Komparasi Deteksi Kecurangan pada Data Klaim Asuransi Pelayanan Kesehatan Menggunakan Metode *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting* (XGBoost)

Alan Catur Nugraha, dan Mohammad Isa Irawan
Departemen Matematika, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: mii@its.ac.id

***Abstrak*—Pada era informasi ini banyak proses digitalisasi di berbagai bidang kehidupan maka semakin penting juga informasi yang didapatkan dari kumpulan data yang ada. Dampak dari perkembangan ini adalah semakin mudah terlihat kejanggalan pada data yang biasa terjadi dikarenakan adanya praktek kecurangan atau *fraud*. Deteksi adanya *fraud* pada layanan kesehatan penting dilakukan untuk dalam pengambilan keputusan yang diambil penyedia layanan kesehatan. *Fraud* pada layanan kesehatan itu sendiri merupakan masalah utama yang sering dialami penyedia layanan kesehatan saat ini yang merugikan banyak pihak di dalamnya. Oleh karena itu, penelitian ini membahas bagaimana cara mendeteksi *fraud* pada pelayanan kesehatan dengan cara *machine learning*. *Machine learning* adalah cara peningkatan kemampuan mesin dalam menyelesaikan masalah yang baru. Metode *machine learning* yang digunakan adalah klasifikasi *Support Vector Machine* (SVM) dan metode klasifikasi *Extreme Gradient Boosting* (XGBoost) yang hasilnya dibandingkan untuk melihat model yang lebih baik. Hasil yang didapatkan adalah hasil yang berhasil mendeteksi data *fraud* pada data pelayanan kesehatan tersebut dengan performa klasifikasi yang baik dalam membantu memberikan referensi pada penyedia layanan dalam mendeteksi *fraud* . Metode XGBoost menghasilkan performa klasifikasi yang baik dengan menghasilkan nilai *Balanced Accuracy* dan nilai *Recall* sebesar 0.9995 dan 0.9994.**

***Kata Kunci*—Extreme Gradient Boosting, Fraud Detection, Healthcare, Support Vector Machine.**

I. PENDAHULUAN

PADA era informasi ini, salah satu komoditas paling penting di berbagai bidang kehidupan saat ini adalah informasi. Informasi merupakan kunci utama untuk sebuah organisasi untuk terus berkembang. Semakin banyak sebuah organisasi memiliki informasi dan dapat mengolahnya dengan baik, semakin optimal juga organisasi tersebut dalam mencapai hasil yang diinginkan [1]. Di berbagai bidang kehidupan saat ini mulai gencar dalam melakukan digitalisasi dalam memproses informasi tersebut mulai dari pengambilan informasi, dalam kegiatan administrasi, pengolahan informasi, sampai pada penyimpanan informasi tersebut pada *storage digital*. Salah satunya adalah pada bidang pelayanan kesehatan yang contohnya adalah digitalisasi pada sistem medisnya itu sendiri, rekam medis pasien, proses citra medis, dan berbagai informasi lainnya. Keuntungan digitalisasi ini pada penyedia layanan kesehatan contohnya adanya peningkatan pada pelayanan kesehatan yang diberikan penyedia layanan kesehatan terhadap pasien, dan meningkatkan keamanan penyimpanan informasi pasien. Para ahli di bidang terkait menghitung pada tahun 2021 jumlah data yang ada pada bidang layanan kesehatan 44 kali lebih besar dibanding tahun 2009.

Perubahan bentuk sistem ini menyebabkan berbagai informasi yang sebelumnya sulit untuk ditemukan menjadi lebih mudah untuk ditemukan. Contoh informasi yang dapat ditemukan tersebut adalah kecurangan-kecurangan yang pernah terjadi pada sistem pelayanan kesehatan. Kecurangan atau yang sering dikenal dengan nama *fraud* adalah tindakan disengaja dengan tujuan mendapatkan keuntungan ilegal baik ditujukan pada pribadi atau sebuah kelompok/organisasi dengan melakukan tindakan penipuan, memberikan saran palsu, atau menahan kebenaran. Di berbagai negara di dunia terjadi peningkatan dalam hal pengeluaran keuangan dalam bidang pelayanan kesehatan, pada tahun 2011 setidaknya 10\% produk domestik bruto negara dihabiskan di bidang pelayanan kesehatan. Sayangnya, tidak semua dana tersebut tersalurkan sepenuhnya secara benar. Banyak pengeluaran tersebut justru dikarenakan adanya *fraud*, kegiatan yang *wasteful* dan penyalahgunaan wewenang [2]. Bentuk kecurangan pada pelayanan kesehatan yang sering terjadi di Indonesia diantaranya penulisan diagnosis yang berlebihan (*upcoding*), duplikasi klaim data pasien lain (cloning), klaim palsu (*phantom billing*), penggelembungan tagihan obat (*inflated bills*), pemecahan bagian pelayanan (*services unbundling*), rujukan semu (*selfs-referrals*), tagihan berulang (*repeat billing*), memperpanjang lama perawatan, memanipulasi kelas perawatan dan pembatalan tindakan (*cancelled services*) [3].

Pada tanggal Januari 2022 di *New York City* Amerika Serikat, pihak berwenang Amerika Serikat menangkap 13 orang terpidana dengan tuduhan melakukan kecurangan pada layanan kesehatan, praktik cuci uang, dan penyuapan sebesar 100 juta USD. Dari 13 terpidana tersebut 8 diataranya merupakan orang yang bekerja sebagai dokter dan tenaga medis. Adapun kasus kecurangan di layanan kesehatan lainnya adalah pada Oktober 2021 di provinsi Sichuan Tiongkok dimana 47 karyawan rumah sakit swasta diduga melakukan kecurangan pada dana asuransi kesehatan sebesar 1.5 juta USD. Terpidana yang juga termasuk pemilik rumah sakit terkait ditahan atas pemalsuan rekam medis dan pemalsuan akun untuk menutupi biaya operasi.

