

# Perbandingan CatBoost dan Elastic Net untuk Estimasi Komposisi Tubuh Berbasis Antropometri

Aditiya Ari Wicaksono\*, Khanun Roisatul Ummah

Teknik Informatika, Fakultas Komunikasi dan Informatika, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

Email: <sup>1</sup>, \*aditiyaariw@gmail.com, <sup>2</sup>kru344@ums.ac.id

Email Penulis Korespondensi: aditiyaariw@gmail.com\*

Submitted: 09/06/2026; Accepted: 20/06/2026; Published: 30/06/2026

**Abstrak**—Pemantauan komposisi tubuh diperlukan untuk memberi gambaran kondisi tubuh yang lebih informatif dibandingkan BMI, terutama terkait persentase lemak dan massa otot. Pemeriksaan standar seperti *Dual-Energy X-ray Absorptiometry* (DXA) memiliki keterbatasan biaya dan aksesibilitas, sehingga diperlukan pendekatan estimasi berbasis antropometri yang lebih praktis. Penelitian ini bertujuan membandingkan performa CatBoost dan Elastic Net dalam mengestimasi komposisi tubuh menggunakan *dataset* National Health and Nutrition Examination Survey (NHANES) 2017–2018. Data yang digunakan berjumlah 3.551 data setelah proses *cleaning*, terdiri atas delapan fitur *input* antropometri dan dua target berbasis DXA. Metode penelitian mengikuti CRISP-DM, meliputi pemahaman data, persiapan data, pemodelan, evaluasi, dan *deployment* sederhana. Data dibagi menjadi 2.840 data latih dan 711 data uji, kemudian model dievaluasi menggunakan MAE, RMSE, dan R<sup>2</sup>. Hasil pengujian menunjukkan CatBoost memperoleh performa terbaik dengan Mean MAE 1,588891, Mean RMSE 2,044905, dan Mean R<sup>2</sup> 0,862830, sedangkan Elastic Net memperoleh Mean MAE 1,888025, Mean RMSE 2,408882, dan Mean R<sup>2</sup> 0,819884. Selain itu, prototipe berbasis Streamlit dikembangkan untuk simulasi inferensi terhadap data antropometri baru. Hasil penelitian ini masih terbatas pada dataset NHANES 2017–2018 sehingga generalisasi ke populasi lain memerlukan validasi lanjutan.

**Kata Kunci:** antropometri; CatBoost; Elastic Net; komposisi tubuh; machine learning

**Abstract**—Body composition monitoring is needed to provide more informative health information than BMI, especially regarding body fat percentage and muscle mass. Standard examination methods such as *Dual-Energy X-ray Absorptiometry* (DXA) have limitations in cost and accessibility, so a more practical anthropometry-based estimation approach is required. This study compares the performance of CatBoost and Elastic Net in estimating body composition using the National Health and Nutrition Examination Survey (NHANES) 2017–2018 dataset. The data consisted of 3,551 records after cleaning, with eight anthropometric input features and two DXA-based targets. The research method followed CRISP-DM, including data understanding, data preparation, modeling, evaluation, and simple deployment. The data were divided into 2,840 training samples and 711 testing samples, and the models were evaluated using MAE, RMSE, and R<sup>2</sup>. The results show that CatBoost achieved the best performance with a Mean MAE of 1.588891, Mean RMSE of 2.044905, and Mean R<sup>2</sup> of 0.862830, while Elastic Net obtained a Mean MAE of 1.888025, Mean RMSE of 2.408882, and Mean R<sup>2</sup> of 0.819884. In addition, a Streamlit-based prototype was developed to simulate inference on new anthropometric data. Further validation on other populations is required.

