pendekatan evaluasi komparatif terhadap lima algoritma: ANN, SVM, Random Forest, XGBoost + SMOTE, dan Logistic Regression. Penelitian ini menggunakan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) sebagai metodologi dalam membangun dan mengevaluasi model [8]. Setiap tahapan, mulai dari pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, hingga implementasi dirancang untuk memastikan proses penelitian yang sistematis dan dapat direplikasi.

Dengan demikian, latar belakang masalah yang telah diuraikan menunjukkan urgensi dan relevansi penelitian ini. Dalam jangka panjang, penerapan teknologi machine learning terutama Artificial Neural Network dan XGBoost + SMOTE yang diharapkan dapat menjadi fondasi bagi sistem deteksi fraud yang lebih canggih dan terintegrasi dengan sistem informasi klaim BPJS Kesehatan, sehingga mampu memperkuat akuntabilitas serta keberlanjutan program JKN secara nasional..

2. METODOLOGI PENELITIAN

Tahapan metodologi penelitian mengadopsi metodologi CRISP-DM [8], [9], yang terlebih dahulu mengenali alur proses bisnis rumah sakit dalam pengolahan data hingga penentuan model algoritma. Gambar 1 menjelaskan tahapan yang dilakukan dalam penelitian ini.



Gambar 1. Alur metodologi penelitian

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari dataset klaim BPJS Kesehatan yang dirilis secara terbuka pada ajang BPJS Hackathon 2022. Data ini diakses secara daring pada September 2023 dan terdiri dari lebih dari 200.000

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

entri klaim rumah sakit yang mencakup berbagai atribut diagnosis, prosedur, demografi, dan kategori layanan. Dataset bersifat seimbang dengan jumlah klaim fraud dan non-fraud yang hampir sama.

2.2 Praproses Data

Praproses data merupakan proses yang sangat penting dalam penerapan supervised learning [10] [11], karena data harus dipastikan konsisten dan seimbang untuk bisa dioptimalkan dalam model machine learning. Pada penelitian ini dilakukan bebera tahapan praproses data.

2.2.1 Pembersihan Data

Tahap pertama ini mencakup identifikasi dan penanganan terhadap nilai yang hilang (*missing value*), data duplikat, serta nilai yang tidak wajar (*outlier*). Proses pembersihan data dilakukan untuk meningkatkan kualitas data, akurasi data dan konsistensi data. Menurut Rofiq pembersihan data dilakukan untuk membuang data *null* dan memperbaiki struktur data yang tidak sesuai dengan standar layanan klaim BPJS [12].

2.2.2. Transformasi dan Normalisasi Data

Transformasi data mencakup perubahan format atribut agar sesuai dengan input model, termasuk konversi format tanggal, penggabungan kolom relevan, serta normalisasi fitur numerik seperti biaya tindakan dan lama rawat inap. Teknik normalisasi seperti Min-Max Scaling atau Z-Score digunakan untuk menyamakan skala antar fitur. Dalam buku pengenalan data mining menyebutkan bahwa normalisasi penting untuk mencegah dominasi fitur tertentu dalam proses pembelajaran, terutama ketika data memiliki distribusi yang sangat beragam [13]. Dan penjelasan pada tabel 2 persiapan data.

Tabel 1. Persiapan Data

Tahap	Metode	Detail
Persiapan Data	Seleksi Fitur yang tidak mempengaruhi SelectKBest	'dx2_koo_k93', 'dx2_u00_u99', 'procv00_v89', 'dati2', 'visit_id', 'kdkc', dan 'typeppk' K= 20
	Transformasi Data	One hot encoding pada atribut jenis kelamin, cmg , typeppk, dan diagprimer. Standardisasi data dengan StandarScaler Data binning pada atribut umur, LoS (Length of Stay)
	Mengurangi dimensi	Principal Component Analisis dengan 95% Variance (90 komponen)
	Pembagian Data	Training=80%, Testing=20% Training dengan 10 Fold Cross Vaidation

Tabel 1 menyajikan tahapan utama dalam proses persiapan data yang digunakan untuk pelatihan model klasifikasi fraud. Tahapan dimulai dari seleksi fitur, transformasi data, hingga proses pembagian data untuk pelatihan dan pengujian model.

a. Seleksi fitur dilakukan untuk menghapus atribut yang tidak memberikan kontribusi signifikan terhadap hasil klasifikasi seperti, visit_id, typeppk, dan dati2, serta dilakukan pemilihan fitur terbaik dengan SelectKBest sebanyak 20 fitur.
b. Transformasi data mencakup penggunaan One-Hot Encoding untuk variabel kategorikal (jenis kelamin, tipe PPK, dll), normalisasi menggunakan StandardScaler, dan teknik data binning untuk variabel seperti umur dan LOS (Length of Stay).
c. Pengurangan dimensi menggunakan Principal Component Analysis (PCA) dengan target 95% variansi untuk menghasilkan 90 komponen utama.
d. Pembagian data dilakukan secara sistematis dengan proporsi 80% untuk data latih dan 20% untuk data uji, serta validasi menggunakan 10-Fold Cross Validation guna menjaga konsistensi hasil evaluasi model.

2.3 Seleksi Fitur

Seleksi fitur dilakukan untuk mengidentifikasi atribut yang paling berkontribusi dalam klasifikasi potensi fraud pada klaim layanan kesehatan [14]. Dalam penelitian ini, dilakukan feature selection menggunakan random feature importance, yaitu metode yang mengukur tingkat kontribusi setiap fitur terhadap prediksi model dengan cara mengacak nilai fitur tertentu dan mengevaluasi dampaknya terhadap performa model. Jika pengacakan fitur menyebabkan penurunan akurasi yang signifikan, maka fitur tersebut memiliki pengaruh yang besar dalam klasifikasi.