Salah satu cara yang saat ini digunakan adalah menggunakan *Machine Learning*. *Machine learning* sendiri adalah cabang ilmu dari *Artificial Intellegence*} yang tujuan utamanya adalah meningkatkan kemampuan mesin dalam mempelajari informasi baru dari data dan mengembangkannya untuk menyelesaikan suatu

permasalahan [4-6]. Terdapat berbagai metode *machine learning* yang digunakan dalam mendeteksi *fraud* diantaranya dengan mengunakan metode *Evolving Clustering Method* (ECM), *K-Means Algorithm*, *Outlier Detection*, *Anamoly Detection* dan *Support Vector Machine* (SVM). Pada penelitian terdahulu tersebut telah digunakan metode klasifikasi SVM pada kasus serupa, pada penelitian tersebut didapatkan akurasi model sebesar 87,91\% dengan *kernel radial basis function* [7]. Metode klasifikasi SVM adalah metode yang digunakan untuk mengklasifikasi data menjadi dua kelas berbeda yang pada kasus ini adalah menentukan data yang merupakan data *fraud* dan yang tidak [8]. Metode SVM merupakan metode *machine learning* yang memiliki keuntungan dapat bekerja baik pada data yang memiliki pembeda yang jelas pada kelas-kelasnya yang dalam hal ini adalah kelas *fraud* dan kelas tidak *fraud* dan SVM relatif bekerja lebih baik pada data yang berdimensi tinggi. Selanjutnya terdapat metode *machine learning* lain yang juga dapat digunakan yaitu dengan metode *Extreme Gradient Boost*. Pada penelitian terdahulu metode XGBoost digunakan pada kasus yang berbeda dengan tipe data yang serupa yaitu pada data *credit card* dalam deteksi *fraud* [9]. Oleh karena itu pada penelitian ini digunakan metode klasifikasi SVM dengan *kernel* RBF untuk menyelesaikan permasalahan serupa dengan data berbeda yang dibandingkan hasil performanya menggunakan metode XGBoost. Data yang digunakan adalah data klaim peserta asuransi yang disediakan penyedia layanan kesehatan.

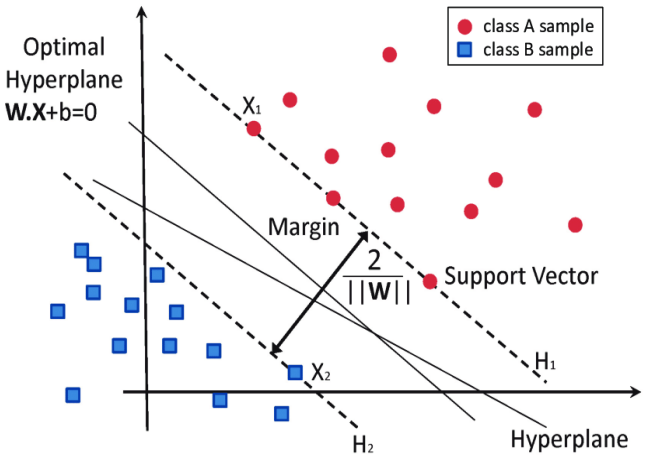
Pada jurnal ini mengklasifikasi klaim data layanan kesehatan untuk mendeteksi adanya kecurangan data asuransi menggunakan metode *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting* (XGBoost) dan menganalisa dan membandingkan hasil dari penerapan klasifikasi *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting* (XGBoost) dalam melakukan deteksi *fraud* pada data asuransi pelayanan kesehatan.

II. TINJAUAN PUSTAKA

A. *Fraud Pada Asuransi Kesehatan*

Menurut *The Association of Certified Fraud Examiners* (ACFE) tindak kecurangan atau dalam hal ini disebut *fraud* adalah tindakan memperkaya diri dengan melalui penyalahgunaan dalam menggunakan aset dan sumber daya organisasi dengan cara yang disengaja [10]. Secara umum *fraud* adalah tindakan disengaja dengan tujuan mendapatkan keuntungan yang tidak sah untuk kepentingan pribadi atau kelompok dengan cara menipu atau dengan cara tidak etis lainnya yang merugikan orang lain. Tindakan merampas keuntungan pribadi atau kelompok dengan segala cara yang tidak sah dapat disebut juga sebagai *fraud*. Beberapa contoh dari *fraud* dalam kehidupan sehari-hari diantaranya adalah segala bentuk penggelapan, pemalsuan arsip, pemalsuan laporan finansial, dan pelanggaran etika lainnya yang merugikan orang lain termasuk perusahaan penyedia asuransi kesehatan [11].

Fraud yang dilakukan penyedia layanan kesehatan salah satu contohnya adalah penyedia layanan kesehatan menagih biaya pada layanan yang sebenarnya tidak dilakukan, sedangkan *fraud* yang dilakukan pasien contohnya adalah pasien memalsukan riwayat pekerjaan untuk mendapatkan keuntungan asuransi, dan *fraud* yang dilakukan penyedia layanan asuransi kesehatan diantaranya adalah asuransi kesehatan memalsukan pernyataan *benefit*. Ketiga contoh *fraud* tersebut adalah contoh yang sering ditemukan pada kegiatan *fraud detection*, namun tidak menutup kemungkinan lebih dari satu pihak tersebut yang bertanggung jawab atas tindakan *fraud* pasien bekerja sama dengan penyedia layanan kesehatan dalam memalsukan transaksi, memalsukan riwayat untuk mendapatkan dana asuransi [12].



