DETEKSI POTENSI *INFLATED BILLS* PADA DATA KLAIM LAYANAN BPJS MENGGUNAKAN ALGORITMA *EXTREME GRADIENT BOOSTING* (*XGBOOST*)

Kayla Alysa Adra

Program Studi Teknologi Informasi

Fakultas Ilmu Komputer dan Teknologi Informasi

Universitas Sumatera Utara

Abstract

Inflated bills or bills that exceed the actual cost are a common type of fraud happen in advanced referral health facilities in the implementation of bpjs health program. Such fraud can lead to financial losses for BPJS Health, trigger budget deficits, and disrupt the sustainability of the National Health Insurance (JKN) program. This study aims to develop a system that capable to automatically detecting potential inflated bills to prevent such budget deficits. The data used consists of a sample of health service claims in FKRTL from 2016 to 2022, comprising 1,176,438 rows. The data is divided into training, validation, and testing subsets with 25 variables selected based on the Indonesian Minister of Health Regulation No. 3 of 2023 about Healthcare Tariff Standards In the JKN Program, The parameter turning process was conducted 48 times, testing the combination of n estimators, learning rate, and max depth parameters, The best combination achieved was n estimators at 100, leaning rate at 0,1 and max depth at 8 with Mean Absolute Error (MAE) of 0,005 on the training data and 0,008 on the validation data. The testing result in this study

indicate that the model achieved an MAE of 0,008, MSE 0,015, and RMSE of 0,12 on the testing data. These values demonstrate that the model has a low error rate in predicting the verification costs and can be used as a tool for detecting potential inflated bills.

***Keywords**: Inflated Bills, BPJS Health, Extreme Gradient Boosting.*

JEL Classification **:** I13 (*Health Insurance*), C53(*Prediction Methods*).

I.Pendahuluan

1.1 Latar Belakang

Setiap warga negara Indonesia memiliki hak untuk memperoleh layanan kesehatan yang layak, yang dijamin melalui program Jaminan Kesehatan Nasional (JKN) yang diselenggarakan oleh BPJS Kesehatan. Namun, dalam implementasinya, program ini menghadapi tantangan serius, salah satunya adalah praktik kecurangan (*fraud*) yang dilakukan oleh fasilitas Kesehatan. Salah satu bentuk kecurangan yang terjadi adalah *inflated bills*, yaitu tindakan mengenakan biaya lebih tinggi dari orang yang seharusnya untuk layanan medis dan alat Kesehatan (Gede et al.,2024). Praktik ini menjadi salah satu penyebab deficit keuangan BPJS Kesehatan yang pada tahun 2023 lalu mencapai Rp 7,9 Triliun (Kompas,2024).

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah mendeteksi potensi inflated bills pada data klaim layanan BPJS Kesehatan menggunakan algoritma *Extreme Gradient Boosting* (XGBoost).

II. Studi Literatur

2.1 *Inflated Bills* dalam Layanan Kesehatan

Inflated bills didefinisikan sebagai pengajuan tagihan yang secara sengaja dilebih-lebihkan dari biaya actual layanan kesehatan yang diberikan(Sarah et al., 2021). Dalam konteks BPJS Kesehatan, praktik ini termasuk dalam kategori kecurangan (*fraud*) yang diatur dalam peraturan Menteri Kesehatan Nomor 16 tahun 2019. Kecurangan ini tidak hanya merugikan secara finansial tetapi juga mengancam sustainability program JKN (Raisa et al., 2021).

2.2 Deteksi Kecurangan dengan *Machine Learning*

Berbagai penelitian terdahulu telah mengeksplorasi penggunaan *machine learning* untuk mendeteksi kecurangan dan inefisiensi dalam klaim asuransi kesehatan. Kevin er al. (2023) menggunakan *Light Gradient Boosting Machine* (LGBM) untuk memprediksi inefisiensi klaim, menghasilkan *Area Under Curv*e (AUC) sebesar 0,75. Rofiq (2023) menemukan bahwa kombinasi algoritma *Random Forest* dengan teknik Tomek *Links* menghasilkan *F-1Score* terbaik sebesar 19,53. Sementara itu, Alan & Mohammad (2023) membandingkan *Support Vector Machine* (SVM) dan *XGBoost*, dimana *XGBoost* unggul dengan *recall* 0,9994 dan *balanced accuracy* 0,9995. Keunggulan *XGBoost* dalam menangani data besar, *missing value*, dan mencegah *overfitting* melalui regulasi menjadikanya pilihan yang kuat untuk tugas deteksi kecurangan (Chen & Guestrin, 2016; Jafar et al., 2020).

