Penerapan Pendekatan Sains Data untuk Pemodelan Klasifikasi Desain Keamanan Rangka Sepeda Melalui Pembelajaran Mesin Otomatis

Adriyan dan Budi Istana
Department of Mechanical Engineering, Universitas Muhammadiyah Riau, Indonesia
E-mail: adriyan@umri.ac.id

Abstract

Recent developments in engineering design have incorporated the domain of artificial intelligence due to the availability of historical data on the design of an object, such as a bicycle frame. Based on this historical data, a model can be built to derive a data-driven design decision. Therefore, this paper discusses this combination through a data science methodology approach in creating a design decision model using automatic machine learning (AutoML). The model is a multi-class classification model in determining the safety of a bicycle frame whether it passes two types of load tests or not. Exploratory data analysis techniques were applied to assess the features used in the modeling. This dataset has a disproportionate number of classes expressed by an unbalanced ratio of 10.77. The classification model was trained with a subset of the training data through stratifed K-fold cross-validation using the XGBoost algorithm for default values of its parameters. Two models were built based on original and prepared training data. Both models provide performance that is not too significantly different through the multi-class Matthews correlation coefficient value. However, the confusion matrix shows that the model with data preparation can better classify the minority classes. Finally, this study provides an initial reference for modeling safety-related design decisions on bicycle frames.

Keywords: *safety design of bike frame, data science approach, multi-class classification, and AutoML.*

Abstrak

Perkembangan terkini dalam desain kerekayasaan dilakukan dengan memadukan ranah kecerdasan buatan karena tersedianya data historis rancangan suatu objek, misalnya rangka sepeda. Berdasarkan data historis ini, sebuah model dapat dibangun untuk memperoleh keputusan desain berbasis data-driven. Untuk itu, kajian ini membahas perpaduan ini melalui pendekatan metodologi data sains dalam membuat sebuah model keputusan desain menggunakan pembelajaran mesin otomatis (automatic machine learning yang disingkat AutoML). Model yang dibangun merupakan model klasfikasi multi kelas dalam menentukan keamanan rangka sepeda dengan lulus atau tidaknya suatu rangka sepeda dalam dua jenis uji beban. Teknik exploratory data analysis diterapkan untuk menilai fitur-fitur yang digunakan dalam pemodelan. Dataset ini memiliki jumlah kelas yang propersinya tidak berimbang yang dinyatakan dengan rasio tak-seimbang sebesar 10.77. Model klasifikasi dilatih dengan bagian data latih melalui validasi silang stratified K-fold menggunakan algoritma XGBoost dengan parameter untuk nilai default-nya. Ada dua model yang dibangun berdasarkan data latih tanpa persiapan dan dengan persiapan data. Kedua model memberikan perfomansi yang tidak terlalu signifikan berbeda melalui nilai multi-class Matthews correlation coefficient. Namun, matriks konfusi menunjukkan bahwa model dengan persiapan data dapat lebih baik mengklasifikasikan kelas-kelas minoritas. Akhirnya, kajian yang dilakukan ini menjadi sebuah rujukan awal untuk pemodelan keputusan desain terkait keamanan pada rangka sepeda.

Kata kunci: desain keamanan rangka sepeda, pendekatan sains data, klasifikasi multi-kelas, dan AutoML.

1. Pendahuluan

Proses perancangan kerekayasaan umumnya menerapkan ranah-ranah keilmuan rekayasa secara spesifik untuk menghasilkan suatu rancangan berupa produk atau sistem. Pada desain sistem mekanik terdapat parameter atau variabel masukan yang berpengaruh secara signifikan, seperti geometri dan dimensi, kategori komponen, kondisi batas (tumpuan), besar dan jenis pembebanan, serta jenis material yang digunakan. Keseluruhan parameter atau variabel masukan ini dilibatkan untuk menentukan kelayakan variabel keluaran pada objek yang dirancang, misalnya, tegangan minimum yang diizinkan atau defleksi maksimum yang diizinkan. Variabel luaran dapat juga disebut sebagai keputusan desain.

Parameter atau variabel masukan dalam perancangan umumnya memiliki jumlah yang besar dari sisi kuantitas yang perlu dievaluasi untuk menentukan variabel keluaran objek rancangan optimum. Pengevaluasian dapat dilakukan melalui penggunaan simulasi numerik melalui penggunaan metode elemen hingga [1]. Namun, evaluasi setiap kombinasi pada masing-masing parameter atau variabel masukan untuk memperoleh variabel

optimum yang memakan waktu cukup lama dan kurang praktis. Untuk itu, insinyur perancangan terkadang melewati saja tahapan ini dalam mengambil suatu keputusan desain [2].

Untuk beberapa objek desain telah ada sekumpulan data rancangan yang terdiri atas parameter atau variabel masukan dan variabel luarannya. Kumpulan data rancangan ini dapat menjadi cara terbaru dalam merancang sebuah objek desain. Tentunya, kondisi ini didukung dengan perkembangan yang masif dalam ranah pembelajaran mesin (*machine learning*) dan *deep learning*. Pengintegrasian pembelajaran mesin dan/atau *deep learning* ke dalam perancangan dikenal dengan suatu terminologi rancangan berbasis pengetahuan (*knowledge-based design*) [3] atau *data-driven engineering design* [4, 5].