Gambar 1. Ilustrasi SVM.

B. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) adalah salah satu algoritma *supervised machine learning* klasifikasi yang digunakan untuk mengklasifikasikan dua kelas yaitu +1 dan -1 seperti yang diperlihatkan pada Gambar 1 [8].

Dengan *decision rule* untuk mencari *hyperplane* yang membagi data menjadi dua kelas sebagai Persamaan 1 berikut.

$$F(\mathrm{x}) = \mathrm{w} \cdot \mathrm{x} + \mathrm{b} \quad (1)$$

Sehingga diperoleh persamaan untuk setiap sampel data yang bernilai positif.

$$[(w \cdot x_+) + b] \geq 1$$

dan untuk sampel data bernilai negatif.

$$[(w \cdot x_-) + b] \leq -1$$

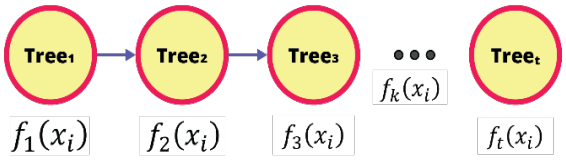
Dimana w merupakan vektor bobot, x_i adalah vektor data, dan b merupakan nilai skalar. Dengan menggunakan rumus untuk menghitung jarak antara H_1 dan H_2 dimana H_1 dan H_2 adalah daerah diantara data *point* yang diilustrasikan pada Gambar 1 dengan daerah H_1 dan H_2 maka akan didapatkan nilai *margin* yang dinotasikan sebagai d dengan $w_1.x_1 + w_2.x_2 + b = 0$, $w = (w_1, w_2), x = (x_1, x_2)$ sehingga didapatkan persamaan *margin*.

$$d = \frac{|(w_i \cdot x_i) + b|}{\sqrt{||w||^2}}$$

Berdasarkan pada persamaan *decision rule* untuk sample positif dan negatif maka didapatkan nilai $|(\boldsymbol{w} \cdot \boldsymbol{x_i}) + b|$ dengan melakukan operasi pengurangan pada persamaan didapatkan persamaan *margin* yang akan dimaksimalkan.

$$d = \frac{2}{||w||} \quad (2)$$

Dari Persamaan 2 diatas memaksimalkan *margin* pada



Gambar 2. Ilustrasi XGBoost.

persamaan $\frac{d^2}{2}$, sama dengan meminimalkan $\frac{1}{2}||w||^2$ dimana memaksimalkan *margin* merupakan kunci utama dari cara kerja SVM. Lalu dengan memisalkan y_i sedemikian hingga adalah +1 untuk sampel data positif dan -1 untuk sampel data negatif maka akan didapatkan persamaan sebagai berikut.

$$y_i[(w_i \cdot x_i) + b] - 1 \geq 0$$

Dimana dari persamaan tersebut didapatkan kendala $y_i[(\mathbf{w} \cdot \mathbf{x_i}) + b] - 1 = 0$ untuk setiap $\mathbf{x_i}$\$ yang ada pada daerah H_1 dan H_2.Sehingga didapatkan permasalahan dimana meminimumkan nilai dari $\frac{1}{2}||w||^2$ dengan kendala $y_i[(w \cdot x_i) + b] - 1 = 0$. Permalahan ini dapat diselesaikan dengan menggunakan optimasi *Lagrange Multiplier* dengan tujuan optimasi yang dilakukan adalah meminimalkan kesalahan prediksi dari pengklasifikasi dengan memaksimalkan jarak *margin* dari *support vector* pada Persamaan 3 berikut [8].

$$L = \frac{1}{2}||\boldsymbol{w_i}||^2 - \sum_{i=1} \alpha_i[y_i(\boldsymbol{w_i} \cdot \boldsymbol{x_i} + b) - 1] \quad (3)$$

Pada permasalahan *dual forms optimization* tersebut dapat terlihat bahwa dalam pengklasfikasi *margin* maksimum hanya bergantung pada *dot product* pada vektor data (x_i, x_j) dan bukan pada vektornya \shortc. Operasi $x_i \cdot x_j$ menunjukkan hasil kali dari vektor x_i dan x_j. Hal ini menunjukkan bahwa tidak harus dibutuhkan data *point* yang tepat, tetapi hanya produk dalam mereka untuk menghitung *hyperplane* terbaik. Sehingga didapatkan Persamaan 4 *hyperplane* sebagai berikut.

$$F(x) = (\sum_{i=1} \alpha_i \, y_i \boldsymbol{x}_i \cdot x) + b \quad (4)$$

Oleh karena itu untuk menyelesaikan permasalahan dimana data tidak dapat diklasifikasi secara linier, dapat digunakan konsep *kernel* untuk melihat relasi data *point* pada ruang berdimensi tinggi hanya dengan menggunakan *dot product* pada vektor data. Pada penelitian ini digunakan *kernel non linier Gaussian Radial Basis Function* (RBF) untuk menyelesaikannya permaslahan klasifikasi data klaim *fraud* pada data kesehatan [7]. *Kernel* RBF adalah *kernel* yang membantu mentransformasi data dari dimensi yang relatif rendah ke dimensi tak hingga dalam mencari *hyperplane* sehingga diperoleh Persamaan 5.