2.3 *Extreme Gradient Boosting* (XGBoost)

XGboost adalah algoritma *ensemble learning* berbasis *boosting* yang bekerja dengan membangun pohon Keputusan secara sekuensial. Setiap pohon berikutnya belajar untuk memberbaiki kesalahan dari pohon sebelumnya (Linggar & Adi, 2022). Prediksi akhir merupakan penjumlahan tertimbang dari semua pohon.

Tabel 1 parameter *XG Boost*

Parameter	Kegunaan
n_estimators	Menentukan jumlah pohon keputusan yang dibangun
learning_rate	Mengatur langkah pembaruan bobot selama pelatihan
max_depth	Menentukan kedalaman maksimum setiap pohon
min_child_weight	Menetapkan bobot minimum yang diperlukan dalam node anak
gamma	Mengontrol pengurangan kerugian yang diperlukan untuk pemisahan
subsample	Mengatur proporsi data pelatihan yang digunakan untuk setiap pohon
colsample_bylevel	Mengatur rasio fitur yang digunakan dalam membangun pohon

Fungsi objektif *XGBoost* menggabungkan fungsi *loss* dan komponen regularisasi :

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} \quad [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \qquad (1)$$

dimana:

$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ (gradien)

$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ (hessian)

$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \quad w_j^2$ (regularisasi)

2.4 Metrik Evaluasi

Untuk evaluasi model regresi digunakan tiga metrik utama :

a. *Mean Squared Error* (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2 \tag{2}$$

b. *Root Mean Squared Error* (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2} \tag{3}$$

c. *Mean Absolute Error* (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \quad | \, y_i - \hat{y}_i \, | \tag{4}$$

III. Data dan Metodologi

3.1 Sumber Data dan Karakteristik

Data yang digunakan adalah dataset sampel histori klaim pelayanan di Fasilitas Kesehatan Rujukan Tingkat Lanjut (FKRTL) untuk periode 2016 hingga 2022, yang dapat diakses melalui portal data JKN.

	PSTV01	PSTV02	PSTV15	FKP02	FKL02	FKL03	FKL04	FKL05	FKL06	FKL07	...	FKL39	FKL40	FKL
0	180205299	180205298	3.662519		100080122V000022	2022-01-03	2022-01-03	PAPUA	JAYAPURA	Pemerintah kab/kota	...		0	NON
1	314289762	314289762	14.453729		100080122V000028	2022-01-02	2022-01-02	PAPUA	JAYAPURA	Pemerintah kab/kota	...		0	NON
2	385562166	314289762	15.544407		100080122V000029	2022-01-02	2022-01-02	PAPUA	JAYAPURA	Pemerintah kab/kota	...		0	NON
3	69505864	69505864	2.836411		100080122V000183	2022-01-05	2022-01-05	PAPUA	JAYAPURA	Pemerintah kab/kota	...		0	NON
4	135231252	135231252	3.361672		100080122V000263	2022-01-07	2022-01-07	PAPUA	JAYAPURA	Pemerintah kab/kota	...		0	NON
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1176433	20647133	20647133	14.917422	345150922Y000546	99591122V000223	2022-11-24	2022-11-24	DKI JAKARTA	KEPULAUAN SERIBU	Pemerintah provinsi	...		0	NON
1176434	71048782	32664628	3.991986	99531122Y000575	99591122V000234	2022-11-25	2022-11-25	DKI JAKARTA	KEPULAUAN SERIBU	Pemerintah provinsi	...		0	NON
1176435	60464841	60464841	74.061844	345151122Y000751	99591122V000238	2022-11-25	2022-11-25	DKI JAKARTA	KEPULAUAN SERIBU	Pemerintah provinsi	...		0	NON
1176436	466097827	60655067	5.043636	99551222Y000104	99591222V000136	2022-12-14	2022-12-14	DKI JAKARTA	KEPULAUAN SERIBU	Pemerintah provinsi	...		0	NON
1176437	68356963	68356963	18.174042	345151222Y000683	99591222V000245	2022-12-29	2022-12-29	DKI JAKARTA	KEPULAUAN SERIBU	Pemerintah provinsi	...		0	NON

Gambar 1 Dataset Sampel FKRTL 2016-2022

3.2 Variabel Penelitian

Berdasarkan Peraturan Menteri Kesehatan RI No.3 Tahun 2023 tentang Standar Tarif Pelayanan Kesehatan dalam Program JKN, dipilih 25 variabel yang relevan dengan penentuan biaya.