2.4 Pembagian Data

Setelah proses seleksi dan transformasi fitur, dataset dibagi menjadi dua bagian utama, yaitu data latih (Training Set) sebesar 80% dari total data, yang digunakan untuk melatih model. Data uji (Testing Set) sebesar 20%, yang difungsikan sebagai alat mengevaluasi performa model dalam mengklasifikasikan klaim yang belum pernah dilihat sebelumnya

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

2.5 Model Algoritma

2.5.1 Artificial Neural Network

Artificial Neural Network (ANN), merupakan model yang cara kerjanya meniru kinerja jaringan syaraf otak manusia, ANN dinilai efektif dalam mendeteksi pola-pola non-linear kompleks yang umum terjadi dalam klaim yang berpotensi fraud, seperti manipulasi kode prosedur, kode diagnosa dan anomali tarif tindakan.

ANN adalah arsitektur supervised learning yang menggunakan neuron untuk mentransformasikan informasi yang diperoleh dari input data ke neuron terhubung lainnya. Arsitektur ANN merupakan Fully Connected Network (FCN) yang terdiri atas tiga layer, yaitu input layer, hidden layer, dan output layer [5].

2.5.2 Support Vector Machine

SVM bekerja dengan mencari hyperplane terbaik yang memisahkan data menjadi dua kelas. Algoritma ini unggul dalam menangani data berdimensi tinggi dan sangat cocok ketika jumlah fitur jauh lebih besar daripada jumlah sampel. SVM juga tangguh terhadap outlier dan *overfitting* . Nugraha dan Irawan (2023) menunjukkan bahwa SVM mampu mencapai performa tinggi dalam mendeteksi klaim fraud asuransi kesehatan, dengan presisi dan recall masing-masing di atas 90% [6]. Karim menjelaskan bahwa konsep SVM menitikberatkan pada risk minimization, yaitu estimasi fungsi dengan cara meminimalisir batas dari *generalization error*, sehingga SVM mampu mengatasi overfitting [15]. Adapun, fungsi regresi dari metode SVM adalah sebagai berikut.

$$f(x) = W^T \varphi(x) + b \quad (1)$$

2.5.3 Random Forest

Random Forest adalah algoritma ensemble learning yang terdiri dari banyak pohon keputusan (decision trees) dan menggabungkan prediksi masing-masing pohon untuk menghasilkan keputusan akhir [16]. Kelebihan dari Random Forest adalah kemampuannya mengurangi overfitting dan menangani dataset besar dengan banyak fitur, baik kategorikal maupun numerik.

2.5.4 Extreme Gradient Boosting

XGBoost adalah algoritma boosting yang membangun pohon keputusan secara iteratif, di mana setiap pohon baru dibentuk untuk memperbaiki kesalahan dari pohon sebelumnya. Algoritma ini sangat efisien, akurat, dan mampu menangani data dengan noise. Dalam studi Nugraha dan Irawan (2023), XGBoost menunjukkan performa superior dibanding SVM dalam beberapa skenario dataset fraud, dengan nilai F1-score mencapai 0.89 [6]. XGBoost juga dikenal dengan kemampuannya dalam menangani missing values dan memiliki fitur regularisasi yang membantu mencegah overfitting. Namun, tuning parameter yang kompleks bisa menjadi tantangan bagi pengguna pemula [17].

2.5.5 Logistic Regression

Logistic regression merupakan algoritma klasifikasi machine learning untuk klafisifikasi probabilitas variabel dependen katagoris. Logistik regression merupakan metode statistik yang merupakan bagian dari analisis regresi yang biasanya digunakan untuk memprediksi kelas biner[18] . Metode ini merupakan metode regresi linear umum untuk mempelajari pemetaan dari sejumlah variabel numerik ke variabel biner atau probabilistic [19]. Metode ini digunakan saat variabel predictor(y) memiliki skala kategorik atau nominal yang terdiri dari dua (biner) atau lebih kategori. Sehingga metode ini dibuat untuk memastikan bahwa, apa pun perkiraan yang terjadi, selalu berada di antara 0 dan 1 [20], fungsi logis pada regresi logistik ditunjukan melalui persamaan (2)

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

2.6 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa prediksi terhadap klasifikasi klaim fraud dan non-fraud [21]. Evaluasi dilakukan berdasarkan metrik-metrik berikut:

a. *Accuracy* : proporsi prediksi yang benar terhadap keseluruhan data.
b. *Precision* : proporsi klaim yang diprediksi fraud dan benar-benar fraud.
c. *Recall (Sensitivity)* : proporsi klaim fraud yang berhasil dikenali dengan benar.
d. *F1-Score* : rata-rata harmonis dari precision dan recall.
e. AUC (*Area Under the Curve*) : luas area di bawah kurva ROC, yang mengukur kemampuan model dalam membedakan kelas.

Evaluasi dilakukan secara konsisten untuk seluruh model agar dapat membandingkan performa antar algoritma secara objektif. Hasil evaluasi ini kemudian divisualisasikan dalam bentuk confusion matrix untuk melihat kesalahan klasifikasi secara lebih detail.

Dengan pendekatan evaluasi yang komprehensif ini, diharapkan hasil penelitian dapat memberikan rekomendasi akurat mengenai pemilihan algoritma machine learning yang paling sesuai untuk mendeteksi potensi fraud dalam klaim layanan kesehatan di Indonesia

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Dataset

Data yang digunakan dalam penelitian ini diperoleh dari data klaim BPJS Kesehatan. Yang merupakan data klaim fasilitas Kesehatan yang juga data public dalam format .csv, yang telah dibagi menjadi dua kategori, kategori pertama merupakan data dalam klasifikasi fraud (1) dengan jumlah 100.255 data dan kategori kedua non fraud (0) dengan jumlah 99.962 data. Dalam data ini berjumlah 200.217 data dan memiliki 53 attribut dan dalam kondisi seimbang. Informasi data dijelaskan dalam Tabel 2 berikut:

Tabel 2. Deskripsi Data

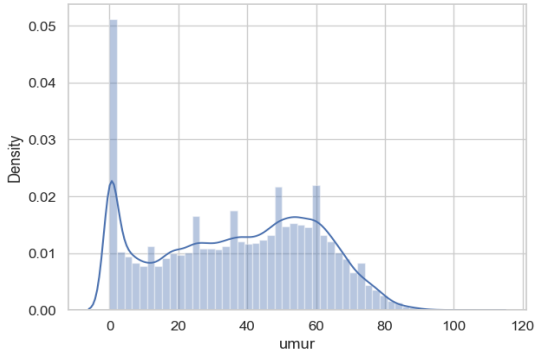
No	Atribut	Keterangan	tipe	nilai
1	visit_id	id kunjungan	int64	Id numerik
2	kdkc	kode wilayah kantor cabang BPJS Kesehatan	int64	Kode numerik
3	dati2	kode kabupaten/kota	int64	Kode numerik
4	typeppk	kode tipe Rumah Sakit	object	SC,C,B,SD,SB,A,D,I3,KM,KI,I2,I4,KJ,KL,I1,KB, KC,GD,SA,KP,KO,KG,HD,KT,KU
5	jkpst	jenis kelamin peserta JKN-KIS	object	P (Peremepuan), L (Laki-laki)
6	umur	umur peserta saat mendapatkan pelayanan rumah sakit	int64	0 - 109
7	jnspelsep	tingkat pelayanan	int64	1: rawat inap; 2: rawat jalan
8	los	lama peserta dirawat di rumah sakit	int64	0 - 255
9	cmg	klasifikasi CMG (Case Mix Group)	object	'F','E','Q','L','H','W','P','U','K','G','M','N','A','C', 'D', 'Z', 'J', 'O', 'S', 'I', 'V', 'T', 'B'
10	severitylevel	tingkat urgensi	int64	0, 3, 2, 1
11	diagprimer	diagnosa primer	object	'f00_f99', 'e00_e90', 'r00_r99', 'j00_j99', 's00_t98', 'h00_h59','m00_m99','c00_d48','z00_z99','p00_p96 ' , 'h60_h95', 'k00_k93','g00_g99', 'i00_i99', 'l00_l99', 'a00_b99', 'n00_n99', 'o00_o99','d50_d89', 'q00_q99', 'u00_u85'
12	dx2_..._...	kode kelompok procedure	int64	terdapat 22 atribut dimana nilainya adalah 0 - 13
13	proc.._...	kode kelompok procedure	int64	terdapat 19 atribut dimana nilainya adalah 0 - 23
14	label	flag fraud	int64	1:fraud; 0:tidak fraud

3.2 Hasil Praproses Data

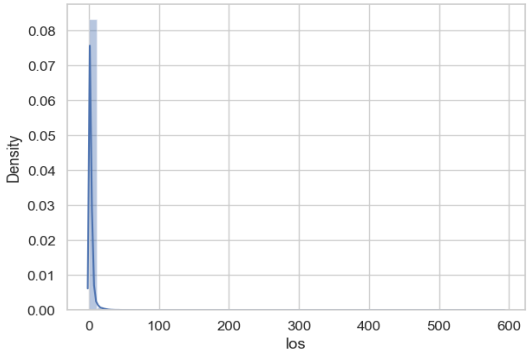
Data yang didapat selanjutnya diproses pada tahapan praproses data dengan tahapan-tahapan yang dipilih untuk meningkatkan kualitas data yang akan diproses dalam Pembangunan model yang diinginkan.

a. Eksplorasi Data

Eksplorasi data dilakukan untuk memahami karakteristik dan distribusi fitur-fitur numerik serta yang berkaitan dengan target klasifikasi indikasi fraud pada klaim layanan kesehatan. Pada Gambar 2 dan Gambar 3.



Gambar 2. Visualisasi distribusi umur



Gambar 3. Visualisasi Distribusi Los

Visualisasi ditribusi umur dan los menunjukan distribusi umur menunjukkan pola yang tidak simetris (non-normal), dengan beberapa lonjakan (*peaks*) pada kelompok usia tertentu. Terlihat bahwa terdapat puncak tajam

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

pada usia mendekati nol, yang kemungkinan besar merepresentasikan klaim untuk pasien bayi atau anak usia dini. Di samping itu, terdapat peningkatan pada rentang usia dewasa muda hingga paruh baya, dan distribusi menurun secara bertahap pada kelompok usia lanjut. Hal ini menunjukkan bahwa klaim layanan kesehatan tersebar di berbagai kelompok usia, namun terdapat dominasi pada usia produktif. Sementara itu, distribusi los sangat condong ke kiri (*positively skewed*), dengan sebagian besar nilai berada di kisaran 0 hingga 10 hari. Puncak distribusi terjadi pada masa rawat inap yang sangat singkat, bahkan mendekati nol hari. Sementara itu, terdapat sejumlah kasus dengan masa rawat inap yang sangat panjang (di atas 100 hari), namun jumlahnya sangat sedikit dan dapat dikategorikan sebagai *outlier*. Bentuk distribusi ini mencerminkan karakteristik umum pelayanan kesehatan, di mana sebagian besar pasien hanya memerlukan perawatan jangka pendek, dan hanya sebagian kecil yang menjalani rawat inap jangka panjang karena kondisi tertentu.

Bentuk distribusi kedua variabel ini penting untuk diperhatikan dalam proses pra-pemrosesan data. Distribusi yang miring dan adanya outlier dapat memengaruhi performa model pembelajaran mesin, terutama model yang sensitif terhadap skala dan sebaran data seperti regresi logistik atau Support Vector Machine. Oleh karena itu, normalisasi atau transformasi seperti log-scaling, serta penanganan outlier, dapat dipertimbangkan untuk meningkatkan akurasi dan kestabilan model.

b. Pembersihan Data

Dalam pembersihan data kali ini penulis menekan pada pembersihan terhadap *missing value* dan *outlier*. Baris yang terdapat nilai kosong maka akan dihapus. Sedangkan untuk *outlier* menggunkan metode *Interquatile Range* (IQR).

c. Tranformasi dan Normalisasi data

Dataset terdapat beberapa kolom bertype kategorikal seperti typeppk, jkpst, dan cmg, kolom-kolom ini dikonversikan ke bentuk numerik menggunakan teknik *Label Encoding*. Sedangkan untuk normalisasi penulis menggunakan Min-Max Scaling. Bertujuan agar semua fitur memiliki skala nilai yang seragam dalam rentang [0, 1] sehingga tidak ada fitu yang mendominasi dalam proses pelatihan model.