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2} \quad (5)$$

C. *Extreme Gradient Boosting*

Extreme Gradient Boosting atau XGBoost adalah salah satu metode supervised learning yang dapat digunakan dalam kasus klasifikasi atau regresi. XGBoost adalah algoritma machine learning system pada tree boosting yang dapat berkembang atau dibuat untuk sistem tree yang lebih besar. XGBoost merupakan metode lanjutan Gradient Boosting yang merupakan metode ensemble dari model yang digunakan pada decision tree yang dikembangkan untuk mendapatkan running time yang lebih cepat walaupun dalam memproses data besar. Extreme Gradient Boosting bekerja seperti metode learning boosting yang lainya yaitu dengan menggabungkan berbagai metode pengklasifikasi lemah menjadi kuat dengan melatih model satu demi satu secara berurutan menggunakan hasil klasifikasi yang didapat dari model sebelumnya yang disebut residuals atau error [13].

Dari Gambar 2 ditunjukkan bahwa $f_t(x_i)$ menggambarkan model pohon dimana nilai prediksi pada t diumpamakan $\hat{y}_i^{(t)}$ dengan Persamaan 6.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \quad (6)$$

Dimana $\hat{y}_i^{(t)}$ merupakan model *tree* terakhir, $\hat{y}_i^{(t-1)}$ adalah model pohon yang dihasilkan sebelumnya, $f_t(x_i)$ merupakan model baru yang dibuat sedangkan t adalah jumlah total model *tree* yang dibangun dari *base tree models*. Pada algoritma XGBoost sangat penting untuk dapat menentukkan banyaknya *tree* dan *depth* yang dimilikanya. Permasalahan dalam menemukan algoritma yang optimum dapat diubah dengan pencarian klasifikasi baru yang dapat mengurangi fungsi *loss* dengan target fungsi kerugian ditunjukkan pada Persamaan 7 yang menggambarkan *learning function* (Obj) berikut.

$$Obj(t) = \sum_{i=1}^{t} l(\hat{y}_i^t, y_i) + \sum_{i=1}^{t} \Omega(f_i) \quad (7)$$

Dimana $\hat{y}_i^{(t)}$ merupakan nilai prediksi dan y_i adalah nilai aktual dan l adalah fungsi *loss* dan Ω adalah fungsi regularisasi, dimana Ω adalah

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$$

Dengan λ merupakan parameter regularisasi yang memiliki nilai *default* sebesar 1 dan T adalah jumlah *'leaves'* yang ada pada *Tree*. Sedangkan Ω digunakan untuk menentukan kompleksitas yang dimiliki model dengan menentukan nilai γ, ω adalah bobot *leaves* yang digunakan sebagai *output value*.

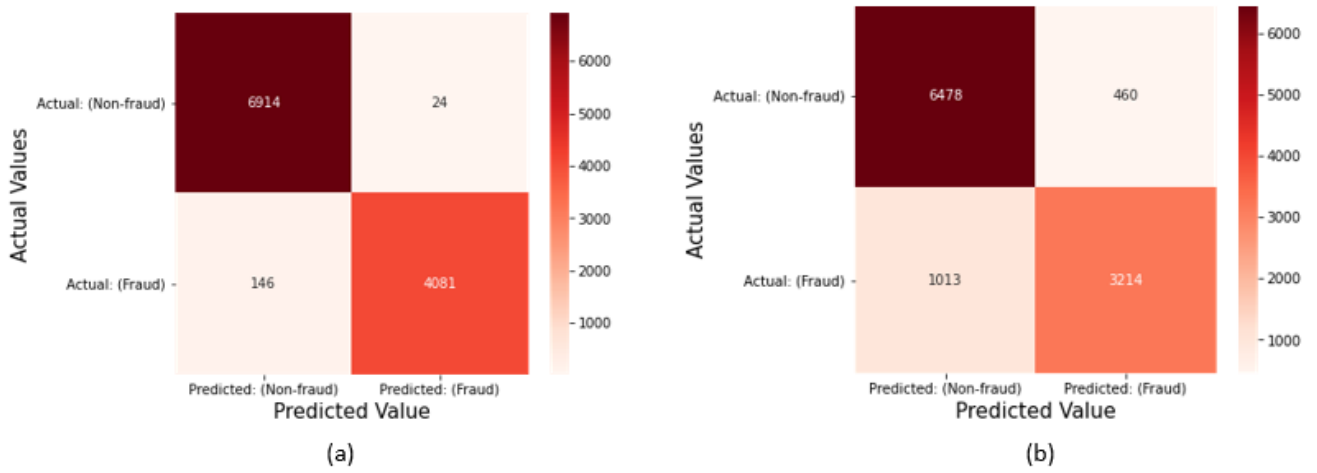
III. METODE PENELITIAN

A. *Deskripsi Data*

Data yang digunakan didapat dari *Kaggle.com* yang terdiri dari empat dataset untuk data latih dan data uji. Data yang pertama adalah data dari penyedia layanan kesehatan yang berpotensi melakukan *fraud*. Data selanjutnya dataset yang menyediakan detil riwayat penyakit dan detail pembayaran asuransi dari *beneficiary* atau pasien. Data ketiga adalah dataset yang menyediakan klaim dari pasien yang masuk kedalam rumah sakit untuk pengobatannya atau bisa disebut rawat inap. Data keempat adalah dataset yang menyediakan klaim dari pasien yang datang ke rumah sakit namun tidak dimasukkan kedalam rumah sakit untuk pengobatannya atau bisa disebut rawat jalan.