Perancangan sepeda merupakan salah satu objek desain yang menarik dalam penerapan pembelajaran mesin dan *deep learning*. Usaha oleh Regenwetter, dkk [6] dalam menyediakan dataset rancangan sepeda dengan nama BIKED. Berdasarkan dataset ini memungkinkan kajian untuk mengeksplorasi pengembangan algoritma untuk pemodelan sepeda. Peneliti yang sama juga memperkenalkan FRAMED sebagai dataset turunan dari BIKED [7]. Dataset ini merupakan data parametrik untuk sepeda yang terdiri atas dataset untuk validitas geometri dan keamanan struktur rangka sepeda.

Dataset FRAMED telah digunakan untuk membuat model *surrogate* klasifikasi dan regresi melalui penerapan pembelajaran mesin otomatis atau *automatic machine learning* (AutoML) dengan *framework* AutoGluon. Penggunaan AutoML dengan *weighted ensembled classifier* menghasilkan model *surrogate* pengklasifikasian validitas geometri yang terbaik dibanding penerapan algoritma lainnya. Hal ini ditandai dengan nilai performansi yang tinggi (di atas 90%) baik skor F1 (dengan perataan makro) dan skor akurasi.

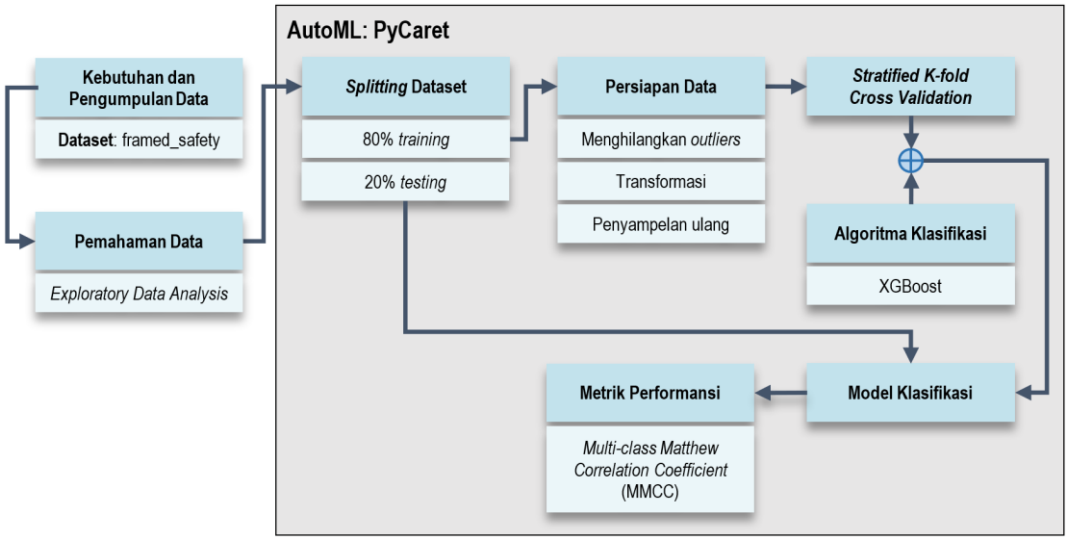
Selanjutya, model *surrogate* regresi dengan performansi terbaik untuk sepuluh jenis target diperoleh melalui penggunaan AutoML dengan *weighted ensembled regressor*. Sepuluh target dalam pemodelan ini diantaranya yaitu defleksi di beberapa posisi bagian sepeda, faktor keamanan, dan massa sepeda. Performansinya dinyatakan ke dalam nilai koefisien determinasi (R^2), nilai galat kuadrat rata-rata (MSE), dan galat absolut rata-rata (MAE). Perfomansi kedua model *surrogate* klasifikasi dan regresi diperoleh setelah model dioptimasi dengan optimisasi Bayesian.

Kajian yang telah dilakukan ini menunjukkan penggunaan AutoML dapat diandalkan untuk perancangan berbasis *data-driven*. Berdasarkan hal demikian, kajian yang diajukan dalam manuskrip ini juga menerapkan metodologi sains data dengan AutoML pada dataset FRAMED. Dataset yang dimaksud berupa klasfikasi multikelas hasil dua jenis uji beban rangka sepeda yang dikenal dengan nama *framed safety* dengan deskripsi detailnya pada [7]. Berbeda dengan kajian oleh Regenwetter, dkk [7] serta Piccard dan Ahmed [2] yang menggunakan skor F1 perataan makro untuk metrik performansinya, kajian ini menerapkan *multi-class* Matthews *correlation coefficient* (MMCC) [8] sebagai metrik dalam mengukur performansi model yang dibangun. Melalui kajian ini diharapkan dapat menjadikan pijakan awal pada perancangan keamanan rangka sepeda berbasis *data-driven*.

2. Metodologi

Kajian yang disajikan dalam manuskrip ini menerapkan pendekatan sains data untuk pemodelan klasifikasi desain kemanan rangka sepeda. Pendekatan sains data yang dimaksud yaitu penerapan metodologi fundamental untuk sains data oleh IBM [9]. Metodologi fundamental oleh IBM menyediakan sepuluh tahapan untuk memandu dalam menyelesaikan permasalahan yang menggunakan sains data.