Tabel 2 Informasi Variabel yang digunakan

No	Variabel	Keterangan	Landasan Hukum
1	FKL09	Tipe FKRTL	Pasal 28 Ayat 1
2	FKL10	Tingkat pelayanan di FKRTL	Pasal 28 Ayat 1
3	FKL13	Kelas iuran	Pasal 28 Ayat 3
4	FKL17	Kode dan nama diagnosis primer ICD	Lampiran Permenkes
5	FKL18	Kode ICD-10 diagnosis primer FKRTL	Lampiran Permenkes

6	FKL20	INACBGs - Kode Casemix Main Groups	Pasal 27 Ayat 1
7	FKL21	INACBGs - Tipe kelompok kasus atau case groups	Pasal 27 Ayat 1
8	FKL22	INACBGs - Spesifikasi kelompok kasus	Pasal 27 Ayat 1
9	FKL23	INACBGs - Tingkat keparahan kasus	Pasal 27 Ayat 1
10	FKL30	Jenis prosedur yang diberikan	Lampiran Permenkes
11	FKL31	Tarif regional	Pasal 28 Ayat 2
12	FKL32	Group tarif regional	Pasal 28 Ayat 2
13	FKL33	Kode special sub acute (SA)	Pasal 27 Ayat 2
14	FKL34	Tarif special sub acute (SA)	Pasal 27 Ayat 2
15	FKL35	Kode special procedure (SP)	Pasal 27 Ayat 2
16	FKL37	Tarif special procedure (SP)	Pasal 27 Ayat 2
17	FKL38	Kode special prosthesis (RR)	Pasal 27 Ayat 2
18	FKL40	Tarif special prosthesis (RR)	Pasal 27 Ayat 2
19	FKL41	Kode special investigation (SI)	Pasal 27 Ayat 2
20	FKL43	Tarif special investigation (SI)	Pasal 27 Ayat 2
21	FKL44	Kode special drugs (SD)	Pasal 27 Ayat 2
22	FKL46	Tarif special drugs (SD)	Pasal 27 Ayat 2
23	FKL47	Biaya tagih oleh fasilitas kesehatan	-
24	LOS	Length of Stay (lama masa rawat)	Lampiran Permenkes
25	FKL48	Biaya verifikasi oleh BPJS Kesehatan (target)	-

3.3 Tahapan *Preprocessing* Data

3.3.1 *Data Cleaning*

Tabel 3 Perbandingan Sebelum dan Sesudah *Data Cleaning*

Metrik	Sebelum Cleaning	Sesudah Cleaning
Jumlah baris	1.176.438	723.476
Jumlah kolom	55	26
Missing values	2	0
Duplicated values	452.960	0

Proses pembersihan data dilakukan melalui serangkaian tahapan algoritmik yang sistematis, dimulai dengan inisialisasi daftar kolom yang akan dihapus yang mencakup variabel-variabel tidak relevan seperti data identitas personal dan informasi geografis, dilanjutkan dengan penghapusan kolom-kolom tersebut untuk menyederhanakan struktur *dataset*. Selanjutnya, semua string kosong diganti dengan '*NONE*' untuk menjaga konsistensi format data, kemudian dilakukan penghapusan baris yang mengandung nilai hilang dan seluruh baris data terduplikasi untuk memastikan integritas data dan mencegah bias dalam analisis. Tahap akhir proses ini adalah mereset ideks *dataset* untuk menjamin penomoran baris yang konsisten sebelum data yang telah dibersihkan dikembalikan untuk diproses pada tahapan berikutnya, sehingga menghasilkan dataset optimal.