B. *Pra Proses Data*

Setelah data didapatkan dan dieksplorasi, dilakukan praproses data seperti *merging data* untuk menggabungkan dataset-dataset yang didapatkan, proses pembersihan data



Gambar 3. *Confusion matrix* (a) XGBoost, (b) SVM data latih 80% data uji 20%.

Tabel 1.
Hasil uji coba perbandingan klasifikasi pada komposisi data 80% data latih 20% data uji

Percobaan	BACC	F-1 Score	Precision	Recall
XGBoost	0.981	0.979	0.997	0.979
SVM	0.874	0.87	0.902	0.854

Tabel 2.
Hasil uji coba perbandingan klasifikasi pada komposisi data 70% data latih 30% data uji

Percobaan	BACC	F-1 Score	Precision	Recall
XGBoost	0.979	0.977	0.995	0.977
SVM	0.864	0.859	0.901	0.84

dari data yang memiliki nilai kosong dan juga perubahan nilai dengan merubah data menjadi bentuk *binary* yaitu '1,0', proses *feature engineering* dengan membuat fitur baru dari fitur yang sudah ada dengan tujuan menambah iformasi yang didapatkan dari dataset, proses *encoding data* yang bertujuan untuk mengubah data berbentuk kategori menjadi data bilangan, lalu *splitting data* dengan tujuan memisah data yang dilatih dan diuji, lalu dilakukan proses *oversampling* untuk menanggulangi permasalahan bahwa data tidak seimbang, terakhir dilakukan proses normalisai dengan tujuan merubah nilai fitur memiliki rentang yang sama.

C. Proses Klasifikasi Menggunakan Support Vector Machine

Pada tahapan ini dilakukan proses klasifikasi SVM pada data asuransi pelayanan kesehatan. Klasifikasi metode SVM bekerja dengan membagi kelas data menjadi dua yaitu data *fraud* atau data tidak *fraud* dengan cara mencari *hyperplane* terbaik untuk memisah kedua kelas tersebut. Pada proses klasifikasi digunakan klasifikasi SVM pada data asli, *oversampling*.

D. Proses Klasifikasi Menggunakan XGBoost

Pada tahapan ini dilakukan proses klasifikasi menggunakan metode *Extreme Gradient Boosting*. Hasil hasil dari berbagai kriteria observasi data yang telah dilakukan menggunakan metode SVM dibandingkan nilai performansi dengan metode XGBoost untuk melihat metode yang lebih baik digunakan pada data klaim layanan kesehatan untuk mendeteksi kecurangan.

E. Analisis dan Evaluasi Hasil Prediksi

Pada tahapan ini, prediksi yang dihasilkan oleh metode SVM dianalisis menggunakan *Confusion Matrix* dan dihitung nilai *Balanced Accuracy*, *Precision*, *Recall* dan nilai F-1. Analisis dilakukan untuk melihat seberapa baik SVM dan XGBoost dalam mengklasifikasi dalam mendeteksi *fraud* pada data asuransi kesehatan.

IV. HASIL DAN PEMBAHASAN

A. Deskripsi Uji Coba

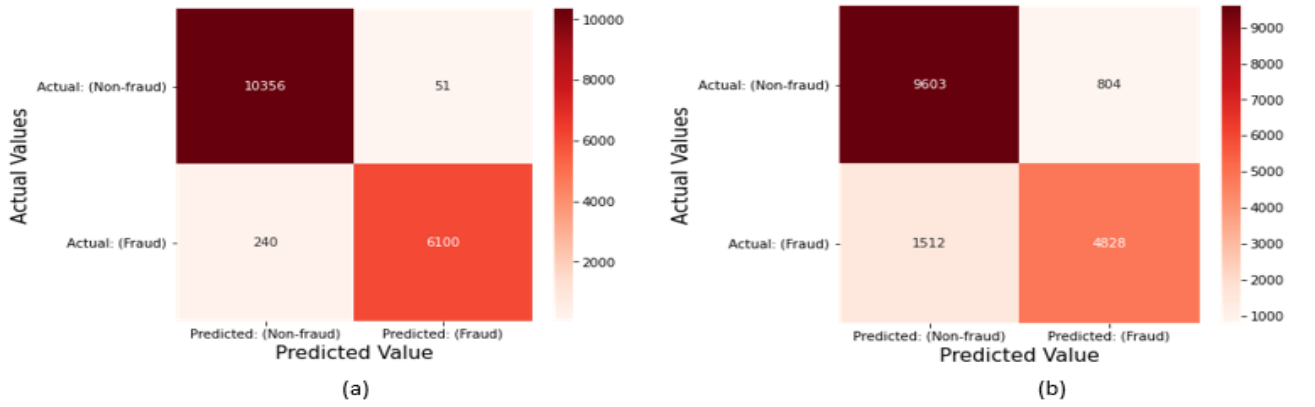
Data uji coba yang digunakan dalam penelitian ini yaitu data asli setelah normalisasi untuk setiap model klasifikasi SVM dan XGBoost. Untuk setiap percobaan pada data uji coba dilakukan uji coba sebanyak 3 kali dengan komposisi data berbeda yaitu data latih 80% dengan 20% data uji, data latih 70% dengan 30\% data uji, dan data latih 60% dengan 40% data uji.