Untuk permasalahan yang dibahas di dalam manuskrip ini hanya mengimplementasikan delapan tahapan awal saja. Dua tahapan awal yaitu *bussiness understanding* dan *analytic approach* telah dijelaskan melalui tujuan penelitian ini. Selanjutnya, artikel ini menekankan pada tahapan yang tersisa, yaitu kebutuhan data, pengumpulan data, pemahaman data, persiapan data, pemodelan, dan evaluasi. Secara umum, alur dari kajian ini ditampilkan dalam suatu diagram alir pada **Gambar 1**.



Gambar 1. Diagram alir pengklasifikasian desain rangka sepeda.

2.1. Dataset

Dataset rangka sepeda ini bernama FRAMED yang tersedia melalui DECODE (*Design Computation and Digital Engineering Lab*), *Mechanical Engineering Department at Massachusetts Institute of Technology* [7]. *Dataset* ini dirilis untuk publik yang merupakan adaptasi dari BIKED [6]. Untuk BIKED sendiri merupakan koleksi dari data awal oleh Rinard [10].

Dataset ini terdiri atas dua jenis yaitu *dataset* yang menunjukkan apakah (1) geometri desain rangka sepeda valid, dan (2) rangka sepeda aman setelah pengujian dengan dua jenis uji beban. *Dataset* (1) dan (2) ini masing-masingnya dinamai dengan `framed_validity` dan `framed_safety`. Kedua *dataset* ini ditujukan untuk proses klasifikasi biner dan multikelas. Untuk kajian ini digunakan *dataset* (2), yaitu `framed_safety`, yang kemudian disebut dengan *dataset* saja hingga akhir manuskrip ini.

2.2. Exploratory Data Analysis dan Persiapan Data

Exploratory data analysis atau EDA merupakan suatu teknik yang digunakan untuk menggali informasi atau wawasan suatu dataset. Dalam hal ini, EDA menjadi sebuah tahapan sentral dalam pemahaman data. Melalui tahapan ini, dataset ditelaah melalui perhitungan statistik deskriptif beserta dengan visualisasi data.

Setelah EDA dilaksanakan dapat ditentukan persiapan data yang sesuai untuk pemodelan. Persiapan yang dimaksud dapat berupa imputasi, membuang data pencilan (*outliers*), transformasi data, melakukan penyampelan ulang jika kelas dataset tidak seimbang, dan pemilihan fitur-fitur yang berkontribusi besar terhadap model.

2.3. Pemodelan dan Evaluasi

Algoritma XGBoost untuk proses klasifikasi multikelas digunakan untuk menghasilkan model dari dataset keamanan rangka sepeda. Algoritma ini dilatih berdasarkan fitur-fitur dan label yang dimiliki oleh dataset. Dataset yang dimaksud di sini adalah bagian data latih (*training dataset*) yang telah dilakukan persiapan data. Sementara, itu untuk data uji (*testing dataset*) sesuai dengan kondisi asalnya tanpa dilakukan perlakuan dalam tahapan persiapan data. Hal ini ditujukan untuk mencegah terjadinya kebocoran informasi (*leakage*) dari data latih ke data uji.

Proporsi pemecahan data menjadi data latih dan uji masing-masingnya adalah 80% dan 20%. Selama proses latihan (*training*), model akan dilatih melalui penerapan validasi silang (*cross validation*) pada data latih. Validasi silang yang dipilih yaitu *stratified K-fold cross validation* untuk 5-*fold*. Penggunaan jenis validasi silang ini ditujukan untuk mengkomodir label yang proporsi kelasnya tidak seimbang. Selanjutnya, proses evaluasi model dilakukan untuk menguji performansi model yang telah dilatih.

Dalam permasalahan klasifikasi seringkali digunakan matrik konfusi (*confusion matrix*) untuk memetakan kelas-kelas yang diprediksi (*predicted label*) terhadap kelas-kelas yang sebenarnya (*true label*). Matrik konfusi untuk klasifikasi label dengan empat kelas ditunjukkan dalam **Gambar 2**. Pada **Gambar 2**, TP_i merupakan nilai kelas C_i yang diprediksi dan benar merupakan kelas C_i tersebut (i = 1, 2, 3, 4). Sementara itu, $K_{i,j}$ ($i \neq j$) menyatakan jumlah sampel yang diklasfikasikan sebagai kelas C_i tetapi sebenarnya merupakan kelas C_j.

		Nilai sebenarnya (*true value*)			
		C_1	C_2	C_3	C_4
Nilai prediksi (*predicted value*)	C_1	TP_1	$K_{2,1}$	$K_{3,1}$	$K_{4,1}$
	C_2	$K_{1,2}$	TP_2	$K_{3,2}$	$K_{4,2}$
	C_3	$K_{1,3}$	$K_{2,3}$	TP_3	$K_{4,3}$
	C_4	$K_{1,4}$	$K_{2,4}$	$K_{3,4}$	TP_4

Gambar 2. Matriks konfusi dengan C_1, C_2, C_3, dan C_4 merupakan kelas dari label.