d. Seleksi Fitur

Untuk meningkatkan efisiensi dan performa model, dilakukan seleksi fitur dengan teknik :

1. *Variance Threshold* Menghapus fitur yang memiliki variansi sangat rendah (di bawah 0.01), karena fitur ini dianggap tidak memberikan kontribusi informasi yang berarti terhadap target.
2. *SelectKBest* dengan ANOVA *F-test*: yaitu dengan memilih 20 fitur terbaik yang memiliki nilai *F-score* tertinggi, mengindikasikan hubungan paling kuat dengan target (label) menggunakan analisis varian (ANOVA).
3. *Random Forest feature Importance :* Model Random Forest digunakan untuk mengukur pentingnya fitur berdasarkan kontribusinya dalam membagi data di node decision tree. Diambil 20 fitur dengan skor tertinggi sebagai fitur yang paling berpengaruh terhadap klasifikasi

Seluruh hasil seleksi fitur dari ketiga metode ini ditampilkan dalam Tabel 3 untuk dianalisis lebih lanjut dan digunakan dalam pelatihan model prediktif.

Tabel 3. Hasil Praproses data

Variance Threshold	SelectKBest (ANOVA)	Random Forest Importance
kdkc	kdkc	umur
dati2	dati2	dati2
typeppk	typeppk	kdkc
jkpst	jkpst	diagprimer
umur	umur	typeppk
jnspelsep	jnspelsep	cmg
los	los	proc80_99
cmg	cmg	jkpst
severitylevel	severitylevel	los
diagprimer	diagprimer	severitylevel
proc80_99	proc58_62	jnspelsep
	proc63_67	dx2_h00_h59
	proc68_70	dx2_m00_m99
	proc71_73	dx2_l00_l99
	proc74_75	dx2_koo_k93
	proc76_77	dx2_j00_j99
	proc78_79	dx2_i00_i99
	proc80_99	dx2_h60_h95
	proce00_e99	dx2_c00_d48

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

Variance Threshold	SelectKBest (ANOVA)	Random Forest Importance
	procv00_v89	dx2_g00_g99

3.3 Hasil Pemodelan

3.3.1 Artificial Neural Network

Pada Pembangunan model ANN penulis menggunakan skenario dengan menerapkan 5 hidden layer visualisasi seperti dalam tabel 5. Untuk mengurangi kemungkinan terjadi *overfitting*, hasil dari seleksi fitur dijalan pada model ANN dan penerapan regulasi Learning Rate, Dropout dan Early Accuracy. Selain itu juga dilakukan tunning hyperparameter guna memperoleh performa terbaik, melakukan uji dengan dengan menggunakan dua hyperparameter dengan kombinasi terbaik yaitu, Learning Rate (LR) diuji dengan nilai : 0.001, 0.0005, 0.0001. dan Dropout diuji dengan nilai 0.2, 0.3, 0.4 Eksperimen dilakukan dengan mencari kombinasi LR dan Dropout yang memberikan nilai akurasi terbaik. Penjelasan hasil pelatihan model ANN pada Tabel 4.

Tabel 4. Hasil Pemodelan ANN

Learning Rate	Dropout	Validation Accuracy
0.001	0.2	76.5%
0.001	0.3	78.3%
0.001	0.4	80.1%
0.0005	0.2	83.8%
0.0005	**0.3**	**84.2%**
0.0005	0.4	83.6%
0.0001	0.2	80.5%
0.0001	0.3	82.8%
0.0001	0.4	83.2%

Tabel 5. Arsitektur ANN

layer type	output shape	Param #
Hidden_Layer_1 (Dense)	(None, 128)	6,656
Dropout_1 (Dropout)	(None, 128)	0
Hidden_Layer_2 (Dense)	(None, 64)	8,256
Dropout_2 (Dropout)	(None, 64)	0
Hidden_Layer_3 (Dense)	(None, 32)	2,000
Hidden_Layer_4 (Dense)	(None, 16)	528
Hidden_Layer_5 (Dense)	(None, 8)	136
Output_Layer (Dense)	(None, 1)	1

Tabel 4 menunjukkan pengaruh variasi learning rate dan dropout mempengaruhi terhadap akurasi validasi pada model klasifikasi. Terlihat bahwa kombinasi parameter ini memiliki dampak signifikan terhadap performa model. Pada learning rate sebesar 0.001, peningkatan nilai dropout dari 0.2 ke 0.4 menghasilkan peningkatan akurasi validasi dari 76.5% menjadi 80.1%, yang menunjukkan bahwa regularisasi melalui dropout membantu mengurangi overfitting pada tingkat pembelajaran tersebut.

Namun, peningkatan paling signifikan dicapai saat learning rate diturunkan menjadi 0.0005. Pada kombinasi ini, akurasi validasi mencapai puncaknya di 84.2% ketika dropout bernilai 0.3. Ini menunjukkan bahwa learning rate yang lebih kecil memberikan waktu lebih banyak bagi model untuk belajar pola data secara bertahap, dan pada saat yang sama dropout sebesar 0.3 mampu menjaga keseimbangan antara underfitting dan overfitting.

Sementara itu, learning rate yang terlalu kecil seperti 0.0001 cenderung menghasilkan akurasi yang lebih rendah dibandingkan 0.0005, meskipun masih berada pada tingkat yang cukup baik (maksimal 83.2%). Hal ini bisa disebabkan karena proses pembelajaran menjadi terlalu lambat, sehingga model tidak mencapai konvergensi optimal dalam jumlah epoch yang ditetapkan. Secara keseluruhan, hasil ini mengindikasikan bahwa kombinasi learning rate 0.0005 dan dropout 0.3 merupakan konfigurasi yang paling optimal untuk model yang diuji, karena memberikan performa validasi terbaik sebesar 84.2%.