3.3.2. *Feature Engineering*

Proses feature engineering menghasilkan fitur baru dan menyaring data:

Tabel 4 Hasil *Feature Enggineering*

Metrik	Sebelum *Engineering*	Sesudah *Engineering*

Jumlah baris	723.476	251.341
Jumlah kolom	26	25

Fitur *Length of Stay* (LOS) dibuat dengan rumus:

$$LOS = FKL04 - FKL03 \quad (4)$$

Dimana FKL04 adalah tanggal pulang dan FKL03 adalah tanggal datang., Distribusi Data Duplikat**:**

a. Data duplikat: 65.3%

b. Data non-duplikat: 34.7%

3.3.3 *Label Encoding*

Proses transformasi data kategorikal ke dalam format numerik dilakukan menggunakan Teknik *Label Encoding* untuk memungkinkan pemrosesan oleh algoritma *Extreme Gradient Boosting*, dimana setiap nilai unik dalam variabel kategorik dikonversi menjadi representasi bilangan bulat yang terurut.

Tabel 5 Contoh *Label Encoding*

Nilai Asli	Nilai Setelah *Encoding*
Rawat Jalan Tingkat Lanjut (RJTL)	0
Rawat Inap Tingkat Lanjut (RITL)	1

3.3.4 Normalisasi Data

Menggunakan *StandardScaler* (*Z-Score Normalization):*

Rumus Normalisasi:

$$x' = \frac{x_i - \bar{x}}{\sigma} \quad (5)$$

Contoh Perhitungan:

Untuk nilai FKL22 = 13 dengan $\bar{x}_{FKL22} = 17.6$ dan $\sigma_{FKL22} = 4.963$:

$$x' = \frac{13 - 17.6}{4.963} = -0.93$$

3.4 Pembagian Dataset

Data dibagi dengan proporsi :

Tabel 6 pembagian *dataset*

Subset	Persentase	Jumlah Data
Training	70%	175.939
Validation	20%	50.268
Testing	10%	25.134
Total	100%	251.341

Pembagian *dataset* dengan proporsi 70:20:10 untuk *training, validation,* dan *testing* dipilih berdasarkan pertimbangan yang lazim dalam pengembangan model *machine learning*. Proporsi 70% untuk *data training* dinilai cukup untuk memastikan model mempelajari pola data dengan komprehensif. Sementara itu, alokasi 20% untuk *data validation* memadai untuk melakukan evaluasi yang stabil selama proses parameter *tuning* tanpa mengurangi *data training* secara signifikan. Subset *testing* 10% dipertahankan sebagai data yang benar-benar independen dan belum tersentuh proses

development untuk menguji kemampuan generalisasi model akhir secara tidak bias (Brownlee, 2020). Pembagian ini juga mempertimbangkan ukuran *dataset* yang besar (>250.000 sampel), sehingga subset testing 10% sudah mewakili lebih dari 25.000 sampel yang dianggap cukup untuk evaluasi final.

3.5 Proses *Training* dan Parameter *Tuning*

Dilakukan pencarian parameter optimal menggunakan metode *Grid Search* dengan 48 kombinasi parameter. *Grid Search* dipilih karena kemampuanya untuk mengeksplorasi semua kemungkinan kombinasi dalam ruang pencarian yang telah ditentukan, sehingga memastikan ditemukanya konfigurasi terbaik untuk model *XGBoost* pada dataset ini (Chen &Guestrin, 2016). Kombinasi parameter yang diuji adalah:

Kombinasi Parameter:

a. *n_estimators*: [50, 100, 200, 300]

b. *learning_rate*: [0.01, 0.05, 0.1]

c. *max_depth*: [2, 4, 6, 8]

Rumus Prediksi *XGBoost*:

$$\hat{y}_i = \hat{y}_i^{(0)} + \eta \sum_{k=1}^{K} \quad f_k(x_i) \tag{6}$$

dimana:

$\hat{y}_i^{(0)}$ = prediksi awal (rata-rata target)

η = *learning_rate*

f_k = pohon ke-k

Contoh Perhitungan *Training*:

Dengan 3 sampel *training* dan *target values*: [-0.1199, -0.139, 2.81]

$$\text{Base prediction: } \hat{y}^{(0)} = \frac{-0.1199+(-0.139)+2.81}{3} = 0.85 \quad (7)$$

Gradient: $g = \hat{y}^{(0)} - y = [0.85 - (-0.1199), 0.85 - (-0.139), 0.85 - 2.81] = [0.97, 0.99, -1.96]$

Proses *boosting* berlanjut hingga konvergensi dengan menambahkan pohon baru yang memperbaiki residual.