Mengacu pada matriks konfusi ini dapat dikalkulasi performansi model. Performansi model klasifikasi dapat ditentukan dengan berbagai metrik. Matthews *correlation coefficient* (MCC) merupakan salah satu metrik klasifikasi yang dapat digunakan untuk klasifikasi biner. MCC dapat diekstensi ke permasalahan klasifikasi multi kelas tak-seimbang melalui perataan nilai MCC untuk setiap kelas sebagaimana yang didefinisikan oleh [8] dengan nama *multi-class* MCC

$$\mathrm{MMCC} = \frac{2}{C(C-1)} \sum_{i<j} \mathrm{MCC}(i,j). \qquad (1)$$

Berdasarkan persamaan (1) ini C merupakan jumlah kelas dan MCC(i, j) menyatakan nilai MCC antara kelas i dan j. Sebagai metrik yang merupakan koefisien korelasi, MCC memiliki nilai dalam rentang -1 (berkorelasi negatif secara kuat) s/d +1 (berkorelasi positif secara kuat), dengan formulasi MCC sendiri dapat merujuk pada persamaan (33) pada [8].

2.4. AutoML

Automatic machine learning (AutoML) atau pembelajaran mesin otomatis ditujukan untuk mengotomasi proses persiapan data, pemodelan, evaluasi, dan optimasi (*hyperparameter tuning*) model. Untuk itu, AutoML dapat mempercepat proses pembelajaran mesin tanpa harus menuliskan kode-kode yang cukup panjang dari proses-proses yang disebutkan. Melalui bahasa pemrograman python terdapat beberapa pustaka yang mendukung untuk AutoML ini seperti auto-sklearn [11], pycaret [12], H2O AutoML [13], dan autogluon [14].

2.5. XGBoost

XGBoost (e**X**treme **G**radient **Boost**ing) merupakan algoritma pembelajaran mesin kelompok *ensemble learning* yang diperkenalkan di tahun 2016 [15]. Algoritma ini didesain dengan menerapkan *gradient boosting ensemble learning* yang dapat bekerja secara terdistribusi/paralel, cepat, efisien, dan performansi tinggi. Oleh karena itu, algoritma ini dapat digunakan dengan parameter-parameter *default*-nya pada tahap awal tanpa memerlukan proses *hyperparameter tuning* lanjut. Namun, proses *hyperparameter tuning* juga dapat dilakukan dengan mengacu pada panduan terkait parameter-parameter XGBoost melalui laman dokumentasinya [16]. Umumnya, algoritma XGBoost tersedia di dalam setiap *framework* AutoML.

2.6. Perangkat Lunak dan Keras

Kajian penerapan penyampelan sintetis pada pemodelan data keamanan rangka sepeda dengan *ensemble learning* menggunakan python 3.11.9 dan pustaka-pustaka python untuk pembelajaran mesin otomatis yaitu pycaret 3. Sebelum diumpankan ke dalam pembelajaran mesin otomatis, data dieksplorasi terlebih dahulu melalui teknik EDA menggunakan pustaka numpy 1.26.4, pandas 2.1.4, matplotlib 3.7.5 dan seaborn 0.13.2. Kode program yang ditulis berdasarkan pustaka-pustaka ini dijalankan pada *notebook* dengan prosesor Intel i7 11800H yang berjalan pada *clock* maksimum 4.3 GHz dengan memori 32 GB.

3. Hasil dan Pembahasan

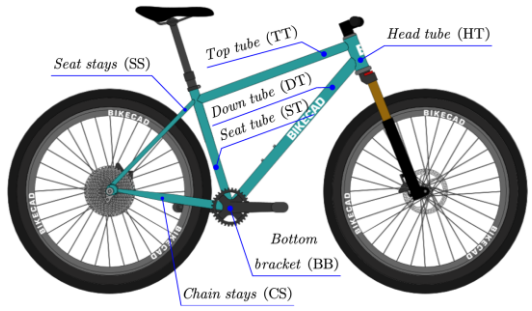
3.1. Dataset dan EDA

Dataset `framed_safety` terdiri atas 39 fitur dan 1 label serta 4046 baris data (observasi). Dataset ini tidak memiliki nilai yang hilang (*missing value*) baik di setiap fitur ataupun labelnya. Fitur sejumlah 39 merupakan penjabaran dari nama bagian rangka berdasarkan sifat geometri (*geometrical properties*) rangka sepeda serta material rangka seperti yang ditampilkan dalam **Tabel 1**. Nama-nama bagian rangka sepeda ini ditunjukkan melalui **Gambar 3**. Untuk pemodelan digunakan keseluruhan fitur pada dataset ini.

Sementara itu, untuk label yang terdiri atas empat kelas yang merupakan hasil uji dua jenis uji beban (*load test*). Keempat kelas hasil uji beban ini dinyatakan dengan gagal kedua uji beban (*both test failed*), lulus uji beban pertama (*first test passed*), lulus uji beban kedua (*second test passed*), dan lulus kedua uji beban (*both test passed*). Banyaknya observasi untuk masing-masing kelas ditampilkan dalam **Tabel 2**. Kolom Kode ditujukan untuk mengkode masing-masing kelas ke dalam nilai numerik.

Tabel 1.
Fitur dataset `framed_safety` yang dikelompokkan berdasarkan geometri rangka pada **Gambar 3**.