3.3.2 Support Vector Machine

Model Support Vector Machine (SVM) digunakan sebagai salah satu baseline dalam eksperimen klasifikasi potensi fraud. SVM memiliki keunggulan dalam menangani data berdimensi tinggi dan menjaga margin pemisah antar kelas, yang menjadikannya cocok untuk klasifikasi dengan distribusi data yang tidak seimbang maupun kompleks. Dalam eksperimen ini, digunakan kernel RBF (Radial Basis Function) yang bersifat non-linear, karena mampu menangkap pola-pola yang kompleks dalam data fraud.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

Akurasi validasi terbaik yang diperoleh dari model SVM mencapai 98%, yang menunjukkan bahwa SVM dapat mengidentifikasi pola fraud dengan cukup baik. Kelebihan dari model SVM adalah stabilitasnya terhadap outlier dan kemampuannya untuk menemukan margin optimal.

Dengan demikian, model SVM tetap layak dipertimbangkan sebagai model pembanding yang kuat dalam eksperimen klasifikasi fraud, khususnya ketika diperlukan model dengan performa yang stabil dan mudah dikendalikan. Tabel 6 merupakan hasil dari pemodelan SVM.

Tabel 6. Hasil pemodelan SVM

Fold	Precision	Recall	F1-Score	Accuracy
0	0.9818	0.9848	0.9879	0.9901
1	0.9880	0.9899	0.9888	0.9833
2	0.9820	0.9026	0.9821	0.9870
3	0.9875	0.9845	0.9859	0.9815
4	0.9899	0.9893	0.9894	0.9845
5	0.9806	0.9831	0.9817	0.9861
6	0.9707	0.9851	0.9875	0.9790
7	0.9843	0.9854	0.9847	0.9895
8	0.9808	0.9852	0.9879	0.9842
9	0.9830	0.9811	0.9919	0.9874
Rata-rata	0.9899	0.9821	0.9808	0.9853

3.3.3 Random Forest

Model *Random Forest Classifier* dievaluasi dengan menggunakan skema Repeated 5 fold dengan 2 pengulangan dan mendapatkan 10 iterasi. Evaluasi menghasilkan Presisi (precision) 67.99%, Recall 70.21%, F1-score 69.08% dan Akurasi sebesar 68.53%. model random forest memberikan kinerja klasifikasi yang kompetitif dengan f1-score mendekati 70%. Dengan kemampuan menangani data mutivariabel dan fitur kategarikal, model ini layak dipertimbangkan sebagai baseline dalam sistem prediksi fraud dalam klaim layanan kesehatan. Tabel 7 merupakan hasil dari pemodelan Random Forest.

Tabel 7. Hasil pemodelan Random Forest

Fold	Precision	Recall	F1-Score	Accuracy
0	0.6818	0.7148	0.6979	0.6901
1	0.6780	0.6999	0.6888	0.6833
2	0.6820	0.7026	0.6921	0.6870
3	0.6775	0.6945	0.6859	0.6815
4	0.6799	0.6993	0.6894	0.6845
5	0.6806	0.7031	0.6917	0.6861
6	0.6707	0.7051	0.6875	0.6790
7	0.6843	0.7054	0.6947	0.6895
8	0.6808	0.6952	0.6879	0.6842
9	0.6830	0.7011	0.6919	0.6874
Rata-rata	0.6799	0.7021	0.6908	0.6853

3.3.4 XGBoost

Model XGBoost Classifier dievaluasi menggunakan metode Repeated Stratified K-Fold Cross Validation dengan 5 fold dan 2 pengulangan, menghasilkan total 10 iterasi pelatihan dan pengujian model. Evaluasi dilakukan terhadap metrik klasifikasi yang meliputi precision, recall, f1-score, dan accuracy untuk mengukur kinerja model dalam mendeteksi klaim yang terindikasi fraud. Menghasilkan presisi (*Precision*) yang menunjukkan hasil klasifikasi fraud 71.36%. *Recall* 71.73%, *F1-score* 71.73% dan Akurasi (*Accuracy*) 71.73%. Seperti pada Tabel 8 berikut

Tabel 8. Hasil XGBoost

Fold	Precision	Recall	F1-Score	Accuracy
0	0.7135	0.7240	0.7187	0.7161
1	0.7067	0.7115	0.7091	0.7075
2	0.7122	0.7173	0.7147	0.7131
3	0.7121	0.7136	0.7128	0.7120
4	0.7129	0.7225	0.7177	0.7152
5	0.7142	0.7193	0.7167	0.7152
6	0.7132	0.7123	0.7127	0.7123
7	0.7139	0.7228	0.7184	0.7161
8	0.7131	0.7169	0.7150	0.7137

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

Fold	Precision	Recall	F1-Score	Accuracy
9	0.7139	0.7195	0.7166	0.7150
Rata-rata	0.7126	0.7180	0.7152	0.7136

3.3.5 Logistic Regression

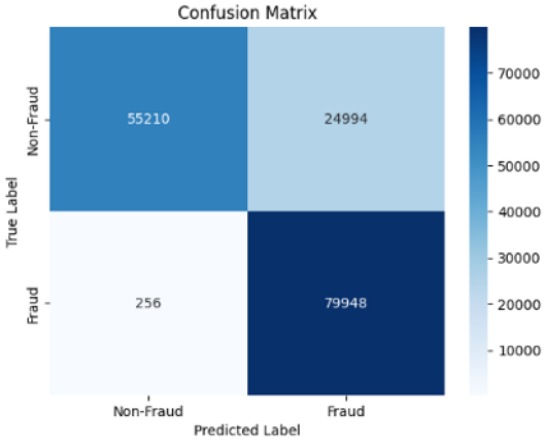
Model Logistic Regression telah dievaluasi menggunakan teknik Repeated Stratified K-Fold Cross Validation sebanyak 10 kali, setelah dilakukan peningkatan kualitas data melalui proses seleksi fitur menggunakan SelectKBest (30 fitur teratas), normalisasi fitur dengan StandardScaler, serta penyeimbangan kelas menggunakan metode SMOTE. Hasil evaluasi menunjukkan bahwa model ini menghasilkan rata-rata precision sebesar 55.41%, recall sebesar 56.35%, f1-score sebesar 55.87%, dan akurasi sebesar 55.50%. Meskipun Logistic Regression memiliki keterbatasan dalam menangani hubungan non-linear, hasil ini cukup kompetitif setelah dilakukan optimasi data. Tabel 9 berikut menunjukkan hasil evaluasi pada tiap fold