IV. Hasil dan Pembahasan

4.1 Hasil Parameter Tuning

Proses *tuning parameter* dilakukan melalui 48 kombinasi berbeda menghasilkan konfigurasi optimal pada *n_estimators* = 100*, learning_rate* = 0,1 dan *max_depth* = 8, Kombinasi ini menghasilkan nilai *Mean Absolute Error* (MAE) sebesar 0,00597 pada *data training* dan 0,00867 pada *data validation*.

Tabel 7 Parameter Optimal

Parameter	Nilai Optimal	Keterangan
n_estimators	100	Jumlah pohon optimal
learning_rate	0.1	Laju pembelajaran
max_depth	8	Kedalaman pohon

4.2 Perhitungan Metrik Evaluasi

pada data testing menunjukan performa model yang konsisten dengan nilai MAE 0,008, MSE 0,015, dan RMSE 0,12. Perhitungan metrik evaluasi berdasarkan data testing dengan n = 25.134 adalah sebagai berikut :

a. *Mean Absolute Error* (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \quad | \, y_i - \hat{y}_i \, | = \frac{222.82}{25.134} = 0.00886 \approx 0.008 \tag{8}$$

b. *Mean Squared Error* (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2 = \frac{389.05}{25.134} = 0.01547 \approx 0.015 \tag{9}$$

c. *Root Mean Squared Error* (RMSE):

$$RMSE = \sqrt{MSE} = \sqrt{0.01547} = 0.12439 \approx 0.12 \tag{10}$$

Tabel 8 Detail Perhitungan Metrik Evaluasi pada *Data Testing*

Metric	Total Error	n Samples	Formula	Result
MAE	222.82	25,134	222.82/25,134	0.00886 ≈ 0.008
MSE	389.05	25,134	389.05/25,134	0.01547 ≈ 0.015
RMSE	-	-	√0.01547	0.12439 ≈ 0.12

Total error dihitung dari akumulasi selisih absolut (MAE) dan kuadrat (MSE) antara nilai prediksi dan actual untuk seluruh 25.134 sampel testing.

4.3 *Feature Importance*

Analisis *feature importance* mengungkapkan sepuluh fitur paling berpengaruh terhadap prediksi biaya verifikasi. Biaya tagih (FKL47) menempati posisi tertinggi dengan nilai 1782.0, diikuti oleh group tarif regional (FKL32) dan tipe FKRTL (FKL09). Variabel klinis dan komponen INACBGs juga menunjukan kontribusi yang signifikan dalam model, mencerminkan struktur determinan biaya Kesehatan yang kompleks dan multidimensi.

Tabel 9 Sepuluh Fitur dengan *Nilai Feature Importance* Tertinggi

Peringkat	Fitur	Keterangan	Nilai Importance
1	FKL47	Biaya Tagih	1782.0
2	FKL32	*Group Tarif* Regional	1673.0
3	FKL09	Tipe FKRTL	1594.0
4	FKL17	Kode Diagnosis Primer ICD-10	817.0
5	LOS	*Length of Stay* (Lama Rawat)	678.0
6	FKL22	INACBGs - Spesifikasi Kelompok Kasus	670.0
7	FKL13	Kelas Iuran	565.0
8	FKL30	Jenis Prosedur	432.0
9	FKL23	INACBGs - Tingkat Keparahan	377.0
10	FKL20	INACBGs - Kode Casemix Main *Groups*	331.0

Distribusi *feature importance* ini menunjukan bahwa faktor *administrative* dan *structural* memiliki pengaruh lebih dominan dibandingkan variabel klinis murni dalam menentukan biaya verifikasi. Pola tersebut selaras dengan kerangka kebijakan Permenkes No.3 Tahun 2023 yang menempatkan komponen-komponen *administrative* sebagai penentu utama dalam standar tarif pelayanan Kesehatan.