Fitur	Jumlah fitur	Tipe data
Bottom bracket (BB)	4	Numerik
Chain stays (CS)	6	Numerik
CSB include	1	Kategori biner
Down tube (DT)	3	Numerik
Head tube (HT)	6	Numerik
Seat stays (SS)	6	Numerik
SSB include	1	Kategori biner
Top tube (TT)	2	Numerik
Seat tube (ST)	5	Numerik
Dropout Offset	1	Numerik
Stack	1	Numerik
Material	3	Kategori biner



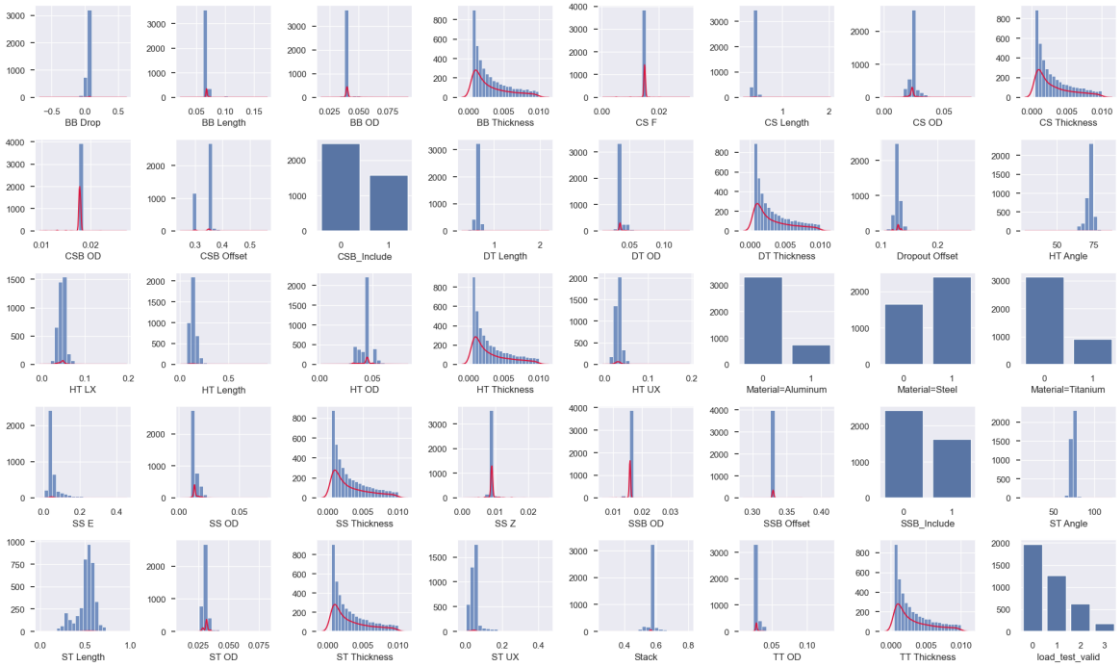
Gambar 3. Nama-nama bagain rangka sepeda yang diadopsi dari [7]. Gambar ini merupakan tangkapan layar dari perangkat lunak BikeCAD *online* di https://www.bikecad.ca/demo.php.

Melalui **Tabel 2** dapat diketahui bahwa kelas dari label dataset `framed_safety` tidak seimbang. Kelas gagal kedua uji beban merupakan kelas mayoritas sedangkan kelas lulus kedua uji beban adalah kelas minoritas. Selanjutnya, rasio tak-seimbang (*imbalanced ratio* yang disingkat IR) merupakan rasio kelas mayoritas terhadap kelas minoritas pada dataset ini sebesar 10.77. Untuk itu, teknik SMOTE (*synthetic minority oversampling technique*) [17] diterapkan untuk penyampelan ulang secara sintetis pada bagian data latih. Penyampelan ulang ini dilakukan secara otomatis nantinya dengan pustaka AutoML: PyCaret.

Tabel 2.
Keempat kelas hasil uji beban.

Kelas	Jumlah	Proporsi	Kode
gagal kedua uji beban	1971	48.7%	0
lulus uji beban pertama	1269	31.4%	1
lulus uji beban kedua	623	15.4%	2
lulus kedua uji beban	183	4.5%	3

Selanjutnya, penerapan teknik EDA dapat menunjukkan distribusi dari 39 fitur dan label. Untuk fitur dengan tipe data numerik akan divisualisasikan melalui histogram dan *kernel density estimation* (KDE). Sementara itu, fitur-fitur dengan kategori biner dapat divisualisasikan ke dalam diagram batang. Visualisasi fitur-fitur dan label ini ditampilkan dalam **Gambar 4**. Di samping itu, perhitungan statistik deskriptif fitur-futir dengan tipe data numerik diberikan melalui **Tabel 3**.



Gambar 4. Visualisasi distribusi fitur dan label pada data `framed_safety`.

Tabel 3.
Statistik deskriptif dataset `framed_safety` untuk fitur-fitur dengan tipe data numerik.