Tabel 9. Hasil model Logistic Regression

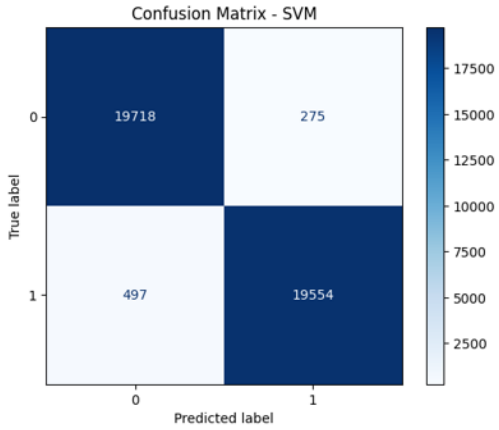
Fold	Precision	Recall	F1-Score	Accuracy
0	0.554175	0.569348	0.561659	0.555658
1	0.553720	0.561618	0.557641	0.554486
2	0.556301	0.566206	0.561210	0.557304
3	0.551475	0.558626	0.555027	0.552142
4	0.553644	0.561817	0.557701	0.554436
5	0.553925	0.568401	0.561069	0.555334
6	0.554906	0.558227	0.556561	0.555234
7	0.554817	0.570395	0.562498	0.556356
8	0.554492	0.560271	0.557366	0.555060
9	0.553120	0.560072	0.556574	0.553788
Rata-rata	0.554057	0.563498	0.558731	0.554980

3.4 Evaluasi Model

Selanjutnya dilakukan Evaluasi model dilakukan terhadap empat algoritma klasifikasi, yaitu ANN, Logistic Regression, Random Forest, XGBoost, dan Support Vector Machine (SVM) seperti pada gambar 4 dan 5. Model kemudian dilatih dan dievaluasi menggunakan skema Repeated Stratified K-Fold Cross Validation untuk menghindari bias akibat pemisahan data. Hasil metrik evaluasi seperti precision, recall, f1-score, dan accuracy telah disajikan sebelumnya. Untuk mendapatkan gambaran yang lebih konkret terhadap performa model dalam mendeteksi kasus fraud, dilakukan analisis berdasarkan confusion matrix masing-masing model.



Gambar 4. Confusion Matrix ANN

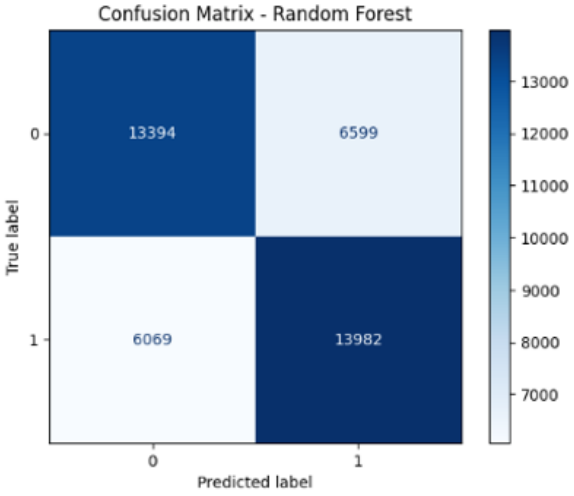


Gambar 5. Confusion Matrix SVM

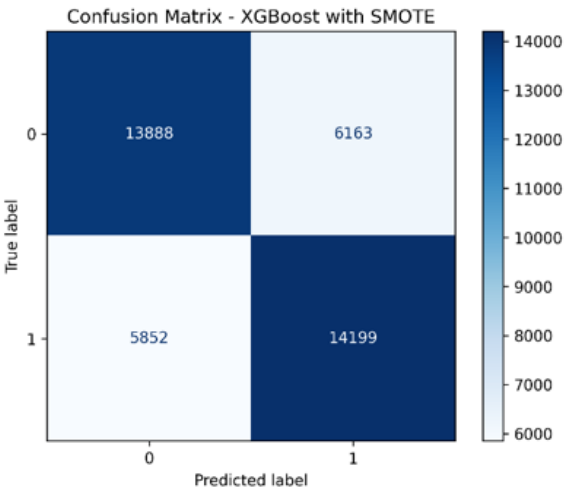
Confusion matrix menggambarkan kemampuan model dalam memetakan data aktual menjadi prediksi yang benar (True Positive & True Negative) maupun salah (False Positive & False Negative). Visualisasi confusion matrix dari masing-masing model menunjukkan perbedaan kinerja dalam mengklasifikasikan data fraud dan non-fraud.

Random Forest menunjukkan performa yang lebih seimbang, dengan jumlah True Positive dan True Negative yang lebih tinggi. Model ini unggul dalam mengenali pola-pola kompleks dengan kombinasi fitur yang beragam. XGBoost menghasilkan confusion matrix terbaik, dengan rasio True Positive yang tinggi dan False Positive yang lebih rendah dibanding model lain. Hal ini menegaskan keunggulan XGBoost dalam menangani dataset klasifikasi biner pada data yang tidak seimbang secara distribusi. Seperti terlihat pada Gambar 6 dan Gambar 7.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