4.4 Implementasi Deteksi *Inflated Bills*

Mekanisme deteksi *inflated bills* diimplementasikan dengan membandingkan biaya tagih aktual (FKL47) dengan prediksi biaya verifikasi ($\hat{y}$). Sistem akan menandai klaim berpotensi *inflated bills* jika memenuhi kondisi FKL47:

$$\text{Potensi } \textit{Inflated Bills} = \begin{cases} \text{Ya}, & \text{jika FKL47} > \hat{y} \\ \text{Tidak}, & \text{jika FKL47} \leq \hat{y} \end{cases}$$

Dalam implementasi operasional, dapat dipertimbangkan penerapan ambang batas (threshold) toleransi untuk menghindari false positive akibat ketidakakuratan model yang sangat kecil. Misalnya dengan kondisi: (FKL47 - ŷ) / ŷ > threshold, dimana threshold dapat ditetapkan berdasarkan konsultasi dengan domain expert BPJS Kesehatan.

Dengan model yang memiliki MAE 0,008 (0,8%), sistem ini mampu mendeteksi selisih signifikan antara biaya tagih dan prediksi biaya verifikasi, sehingga dapat menjadi alat bantu yang efektif dalam identifikasi potensi *inflated bills* pada klaim layanan BPJS Kesehatan.

V. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa Algoritma *Extreme Gradient Boosting* (XGBoost) berhasil diterapkan sebagai model prediktif untuk biaya verifikasi klaim BPJS Kesehatan dengan mencapai tingkat akurasi yang tinggi, dibuktikan dengan nilai kesalahan yang sangat rendah : MAE 0,008, MSE 0,012, dan RMSE 0,12 pada *data testing*, menggunakan kombinasi parameter optimal *n_estimators* = 100, *learning_rate* = 0,1, dan *max_depth* = 8. Oleh karena itu, sistem ini memiliki potensi besar untuk diimplementasikan sebagai alat pendeteksi otomatis potensi *inflated bills* (tagihan yang digelembungkan) untuk meningkatkan efisiensi dan efektivitas proses verifikasi klaim.

5.2 Saran

Untuk pengembangan dan implementasi lebih lanjut, direkomendasikan tiga strategi utama, yaitu : pertama, meningkatkan kualitas *input model* dengan memperluas cakupan penggunaan *dataset* yang lebih mutakhir, mencakup periode yang lebih panjang, serta mengeksplorasi penambahan variabel lain

di luar Peraturan Menteri Kesehatan No.3 Tahun 2023, seperti daya riwayat provider, yang mungkin berkolerasi dengan praktik *inflated bills*. Kedua, mengoptimalkan kinerja model dengan melakukan tuning lebih lanjut pada parameter *XGBoost* yang belum dieksplorasi (Seperti *gamma,subsample* dan *colsample_bytree*) guna memastikan akurasi prediksi maksimal dan ketiga, mengintegrasikan sistem secara bertahap ke dalam proses verifikasi klaim BPJS Kesehatan, dimulai dengan menjadikanya alat bantu (*decision support system*) sebelum beralih ke otomatisasi penuh.

VI Daftar Pustaka

Adams, F., Agsar, R., Anggoro, D., Satria, M. B., Oktavia, A. W., & Chamidah, N. (2021). *Perbandingan normalisasi data untuk klasifikasi wine menggunakan algoritma Naïve Bayes, Decision Tree, dan Support Vector Machine*. Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA), Jakarta–Indonesia.

Adiwiguna, Y. P., Purbaseno, H., Syuhara, H., Pribadi, F. J., Sarbani, A., & Kusuma, A. D. (2022). *Membangun tool prediksi yang komprehensif guna meningkatkan akurasi produk akhir pencampuran kualitas batubara menerapkan perbandingan kolaboratif pada algoritma regresi hingga neural net*. Indonesian Mining Professionals Journal, 4(2), 111–122.

Afrina, E., Cut, D., Aidha, N., Ramdlaningrum, H., Wahyudi, D., Silvia, K., Fanggidae, J., Dwi, H., Ningrum, R., Thaariq, R. M., Kartika, W., & Chrisnahutama, A. (2020). *Defisit Jaminan Kesehatan Nasional (JKN): Mengapa dan bagaimana mengatasinya?*

Alghifari, F., & Juardi, D. (2021). *Penerapan data mining pada penjualan makanan dan minuman menggunakan metode algoritma Naïve Bayes*.