B. Hasil Uji Coba Komposisi Data 80% Data Latih 20% Data Uji

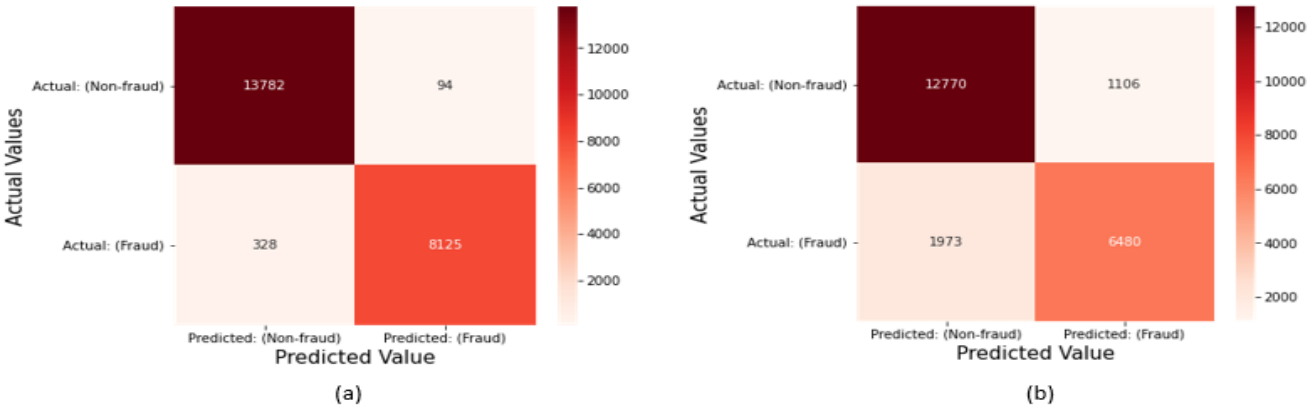
Pada bagian ini dilakukan uji coba klasifikasi kelas *fraud* pada data menggunakan metode SVM dan dibandingkan dengan metode XGBoost dengan perbandingan data latih 80\% data dan data uji 20\% data. Dari contoh *confusion matrix* pada Gambar 3 dengan nilai positif adalah pada klaim data *non-fraud* dan nilai negatif-nya adalah klaim data *fraud* dihitung nilai *Balanced Accuracy*, nilai *Recall*, nilai *Precision*, *F-1 Score* pada hasil klasifikasi XGBoost dan SVM. Berikut merupakan hasil metrik performansi yang didapatkan pada Tabel 1.

Dari Tabel 1 didapatkan metode XGBoost pada data memiliki hasil klasifikasi yang lebih baik dibandingkan dengan metode SVM. Dimana nilai BACC, *F-1 Score*, *Precision*, dan *Recall*} pada XGBoost berturut-turut adalah 0.981, 0.979, 0.997, dan 0.979, sedangkan hasil yang didapatkan klasifikasi SVM berturut-turut adalah 0.874, 0.87, 0.902, 0.854.

Dari Tabel 1 tersebut dan informasi yang didapatkan diketahui bahwa seluruh nilai metrik performansi dari hasil uji coba klasifikasi menggunakan metode SVM memiliki nilai performa yang lebih rendah dibandingkan hasil uji coba menggunakan metode XGBoost. Dapat diketahui bahwa dengan komposisi data dengan 80\% data latih dan 20\% data



Gambar 4. *Confusion matrix* (a) XGBoost, (b) SVM data latih 70% data uji 30%.



Gambar 5. *Confusion matrix* (a) XGBoost, (b) SVM data latih 60% data uji 40%.

Tabel 3.
Hasil uji coba perbandingan klasifikasi pada komposisi data 60% data latih 40% data uji

Percobaan	BACC	F-1 Score	Precision	Recall
XGBoost	0.977	0.975	0.993	0.977
SVM	0.864	0.859	0.897	0.842

uji metode XGBoost lebih cocok digunakan pada data.

C. Hasil Uji Coba Komposisi Data 70% Data Latih 30% Data Uji

Pada bagian ini dilakukan uji coba klasifikasi kelas *fraud* pada data menggunakan metode SVM dan dibandingkan dengan metode XGBoost dengan perbandingan data latih 70% data dan data uji 30% data. Dari contoh *confusion matrix* hasil klasifikasi pada Gambar 4 dihitung nilai *Balanced Accuracy, Recall*, *Precision*, *F-1 Score* pada hasil klasifikasi XGBoost dan SVM. Berikut Tabel 2 yang menunjukkan hasil perbandingan metrik performansi tersebut.

Tabel 2 menunjukkan hasil yang tidak jauh berbeda dengan Tabel 1 dimana didapatkan hasil klasifikasi metode XGBoost lebih baik dibandingkan dengan metode klasifikasi SVM. Dimana nilai BACC, *F-1 Score*, *Precision*, dan *Recall* pada XGBoost berturut-turut adalah 0.979, 0.977, 0.995, dan 0.977, sedangkan hasil yang didapatkan klasifikasi SVM berturut-turut adalah 0.864, 0.859, 0.901, 0.84.

Dari Tabel 2 tersebut dan informasi yang didapatkan diketahui sebagian besar hasil uji coba dari klasifikasi menggunakan metode SVM memiliki nilai performa yang lebih rendah dibandingkan hasil uji coba yang didapatkan menggunakan metode XGBoost. Berdasarkan hasil tersebut dapat diketahui bahwa dengan komposisi data dengan 70\% data latih dan 30\% data uji metode XGBoost lebih cocok digunakan pada data.