Fitur	Rerata	Standar deviasi	Min	25% (Q_1)	50% (Q_2)	75% (Q_3)	Maks	*Skew-ness*	Kurtosis	Persentase pencilan
BB *Drop*	0.0524	0.0320	-0.6650	0.0430	0.0650	0.0700	0.6060	-2.4993	85.8699	9.9852
BB *Length*	0.0691	0.0052	0.0254	0.0680	0.0680	0.0680	0.1700	5.3738	63.1607	12.4567
BB OD	0.0406	0.0041	0.0200	0.0400	0.0400	0.0400	0.0900	6.3915	57.8963	13.7914
BB *Thickness*	0.0032	0.0026	0.0005	0.0011	0.0022	0.0047	0.0100	0.9876	-0.1347	0.0000
CS F	0.0149	0.0014	0.0010	0.0150	0.0150	0.0150	0.0300	-2.0419	45.0814	5.2645
CS *Length*	0.4145	0.0577	0.1999	0.4000	0.4067	0.4250	2.0000	10.0449	199.6943	9.9110
CS OD	0.0239	0.0042	0.0024	0.0233	0.0233	0.0233	0.0700	2.8069	16.6990	42.7583
CS *Thickness*	0.0032	0.0026	0.0005	0.0011	0.0022	0.0047	0.0100	0.9850	-0.1225	0.0000
CSB OD	0.0178	0.0008	0.0100	0.0178	0.0178	0.0180	0.0270	-4.1939	49.9780	2.8917
CSB *Offset*	0.3362	0.0249	0.2600	0.3000	0.3500	0.3500	0.5500	-0.4843	0.6113	0.0494
DT *Length*	0.6493	0.0543	0.3388	0.6298	0.6516	0.6659	2.1598	6.2273	163.1747	9.3673
DT OD	0.0378	0.0054	0.0145	0.0365	0.0365	0.0365	0.1333	5.2655	55.2028	25.0618

DT *Thickness*	0.0032	0.0026	0.0005	0.0011	0.0023	0.0048	0.0100	0.9593	-0.2069	0.0000
Dropout Offset	0.1305	0.0070	0.1090	0.1300	0.1300	0.1350	0.2560	4.2606	61.6109	14.0138
HT *Angle*	71.9666	3.2073	30.0000	71.0000	72.5000	73.0000	85.5000	-3.4277	29.2505	10.1335
HT LX	0.0462	0.0101	0.0005	0.0406	0.0470	0.0500	0.1900	1.6851	16.7529	4.7949
HT *Length*	0.1372	0.0398	0.0271	0.1119	0.1324	0.1520	0.8806	4.4953	65.7004	3.3366
HT OD	0.0432	0.0062	0.0050	0.0400	0.0450	0.0450	0.0800	-0.1590	1.9237	14.1621
HT *Thickness*	0.0031	0.0026	0.0005	0.0011	0.0022	0.0047	0.0100	1.0018	-0.0943	0.0000
HT UX	0.0319	0.0096	0.0012	0.0271	0.0310	0.0351	0.1900	4.8712	57.7101	5.2892
SS E	0.0620	0.0467	0.0050	0.0450	0.0450	0.0600	0.4300	3.0664	11.0545	17.5729
SS OD	0.0142	0.0031	0.0030	0.0130	0.0130	0.0150	0.0715	4.7778	53.2186	10.2323
SS *Thickness*	0.0032	0.0025	0.0005	0.0011	0.0023	0.0047	0.0100	0.9787	-0.1287	0.0000
SS Z	0.0091	0.0014	0.0008	0.0090	0.0090	0.0090	0.0254	3.6636	35.0977	12.7533
SSB OD	0.0158	0.0011	0.0070	0.0158	0.0158	0.0160	0.0360	7.4605	169.9352	4.7207
SSB *Offset*	0.3306	0.0058	0.2900	0.3300	0.3300	0.3300	0.4100	7.3709	78.3605	1.9031
ST *Angle*	73.3818	3.1130	16.2761	73.0000	73.5000	74.0000	120.0000	-3.0376	69.0902	18.3391
ST *Length*	0.5083	0.1050	0.0550	0.4760	0.5289	0.5741	0.9000	-1.0408	1.1732	10.9738
ST OD	0.0317	0.0035	0.0120	0.0302	0.0318	0.0318	0.0800	5.3785	55.8514	13.1488
ST *Thickness*	0.0032	0.0026	0.0005	0.0011	0.0022	0.0047	0.0100	0.9815	-0.1362	0.0000
ST UX	0.0506	0.0406	0.0020	0.0325	0.0467	0.0509	0.4350	4.3717	26.0082	8.5517
Stack	0.5653	0.0266	0.2800	0.5656	0.5656	0.5656	0.7940	-0.9968	19.7522	25.1359
TT OD	0.0298	0.0045	0.0127	0.0286	0.0286	0.0286	0.1300	6.2218	84.0376	36.7523
TT *Thickness*	0.0032	0.0026	0.0005	0.0011	0.0022	0.0047	0.0100	0.9908	-0.1133	0.0000

Melalui visualisasi pada **Gambar 4** dan **Tabel 3** terlihat bahwa fitur-fitur yang bertipe data numerik memiliki distribusi yang tak-normal ataupun hampir normal. Kondisi ini ditandai oleh plot distribusi, nilai *skewness* (kemencengan), dan nilai kurtosis masing-masing fitur. Berdasarkan nilai *skewness* dan plot distribusi terdapat fitur dengan ujung distribusi yang memanjang ke sisi kiri (nilai *skewness* negatif) dan ke sisi kanan (nilai *skewness* positif).

Mayoritas fitur-fitur numerik memiliki karakteristik leptokurtik (nilai kurtosis positif). Sementara itu, terdapat tujuh fitur dengan karakteristik platikurtik (nilai kurtosis negatif). Fitur-fitur dengan karakteristik platikurtik ini memiliki nilai yang hampir mendekati nol, sehingga dapat dinyatakan fitur-fitur tersebut mendekati distribusi normal. Untuk itu, fitur-fitur numerik dapat ditransformasi dengan metode Yeo-Johnson untuk merubah data yang tidak terdistribusi secara normal menjadi mendekati distribusi normal.