Jurnal Ilmiah Informatika, 9(2), 75–81. https://doi.org/10.33884/jif.v9i02.3755

Almaliki, M. F., Ningrum, I. P., & Saputra, R. A. (2023). *Implementasi metode mesin rekomendasi user-based filtering pada sistem penyewaan alat pertambangan*. Jurnal Manajemen Informatika (Jamika), 3(1), 40–51.

Annisa, R., Winda, S., Dwisaputro, E., & Isnaini, K. N. (2021). *Mengatasi defisit dana jaminan sosial kesehatan melalui perbaikan tata kelola*. Integritas: Jurnal Antikorupsi, 6(2), 209–224. https://doi.org/10.32697/integritas.v6i2.664

Browniee, J. (2020). *Data preparation for machine learning: Data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Dharma, A. S., & Tambunan, V. (2021). *Penerapan model pembelajaran dengan metode reinforcement learning menggunakan simulator CARLA*. Jurnal Media Informatika Budidarma, 5(4), 1405. https://doi.org/10.30865/mib.v5i4.3169

Diba, A. F., & Warmana, G. O. (2022). *Sosialisasi BPJS kesehatan pada pegawai security untuk mengetahui pentingnya jaminan kesehatan di Ciputra Surabaya*. Jurnal Pengabdian Kepada Masyarakat, 2(3), 198–202.

Herdian, C., Kamila, A., & Budidarma, I. G. A. M. (2024). *Studi kasus feature engineering untuk data teks: Perbandingan label encoding dan one-hot encoding pada metode linear regresi*. Technologia: Jurnal Ilmiah, 15(1), 93.

Indrawan, G., Lemes, I. N., & Surata, I. N. (2024). *Peranan dinas kesehatan dalam pencegahan dan penanganan kecurangan (fraud) dalam program JKN berdasarkan Permenkes No. 16/2019 di Kabupaten Buleleng*. Kertha Widya, 11(2), 59–84.

Joel, L. O., Doorsamy, W., & Paul, B. S. (2022). *A review of missing data handling techniques for machine learning*. International Journal of Innovative Technology and Interdisciplinary Sciences, 5(3), 971–1005. https://doi.org/10.15157/ijitis.2022.5.3.971-1005

Kurnia, D., Mazdadi, M. I., Kartini, D., Nugroho, R. A., Abadi, F., & Korespondensi, P. (2023). *Seleksi fitur dengan particle swarm optimization pada klasifikasi penyakit Parkinson menggunakan XGBoost*. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 10(5), 1083–1094.

Kurniawan, D. (2022). *Pengenalan machine learning dengan Python*. Elex Media Komputindo.

Kusnaidi, M. R., Gulo, T., & Aripin, S. (2022). *Penerapan normalisasi data dalam mengelompokkan mahasiswa menggunakan K-Means untuk prioritas bantuan UKT*. Journal of Computer System and Informatics (JoSYC), 3(4), 330–338. https://doi.org/10.47065/josyc.v3i4.2112

Maretva Cendani, L., & Wibowo, A. (2022). *Perbandingan metode ensemble learning pada klasifikasi penyakit diabetes*, 13(1).

Mauren Michaela, S., Nurmalasari, M., & Hosizah, H. (2021). *Fraud in healthcare facilities: A narrative review*. Public Health of Indonesia, 7(4), 166–171. https://doi.org/10.36685/phi.v7i4.465

Mienye, I. D., & Sun, Y. (2022). *A survey of ensemble learning: Concepts, algorithms, applications, and prospects*. IEEE Access, 10, 99129–99149. https://doi.org/10.1109/access.2022.3207287

Mitigasi Defisit JKN. (2024). *Kompas*. https://www.kompas.id

Nabrawi, E., & Alanazi, A. (2023). *Fraud detection in healthcare insurance claims using machine learning*. Risks, 11(9). https://doi.org/10.3390/risks11090160

Parthasarathy, S., Lakshminarayanan, A. R., Khan, A. A. A., Sathick, K. J., & Jayaraman, V. (2023). *Detection of health insurance fraud using Bayesian-optimized XGBoost*. International Journal of Safety and Security Engineering, 13(5), 853–861. https://doi.org/10.18280/ijsse.130509

Peraturan Menteri Kesehatan Republik Indonesia No. 16 Tahun 2019 tentang Pencegahan dan Penanganan Kecurangan (Fraud).