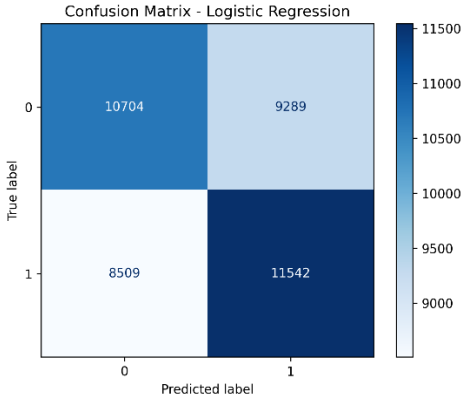


Gambar 6. Confusion Matrix RF



Gambar 7. Confusion Matrix XGBoost

Logistic Regression cenderung menghasilkan lebih banyak False Negatives, yang berarti model ini masih melewatkan sejumlah kasus fraud sebenarnya karena keterbatasannya dalam menangani pola data non-linear. Model Logistic Regression menunjukkan kemampuan yang seimbang antara mendeteksi klaim fraud dan non-fraud, namun dengan tingkat kesalahan yang cukup signifikan, terutama pada kelas fraud yang salah diklasifikasikan (FN = 8.509). Ini menunjukkan bahwa meskipun Logistic Regression mampu memberikan klasifikasi yang kompeten secara umum, model ini memiliki keterbatasan dalam mendeteksi fraud secara akurat, yang merupakan target kritikal dalam studi ini. Angka False Positive (9.289) yang tinggi juga menunjukkan bahwa model sering menganggap klaim sah sebagai fraud, yang dapat menyebabkan verifikasi manual yang tidak perlu dan berpotensi mengganggu layanan



Gambar 8. Confusion Matrix LR

3.5 Waktu Kinerja Model

Efisiensi komputasi merupakan aspek penting dalam sistem operasional berskala besar seperti BPJS Kesehatan. Oleh karena itu, dilakukan pengukuran rata-rata waktu pelatihan untuk setiap model:

Tabel 9. Waktu Pelatihan Model

Model	Waktu Pelatihan (per fold)	Karakteristik Komputasi
Logistic Regression	± 10 Menit	Sangat efisien, cocok untuk sistem real-time
Support Vector Machine	± 10 Menit	Cepat, dengan akurasi dan stabilitas tinggi
Random Forest	± 19 menit	Waktu sedang, hasil cukup baik namun perlu tuning
XGBoost + SMOTE	± 15 menit	Cukup berat, namun mampu menangani ketidakseimbangan
ANN	± 25 menit	Waktu paling lama, akurat namun cocok untuk batch mode

3.6 Analisa Kinerja Model

Evaluasi model dilakukan terhadap lima algoritma klasifikasi: Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), XGBoost dengan SMOTE, dan Logistic Regression (LR). Model dievaluasi

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

menggunakan Repeated Stratified K-Fold Cross Validation dan dinilai dengan metrik evaluasi seperti precision, recall, F1-score, accuracy, serta visualisasi confusion matrix.

Dari hasil evaluasi, model Logistic Regression menunjukkan akurasi paling rendah, yaitu 55.5%, meskipun sudah dilakukan preprocessing dan balancing menggunakan SMOTE. Beberapa alasan yang menyebabkan kinerja LR tidak optimal antara lain:

a. Asumsi linearitas pada model Logistic Regression tidak sesuai dengan karakteristik data klaim yang bersifat kompleks dan memiliki relasi non-linear antar fitur.
b. Multikolinearitas antar fitur yang tinggi menyebabkan ketidakstabilan estimasi koefisien pada model.
c. LR juga tidak mampu menangkap interaksi kompleks antar fitur, seperti kombinasi diagnosis, prosedur, dan durasi rawat inap, yang berperan penting dalam identifikasi fraud.

Sementara itu, Random Forest hanya mencapai akurasi 68.5%, padahal secara teori algoritma ini cukup kuat dalam menangani data multivariabel. Rendahnya kinerja RF dapat disebabkan oleh:

a. Adanya fitur yang kurang informatif atau redundan, yang mempersulit pembentukan pohon keputusan optimal.
b. Distribusi noise dalam data, seperti outlier pada variabel LOS atau umur, yang memengaruhi pembentukan split pada decision tree.
c. Model tidak dilakukan hyperparameter tuning secara mendalam, sehingga belum mencapai konfigurasi optimal (jumlah pohon, kedalaman pohon, minimal sample leaf, dll.).

Sebaliknya, SVM mencatat performa paling tinggi (98%), menunjukkan efektivitas algoritma ini dalam mengenali pola fraud pada data berdimensi tinggi. ANN juga unggul dalam recall terhadap kelas fraud, menunjukkan kemampuan dalam menangkap pola non-linear yang kompleks.

4. KESIMPULAN

Penelitian ini mengevaluasi lima algoritma machine learning antara lain Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), XGBoost dengan SMOTE, dan Logistic Regression (LR). untuk klasifikasi potensi fraud pada klaim layanan kesehatan JKN. Berdasarkan evaluasi metrik utama yaitu precision, recall, F1-score, dan AUC-ROC, ANN menunjukkan performa terbaik dalam mendeteksi klaim fraud dengan recall tertinggi, menjadikannya unggul dalam mengidentifikasi kasus yang benar-benar fraud, meskipun dengan risiko false positive yang lebih tinggi. XGBoost yang dikombinasikan dengan SMOTE juga menunjukkan kinerja yang kompetitif dengan F1-score dan AUC yang tinggi, berkat kemampuannya menangani ketidakseimbangan data. Sementara itu, SVM tampil konsisten dengan keseimbangan antara precision dan recall, namun memiliki jumlah true positive yang jauh lebih rendah, menunjukkan bahwa model ini lebih konservatif dalam mendeteksi fraud, yang bisa berdampak pada banyaknya kasus fraud yang terlewat. Random Forest dan Logistic Regression menghasilkan performa yang moderat dan dapat digunakan sebagai baseline, namun terbatas dalam menangani pola fraud yang kompleks. Oleh karena itu, pemilihan model terbaik sangat tergantung pada prioritas sistem deteksi—apakah lebih menekankan pada deteksi maksimum fraud (recall tinggi) seperti ANN dan XGBoost + SMOTE, atau keseimbangan antar metrik dengan toleransi kesalahan lebih rendah seperti pada SVM. Hasil ini menegaskan pentingnya strategi penyeimbangan data dan pemilihan metrik evaluasi yang sesuai dalam merancang sistem klasifikasi fraud yang adaptif, efisien, dan responsif terhadap karakteristik data klaim kesehatan.