Selanjutnya, banyaknya pencilan pada setiap fitur dapat ditentukan melalui perhitungan setiap butir data *x* yang berada diluar kriteria

$$Q_1 - 1.5 \cdot IQR \leq x \leq Q_3 + 1.5 \cdot IQR \qquad (2)$$

dengan IQR (*inter quartile range*) merupakan jarak antar kuartil Q_3 dan Q_1. Banyak pencilan ini dinyatakan ke dalam persentase pencilan pada kolom terakhir **Tabel 3**. Perhitungan dengan menggunakan persamaan (2) menunjukkan mayoritas fitur-fitur numerik dengan pencilan di atas 5%.

Butir-butir data pencilan ini juga dikeluarkan dari fitur-fitur terkait dalam proses pemodelan. Namun, penerapan persamaan (2) tentu akan menghilangkan banyaknya observasi dalam dataset. Untuk itu, metode deteksi anomali berbasis pembelajaran tidak terbimbing (*unsupervised learning*) digunakan dalam menghindari banyaknya obervasi yang hilang akibat penggunaan persamaan (2). Dalam kajian ini, metode *local outlier factor* (LOF) digunakan dalam mengeluarkan butir-butir data pencilan [18].

3.2. Proses Persiapan Data, Pemodelan, dan Evaluasi dengan AutoML: PyCaret

Tahapan persiapan data telah dideskripsikan pada bagian 3.1 berikut dengan identifikasi perlakuan pada fitur-fiturnya berdasarkan hasil EDA. Selanjutnya, algoritma XGBoost digunakan untuk pemodelan terkait sukses atau tidaknya dua uji beban pada rangka sepeda untuk keseluruhan fitur yang dimiliki. Model yang dilatih dengan algoritma XGBoost ini menggunakan parameter *default*-nya yang diberikan dalam **Tabel 4**.

Persiapan dan pemodelan dataset dinyatakan ke dalam dua buah model. Model pertama menggunakan data latih dengan fitur-fitur tanpa persiapan data (menghilangkan pencilan dan transformasi) untuk labelnya yang tak-seimbang, atau menggunakan dataset asal. Selanjutnya, model kedua dihasilkan untuk data latih yang fitur-fiturnya melalui persiapan data dan labelnya telah diseimbangkan. Untuk setiap tahapan pada kedua model ini dilaksanakan melalui penerapan *framework* AutoML: PyCaret.

Tabel 4.
Parameter *default* XGBoost untuk klasifikasi `frame_safety`.

Parameter	Nilai
learning_rate	0.3
min_split_loss	0
max_depth	6
max_leaves	0
min_child_weight	1
max_delta_step	1
subsample	1
colsample_bytree	1
colsample_bylevel	1
colsample_bynode	1
lambda	1
alpha	0
scale_pos_weight	1

Data latih pada model pertama terdiri atas 3236 observasi dan data ujinya dengan 810 observasi. Sedangkan untuk data latih pada model kedua memiliki 5912 observasi yang merupakan hasil dari proses persiapan data latih yang telah dinyatakan sebelumnya. Sementara data uji untuk model kedua sama dengan data uji pada model pertama.

Setelah kedua model dilatih dapat diperoleh performansi kedua model. Performansi kedua model ini merupakan nilai MMCC selama proses validasi silang. Model pertama dan kedua masing-masingnya memiliki nilai MMCC sebesar 0.6804 ± 0.0145 dan 0.6872 ± 0.0143. Setelah proses persiapan data dapat diketahui bahwa model kedua lebih baik performansinya dibandingkan model pertama. Namun, peningkatan performansi ini (nilai MMCC) tidak terlalu signifikan untuk data latih dengan telah melalui persiapan dan tanpa melalui persiapan (data latih asal).

Selanjutnya, kedua model yang telah dilatih ini digunakan untuk memprediksi performansi (nilai MMCC) pada data uji. Hasil prediksi untuk kedua model ini dinyatakan ke dalam matriks konfusi seperti yang ditampilkan pada **Gambar 5**. Dari kedua matriks konfusi terlihat bahwa kelas mayoritas C_1 (gagal kedua uji beban) memiliki penurunan klasifikasi nilai sebenarnya (11 buah observasi) dibandingkan model kedua. Hal sebaliknya terjadi pada tiga kelas lainnya yang memiliki peningkatan hasil klasifikasi kelas sebenarnya (9 buah observasi untuk ketiga kelas) pada model kedua.

		Nilai sebenarnya (*true value*)			
		C_1	C_2	C_3	C_4
Nilai prediksi (*predicted value*)	C_1	349	39	17	14
	C_2	34	206	14	0
	C_3	11	9	88	13
	C_4	0	0	6	10

(a)

		Nilai sebenarnya (*true value*)			
		C_1	C_2	C_3	C_4
Nilai prediksi (*predicted value*)	C_1	338	35	12	14
	C_2	35	208	9	1
	C_3	14	10	91	8
	C_4	7	1	13	14

(b)

Gambar 5. Matriks konfusi (a) model pertama dan (b) model kedua.