Permana, I., & Salisah, F. N. (2022). *The effect of data normalization on the performance of the backpropagation algorithm*. Indonesian Journal of Informatics Research and Software Engineering, 2(1), 67–72.

Prasetya, M. R. A., Priyatno, A. M., & Nurhaeni. (2023). *Penanganan imputasi missing values pada data time series menggunakan metode data mining*. Jurnal Informasi dan Teknologi, 5(2), 52–62.

Retnoningsih, E., & Pramudita, R. (2020). *Mengenal machine learning dengan teknik supervised dan unsupervised learning menggunakan Python*. Bina Insani ICT Journal, 7(2), 156–165.

Rofiq, H. N. (2023). *Deteksi inefisiensi pada klaim BPJS kesehatan menggunakan machine learning*. Jurnal Jaminan Kesehatan Nasional, 3(1), 83–98. https://doi.org/10.53756/jjkn.v3i1.134

Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). *Prediction of insurance fraud detection using machine learning algorithms*. Mehran University Research Journal of Engineering and Technology, 41(1), 33–40.

Ruswanti, D. (2020). *Pengukuran performa SVM dan neural network dalam meramalkan tingkat curah hujan*. Jurnal Gaung Informatika, 3(1).

Santoso, P., Abijono, H., & Anggreini, N. L. (2021). *Algoritma supervised learning dan unsupervised learning dalam pengolahan data*. UNIRA Malang, 4(2).

Septiana Rizky, P., et al. (2022). *Perbandingan metode LightGBM dan XGBoost dalam menangani data kelas tidak seimbang*. J Statistika, 15(2).

Shahani, N. M., Zheng, X., Liu, C., Hasan, F. U., & Li, P. (2021). *Developing an XGBoost regression model for predicting Young's modulus of intact sedimentary rocks*. Frontiers in Earth Science, 9.

Simanjuntak, H. T. A., Sigiro, M., & Simanungkalit, H. M. (2021). *Deteksi fraud layanan rumah sakit menggunakan neural network*. Journal of Applied Technology and Informatics Indonesia, 1(1).

Singh, D., & Singh, B. (2020). *Investigating the impact of data normalization on classification performance*. Applied Soft Computing.

Singh, K. (2024). *Healthcare fraudulence: Leveraging AI techniques for detection*. International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets49394

Siringoringo, R., Angin, R. P., & Rumahorbo, B. (2022). *Model klasifikasi genetic-XGBoost dengan t-SNE pada peramalan pasar*. Jurnal Times, 11(1), 30–36.

Sitorus, C. M., Rizal, A., & Jajuli, M. (2020). *Prediksi risiko perjalanan transportasi online menggunakan SVM*. Jurnal Teknik Informatika dan Sistem Informasi, 6(2). https://doi.org/10.28932/jutisi.v6i2.2672

Soebroto, A. A. (2019). *Buku ajar AI, machine learning & deep learning*.

Syukron, M., Santoso, R., & Widiharih, T. (2020). *Perbandingan metode SMOTE-Random Forest dan SMOTE-XGBoost untuk klasifikasi hepatitis C*. Jurnal Gaussian, 9(3), 227–236.

Tanha, J., et al. (2020). *Boosting methods for multi-class imbalanced data classification: An experimental review*. Journal of Big Data, 7(1).

Vermaysha, A., & Nurmalitasari, N. (2023). *Prediksi harga rumah menggunakan regresi linear*. Prosiding Seminar Nasional Teknologi Informasi dan Bisnis, 6–11.

Wang, W., Chakraborty, G., & Chakraborty, B. (2021). *Predicting the risk of CKD using machine learning*. Applied Sciences, 11(1). https://doi.org/10.3390/app11010202

Yulianti, S. E. H., Soesanto, O., & Sukmawaty, Y. (2022). *Penerapan metode XGBoost pada klasifikasi nasabah kartu kredit*. Journal of Mathematics: Theory and Applications, 21–26.

Zhao, Y. F., Xie, J., & Sun, L. (2024). *On data quality and imbalance in ML-based design and manufacturing — A systematic review*. Engineering. https://doi.org/10.1016/j.eng.2024.04.024