D. Hasil Uji Coba Komposisi Data 60% Data Latih 40% Data Uji

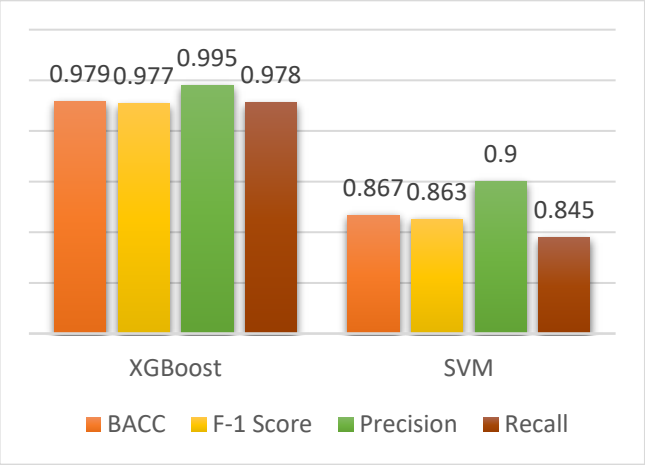
Pada bagian ini dilakukan uji coba klasifikasi kelas *fraud* pada data menggunakan metode SVM dan dibandingkan dengan metode XGBoost dengan perbandingan data latih 60% data dan data uji 40% data.

Dari contoh *confusion matrix* hasil klasifikasi pada Gambar 5 dihitung nilai *Balanced Accuracy*, nilai *Recall*, nilai *Precision*, *F-1 Score* menggunakan persamaan (2.15) pada hasil klasifikasi XGBoost dan SVM. Tabel 3 memperlihatkan hasil perbandingan klasifikasi XGBoost dan SVM dengan data latih 60% data dan data uji 40% data.

Pada Tabel 3 tersebut dapat terlihat hasil perbandingan yang didapat juga tidak jauh berbeda dengan Tabel 1 dan Tabel 2, dimana nilai *Balanced Accuracy*, *F-1 Score*, *Precision*, dan *Recall* pada klasifikasi XGBoost memiliki nilai yang lebih tinggi dibandingkan klasifikasi SVM. Dimana nilai BACC, *F-1 Score*, *Precision*, dan *Recall* pada XGBoost berturut-turut adalah 0.977, 0.975, 0.993, dan 0.977, sedangkan hasil yang didapatkan klasifikasi SVM berturut-turut adalah 0.864, 0.859, 0.897, 0.842.

E. Pembahasan

Pada bagian ini akan ditunjukkan pembahasan berdasarkan hasil klasifikasi yang sudah dilakukan pada subbab sebelumnya didapatkan hasil perbandingan dari metode SVM dan XGBoost pada klasifikasi dua kelas dalam deteksi *fraud* data klaim sistem pelayanan kesehatan. Dari hasil uji coba tersebut diketahui bahwa nilai metrik performansi yang sama pada ketiga uji coba dengan komposisi data latih dan data uji yang berbeda, dari hasil tersebut dapat disimpulkan bahwa



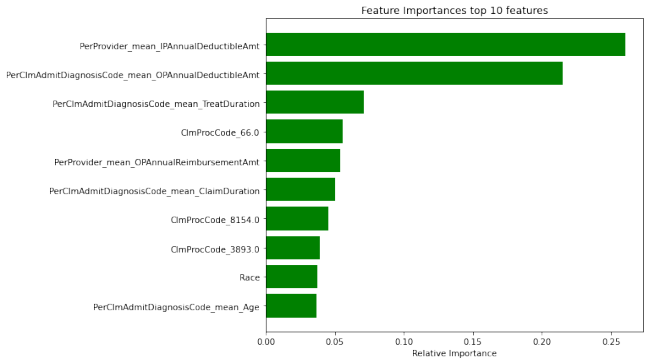
Gambar 6. Perbandingan rata-rata performansi hasil percobaan.

kemampuan dalam *generate* model untuk setiap metode cukup baik di berbagai komposisi data. Oleh karena itu nilai nilai tersebut didapatkan nilai rata-ratanya untuk melihat perbandingan pada setiap uji coba.

Pada Gambar 6 diperlihatkan perbandingan rata-rata nilai metrik performansi dari tiga komposisi data latih dan data uji. Terlihat bahwa semua uji coba klasifikasi metode XGBoost memiliki tingkat performansi yang lebih tinggi dibandingkan dengan SVM. Namun pada penelitian ini penentuan model yang terbaik dengan memfokuskan pada nilai *Balanced Accuracy* dan *Recall*. *Balanced Accuracy* dapat memperlihatkan secara umum bagaimana performa model pada data sedangkan *Recall* sangat penting dikarenakan *Recall* pada penelitian ini merepresentasikan seberapa besar nilai *missclassification* pada klaim data *fraud* yang terklasifikasi *non-fraud* atau bisa disebut *false negative* yang di dalam kehidupan sehari-hari jauh lebih berbahaya dibandingkan *missclassification* pada klaim data *non-fraud* terklasifikasi *fraud* atau bisa disebut *false positive* yang masih dapat dilakukan pemeriksaan kembali berbeda dengan *false negative* yang apabila dibiarkan dapat melanjutkan kegiatan *fraud*-nya yang dapat merugikan berbagai pihak.

Sehingga dengan melihat pada Gambar 6 model XGBoost pada data klaim pelayanan kesehatan memiliki nilai *Balanced Accuracy* dan *Recall* tertinggi adalah model yang dapat dengan sangat baik memprediksi klaim data *fraud* dan klaim data *non-fraud*.