Meskipun terjadi peningkatan pada kelas-kelas non mayoritas namun secara umum tidak berdampak signifikan pada performansi model. Selanjutnya, performansi model pertama dan kedua masing-masingnya memiliki nilai MMCC sebesar 0.6915 dan 0.6919. Kedua model ini memiliki performansi korelasi di angka 0.7 yang menunjukkan korelasi kuat secara positif antara fitur-fitur dengan label [19, 20]. Secara umum, model kedua dapat dinyatakan sedikit lebih baik mengklasfikasikan kelas-kelas minoritas.

Kajian ini telah memberikan deskripsi yang cukup komprehensif dalam hal eksplorasi data dan persiapan pada data latih. Penggunaan metrik MMCC dalam kajian ini didasari pada pernyataan Tanha, dkk [8] dalam menilai performansi klasifikasi multi-kelas tak seimbang. Namun, penggunaan MCC juga tidak direkomendasikan oleh Zhu [21] sebagai metrik klasifikasi dataset tak-seimbang.

4. Simpulan

Kajian dalam artikel ini telah menunjukkan pendekatan sains data dengan AutoML dapat digunakan dalam pengambilan keputusan desain terkait keamanan rangka sepeda. Keputusan ini didasari pada keseluruhan fitur yang dimiliki pada datasetnya. Penggunaan persiapan pada bagian data latih tidak memberikan peningkatan performansi yang signifikan dalam model klasifikasi yang dibangun. Hal ini disebabkan karena proses persiapan data latih menurunkan jumlah kelas mayoritas diklasifikasikan dengan tepat dan menurunkan kelas-kelas minoritas. Kajian ini dapat menjadi sebuah rujukan awal untuk pemodelan keputusan desain terkait keamanan pada rangka sepeda. Untuk masa mendatang, kajian akan melibatkan penggunaan algoritma-algoritma klasifikasi lainnya seperti berbasis pada *voting ensemble* dan *stacking ensemble*. Proses persiapan data yang belum diterapkan dan penggunaan jenis metrik lainnya untuk menilai performansi pada dataset yang tidak seimbang.

Daftar Pustaka

[1] Xu W, Neumann I. Finite element analysis based on a parametric model by approximating point clouds. *Remote Sens* 2020; 12: 1–26.

[2] Picard C, Ahmed F. Fast and Accurate Zero-Training Classification for Tabular Engineering Data. *J Mech Des*; 146. Epub ahead of print 2024. DOI: https://doi.org/10.1115/1.4064811.

[3] Maher ML. Machine Learning in Engineering Design: Learning Generalized Design Prototypes from Examples. In: *Tasso, C., de Arantes e Oliveira, E.R. (eds) Development of Knowledge-Based Systems for Engineering. International Centre for Mechanical Sciences, vol 333*. Vienna: Springer, 1998.

[4] Jenis J, Ondriga J, Hrcek S, et al. Engineering Applications of Artificial Intelligence in Mechanical Design and Optimization. *Machines* 2023; 11: 577.

[5] Brown NK, Garland AP, Fadel GM, et al. Deep reinforcement learning for engineering design through topology optimization of elementally discretized design domains. *Mater Des* 2022; 218: 110672.

[6] Regenwetter L, Curry B, Ahmed F. BIKED: A Dataset for Computational Bicycle Design with Machine Learning Benchmarks. *J Mech Des*

2022; 144: 1–19.
[7] Regenwetter L, Weaver C, Ahmed F. FRAMED: An AutoML Approach for Structural Performance Prediction of Bicycle Frames. *Comput Des* 2022; 156: 1–33.
[8] Tanha J, Abdi Y, Samadi N, et al. Boosting methods for multi-class imbalanced data classification: an experimental review. *J Big Data*; 7. Epub ahead of print 2020. DOI: 10.1186/s40537-020-00349-y.
[9] IBM. *Foundational Methodology for Data Science*. Somers, New York, 2015.
[10] Rinard D. Frame deflection test, https://www.sheldonbrown.com/rinard/rinard%7B_%7Dframetest.html (1996, accessed 18 April 2024).
[11] Feurer M, Klein A, Jost KE, et al. Efficient and Robust Automated Machine. In: *Automated machine learning: methods, systems, challenges. Advances in Neural Information Processing Systems 28 (NIPS 2015)*. 2015, pp. 113–34.
[12] Ali M. PyCaret: An open source, low-code machine learning library in Python.
[13] Ledell E, Poirier S. H2O AutoML: Scalable automatic machine learning. *7th ICML Work Autom Mach Learn* 2020; 1–16.
[14] Erickson N, Mueller J, Shirkov A, et al. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. In: *7th ICML Workshop on Automated Machine Learning*. 2020.
[15] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.
[16] Developers Xgb. Notes on Parameter Tuning in XGBoost Tutorials. *XGBoost Documentation*, https://xgboost.readthedocs.io/en/stable/tutorials/param_tuning.html (2022, accessed 18 April 2024).
[17] Chawla N V, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002; 16: 321–357.
[18] Breunig MM, Kriegel H-P, Ng RT, et al. LOF: Identifying Density-Based Local Outliers. In: *Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX*. 2000, pp. 1–12.
[19] Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg*; 126. Epub ahead of print 2018. DOI: 10.1213/ANE.0000000000002864.
[20] Akoglu H. User's guide to correlation coefficients. *Turkish J Emerg Med* 2018; 18: 91–93.
[21] Zhu Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognit Lett* 2020; 136: 71–80.