REFERENCES

[1] P. Gunadi, S. Hasan, I. N. Arda, E. D. Rahayu, and Z. Andika, “Implementasi Sistem Verifikasi Digital Untuk Pencegahan Fraud Pada Program Jkn (Studi Kasus Rs Syariah Jakarta),” *Media Riset Bisnis Ekonomi Sains dan Terapan*, vol. 2, no. 4, pp. 1–8, Jan. 2025, doi: 10.71312/mrbest.v2i4.208.

[2] B. Santoso, J. Hendrartini, B. U. Djoko Rianto, and L. Trisnantoro, “System For Detection Of National Healthcare Insurance Fraud Based On Computer Application,” *Public Health of Indonesia*, vol. 4, no. 2, pp. 46–56, Jun. 2018, doi: 10.36685/phi.v4i2.199.

[3] R. Annisa, S. Winda, E. Dwisaputro, and K. N. Isnaini, “Mengatasi Defisit Dana Jaminan Sosial Kesehatan Melalui Perbaikan Tata Kelola,” *INTEGRITAS: Jurnal Antikorupsi*, vol. 6, no. 2, pp. 209–224, 2020, doi: 10.32697/integritas.v6i2.664.

[4] I. Sugiarti, I. Masturoh, and F. Fadly, “Menelusuri Potensi Fraud dalam Jaminan Kesehatan Nasional melalui Rekam Medis di Rumah Sakit,” *Jurnal Kesehatan Vokasional*, vol. 7, no. 1, p. 42, Feb. 2022, doi: 10.22146/jkesvo.69056.

[5] M. S. M. S. Humasak Tommy Argo Simanjuntak, “Deteksi Fraud Pada Klaim Layanan Rumah Sakit Menggunakan Model Neural Network,” *Journal of Applied Technology and Informatics Indonesia* , vol. vol 1, 2021, doi: 10.54074/jati.v1i1.30.

[6] A. C. Nugraha and M. I. Irawan, “Komparasi Deteksi Kecurangan pada Data Klaim Asuransi Pelayanan Kesehatan Menggunakan Metode Support Vector Machine (SVM) dan Extreme Gradient Boosting (XGBoost),” *Jurnal Sains dan Seni ITS*, vol. 12, no. 1, May 2023, doi: 10.12962/j23373520.v12i1.107032.

[7] E. Nabrawi and A. Alanazi, “Fraud Detection in Healthcare Insurance Claims Using Machine Learning,” *Risks*, vol. 11, no. 9, Sep. 2023, doi: 10.3390/risks11090160.

[8] C. E. D. Vanegas, J. C. G. Mejía, F. A. V. Agudelo, and D. E. S. Duran, “A Representation Based on Essence for the CRISP-DM Methodology,” *Computacion y Sistemas*, vol. 27, no. 3, pp. 675–689, 2023, doi: 10.13053/CyS-27-3-3446.

[9] R. Winurputra and D. E. Ratnawati, “Peramalan Penjualan Produk Menggunakan Extreme Gradient Boosting (Xgboost) Dan Kerangka Kerja Crisp-Dm Untuk Pengoptimalan Manajemen Persediaan (Studi Kasus : Ub Mart) Product Sales

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 296−307
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7349

Forecasting Using Extreme Gradient Boosting (Xgboost) And Crisp-Dm ," vol. 12, no. 2, pp. 417–428, 2025, doi: 10.25126/jtiik.2025129451.

[10] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," *J Big Data*, vol. 10, no. 1, p. 154, 2023, doi: 10.1186/s40537-023-00821-5.

[11] C. Li, "Preprocessing Methods and Pipelines of Data Mining: An Overview," Jun. 2019, doi: https://doi.org/10.48550/arXiv.1906.08510.

[12] H. N. Rofiq, "Deteksi Inefisiensi pada Klaim BPJS Kesehatan dengan menggunakan Machine Learning," *Jurnal Jaminan Kesehatan Nasional*, vol. 3, no. 1, Jun. 2023, doi: 10.53756/jjkn.v3i1.134.

[13] L. N. Hapsari and N. Rokhman, "Anomaly Detection of Hospital Claim Using Support Vector Regression," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 18, no. 1, p. 1, 2024, doi: 10.22146/ijccs.91857.

[14] E. Nabrawi and A. Alanazi, "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," *Risks*, vol. 11, no. 9, Sep. 2023, doi: 10.3390/risks11090160.

[15] A. Karim, "Perbandingan Prediksi Kemiskinan di Indonesia Menggunakan Support Vector Machine (SVM) dengan Regresi Linear," *Jurnal Sains Matematika dan Statistika*, vol. 6, no. 1, p. 107, 2020, doi: 10.24014/jsms.v6i1.9259.

[16] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/bjml/2024/007.

[17] S. Parthasarathy, A. Raj Lakshminarayanan, A. Abdul Azeez Khan, K. Javubar Sathick, and V. Jayaraman, "Detection of Health Insurance Fraud using Bayesian Optimized XGBoost," *International Journal of Safety and Security Engineering*, vol. 13, no. 5, pp. 853–861, Nov. 2023, doi: 10.18280/ijsse.130509.

[18] R. Tyasnurita and A. Y. M. Pamungkas, "Deteksi Diabetik Retinopati menggunakan Regresi Logistik," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 130–135, Aug. 2020, doi: 10.33096/ilkom.v12i2.578.130-135.

[19] R. Sahila, T. Widiharih, and I. T. Utami, "Analisis Klasifikasi Menggunakan Regresi Logistik Biner Dan Algoritma Naïve Bayes Classifier Pada Penyakit Hipertensi," *Jurnal Gaussian*, vol. 13, no. 2, pp. 319–327, Nov. 2024, doi: 10.14710/j.gauss.13.2.319-327.

[20] F. R. Suprihati, "Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression," *Jurnal Sistem Cerdas*, vol. 4, no. 3, pp. 155–160, 2021, doi: 10.37396/jsc.v4i3.166.

[21] A. M. Syahbani, W. Firdaus, and K. A. Musodo, "A Comparative Study of Data Mining Algorithms for Fraud Detection in Financial Transactions," *Sinkron*, vol. 9, no. 2, pp. 814–821, Apr. 2025, doi: 10.33395/sinkron.v9i2.14645.