Untuk melihat pola yang dilakukan penyedia layanan kesehatan, pasien, asuransi kesehatan, ataupun dokter yang melakukan tindak kecurangan maka dilakukan proses *feature importance* untuk melihat fitur apa yang berperan penting atas klasifikasi kecurangan tersebut. Pada Gambar 7 ditunjukkan sepuluh teratas *feature importance* pada dataset. Fitur yang pertama adalah fitur yang menjelaskan rata-rata jumlah yang harus dibayarkan pasien rawat jalan setiap tahun untuk setiap penyedia layanan kesehatan, semakin besar nilainya maka semakin ada kemungkinan terjadinya kecurangan. Fitur selanjutnya adalah rata-rata jumlah yang harus dibayarkan pasien rawat inap setiap tahun untuk setiap kelompok diagnosis penyakit, contohnya adalah diagnosis untuk proses melahirkan dan penyakit gagal jantung, semakin tinggi nilai semakin ada kemungkinan terjadinya. Dari hasil observasi tersebut dapat dilakukan pertimbangan kebijakan yang dikeluarkan pemerintah maupun lembaga penyedia layanan kesehatan dalam melakukan kebijakan untuk



Gambar 7. Feature importance pada model klasifikasi.

mendeteksi adanya tindak *fraud.*

V. KESIMPULAN/RINGKASAN

A. *Kesimpulan*

Metode *Support Vector Machine* dan *Extreme Gradient Boosting* dapat digunakan untuk klasifikasi data *fraud* pada data klaim layanan kesehatan dengan melakukan pra proses pembersihan data, *feature engineering*, melakukan proses *encoding one-hot* pada beberapa fitur kategorikal, *oversampling* pada data dikarenakan kelas target memiliki jumlah yang tidak seimbang untuk klasifikasi dengan SVM, melakukan normalisasi data dengan mencari nilai *Z-Score*, lalu mengklasifikasi data menjadi dua kelas berbeda yaitu kelas *fraud* dan *non-fraud* dengan menggunakan metode SVM dan XGBoost. Hasil perbandingan klasifikasi SVM dan XGBoost didapatkan bahwa metode XGBoost lebih baik dibandingkan dengan SVM dilihat dari nilai *Balanced Accuracy* dan *Recall*-nya terbaik didapatkan dengan menggunakan metode XGBoost dengan nilai 0.98 dan 0.984 pada data dibandingkan SVM dengan nilai 0.874 dan 0.854.

B. *Saran*

Beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya yaitu Dapat menggunakan model *deep learning* untuk menemukan nilai akurasi yang lebih baik, dapat menggunakan metode *encoding* yang lebih cocok pada model seperti *Response Encoding* untuk menemukan nilai akurasi yang lebih baik, dapat menggunakan perangkat keras GPU untuk mendapatkan hasil dengan *running time* yang lebih cepat.

Lalu saran bagi penyedia layanan kesehatan diantaranya adalah lembaga penyedia layanan kesehatan dapat melakukan evaluasi pada pembayaran pasien melalui asuransi yang dilakukan untuk satu tempat layanan saja, apabila jumlahnya sangat banyak ada kemungkinan terjadi praktik kecurangan melibatkan penyedia layanan kesehatan dengan pihak asuransi, dan lembaga penyedia layanan kesehatan dapat melakukan evaluasi pada pembayaran pasien rawat inap melalui asuransi mengenai diagnosis kelompok seperti pasien melahirkan dan yang lain lain, apabila jumlahnya sangat banyak ada kemungkinan terjadi praktik kecurangan melibatkan penyedia layanan kesehatan dengan pihak asuransi.

DAFTAR PUSTAKA

[1] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *J. Big Data*, no. 1, 2019, doi: 10.1186/s40537-019-0217-0.

[2] H. Joudaki *et al.*, "Improving fraud and abuse detection in general physician claims: a data mining study," *Int. J. Heal. Policy Manag.*, vol. 5, no. 3, pp. 165–172, 2015.

[3] Media Internal Resmi BPJS Kesehatan, "Tindak Kecurangan (Fraud) Merugikan Program JKN (Negara)," *INFOBPJS Kesehatan*, Jakarta, pp. 2–6, Nov. 2015. [Online]. Available: https://persi.or.id/wp-content/uploads/2020/11/fraud_majalahbpjs.pdf

[4] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using machine learning methods," *2017 16th IEEE Int. Conf. Mach. Learn. Appl.*, pp. 858–865, 2017, doi: 10.1109/ICMLA.2017.00-48.

[5] S. Russell and P. Norvig, *Artificial Intelligence, Global Edition A Modern Approach*, 4th ed. London: Pearson Education, 2021.

[6] Z. Lv and L. Qiao, "Analysis of healthcare big data," *Futur. Gener. Comput. Syst.*, vol. 109, pp. 103–110, 2020, doi: 10.1016/j.future.2020.03.039.

[7] R. A. Sowah *et al.*, "Decision support system (DSS) for fraud detection in health insurance claims using genetic support vector machines (GSVMs)," *J. Eng.*, vol. 2019, pp. 1–19, 2019, doi: 10.1155/2019/1432597.

[8] S. Suthaharan, *Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification*. Boston: Springer, 2016.

[9] C. Meng, L. Zhou, and B. Liu, "A case study in credit fraud detection with SMOTE and XGBoost," *J. Phys. Conf. Ser.*, vol. 1601, no. 5, p. 052016, 2020, doi: 10.1088/1742-6596/1601/5/052016.

[10] Global Fraud Study, *Report to The Nations on Occupational Fraud and Abuse*. Texas: Association of Certified Fraud Examiners, 2016.

[11] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, 2016, doi: 10.1016/j.jnca.2016.04.007.

[12] Q. Liu and M. Vasarhelyi, "Healthcare Fraud Detection : A Survey and A Clustering Model Incorporating Geolocation Information," in *29th World Continuous Auditing and Reporting Symposium (29WCARS)* 2013.

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.