**Keywords:** anthropometry; CatBoost; Elastic Net; body composition; machine learning

## 1. PENDAHULUAN

Pemantauan komposisi tubuh, yang mencakup persentase lemak dan massa otot, penting dilakukan secara berkala karena dapat memberikan gambaran kondisi tubuh yang lebih informatif dibandingkan penilaian berdasarkan berat badan saja. Perubahan komposisi tubuh dapat terjadi secara bertahap seiring bertambahnya usia, perubahan aktivitas fisik, pola makan, dan kondisi kesehatan. Penurunan massa otot berkaitan dengan risiko sarkopenia, sedangkan penumpukan jaringan lemak yang berlebih, terutama pada area tertentu, dapat berkaitan dengan peningkatan risiko penyakit kardiometabolik [1], [2]. Kondisi tersebut menunjukkan bahwa pemantauan kesehatan tidak cukup hanya menggunakan indikator sederhana yang menggambarkan ukuran tubuh secara umum, tetapi juga perlu memperhatikan proporsi antara massa lemak dan massa otot. Instrumen medis seperti *Dual-Energy X-ray Absorptiometry* (DXA) sering digunakan sebagai metode acuan dalam pengukuran komposisi tubuh karena mampu memberikan informasi yang lebih rinci mengenai massa lemak, massa bebas lemak, dan distribusi komposisi tubuh [3]. Namun, penggunaan DXA masih memiliki keterbatasan dari sisi biaya, ketersediaan alat, dan akses pemeriksaan, sehingga belum selalu memungkinkan untuk pemantauan rutin pada masyarakat umum [4]. Di sisi lain, Indeks Massa Tubuh (BMI) sering digunakan sebagai alternatif sederhana karena hanya membutuhkan data berat badan dan tinggi badan. Meskipun mudah digunakan, BMI belum mampu membedakan proporsi massa lemak dan massa otot secara akurat. Dua individu dengan nilai BMI yang sama dapat memiliki komposisi tubuh yang berbeda, sehingga penggunaan BMI saja dapat memberikan gambaran yang kurang lengkap terhadap kondisi tubuh. Oleh karena itu, pemanfaatan data antropometri yang diproses melalui pendekatan *machine learning* menjadi salah satu alternatif yang relevan untuk mengestimasi komposisi tubuh secara lebih praktis [5]. Pendekatan ini diharapkan dapat menghasilkan estimasi yang lebih informatif dibandingkan BMI, tetapi tetap lebih mudah diakses dibandingkan pemeriksaan berbasis alat khusus seperti DXA.

Pemodelan komputasi berbasis antropometri memiliki tantangan tersendiri karena variabel-variabel ukuran tubuh sering kali saling berkaitan. Berat badan, tinggi badan, BMI, lingkaran pinggang, lingkaran lengan, dan ukuran tubuh lain dapat memiliki hubungan yang kuat karena sama-sama merepresentasikan karakteristik fisik responden. Hubungan antarfitur yang terlalu kuat dapat menyebabkan multikolinearitas dan menurunkan stabilitas model regresi apabila tidak ditangani dengan tepat [6]. Dalam konteks tersebut, Elastic Net menjadi salah satu metode yang relevan karena menggabungkan karakteristik penalti *Lasso* dan *Ridge* dalam satu model regresi. Penalti *Lasso* berperan dalam menyusutkan koefisien fitur yang kurang relevan, sedangkan penalti *Ridge* membantu menjaga kestabilan koefisien ketika fitur saling berkorelasi [7]. Kombinasi tersebut membuat Elastic Net sesuai digunakan pada data antropometri yang memiliki hubungan antarvariabel cukup kuat, sekaligus tetap mempertahankan sifat model regresi yang lebih sederhana dan mudah ditafsirkan. Selain pendekatan linier, penelitian ini juga menggunakan CatBoost sebagai model pembanding. CatBoost merupakan algoritma *gradient boosting* berbasis pohon keputusan yang dapat menangani data tabular, memproses fitur kategorikal, dan mempelajari pola nonlinier antarfitur. Kemampuan tersebut penting karena hubungan antara data antropometri dan komposisi tubuh tidak selalu mengikuti pola linier sederhana. Misalnya, pengaruh BMI terhadap persentase lemak tubuh dapat berbeda berdasarkan jenis kelamin, usia, dan ukuran lingkaran pinggang. CatBoost memiliki kemampuan untuk menangkap interaksi semacam itu melalui proses pembentukan pohon keputusan secara bertahap. Selain itu, metode berbasis *boosting* juga banyak digunakan pada data kesehatan karena mampu mempelajari pola kompleks, meskipun tetap perlu dikendalikan agar tidak mengalami *overfitting* [8]. Kajian lain menunjukkan bahwa CatBoost mampu menghasilkan performa tinggi pada pemodelan tingkat obesitas berbasis data fisik dan kebiasaan hidup [9]. Dengan demikian, Elastic Net dan CatBoost mewakili dua pendekatan yang berbeda, yaitu regresi linier terregularisasi yang stabil terhadap multikolinearitas dan model nonlinier berbasis *ensemble* yang fleksibel dalam mempelajari interaksi antarfitur.

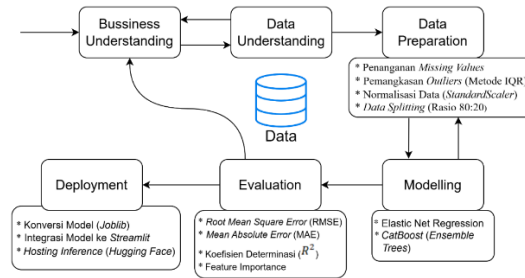
Penelitian sebelumnya telah banyak memanfaatkan *machine learning* dalam bidang kesehatan, terutama untuk memprediksi obesitas, risiko metabolik, kenaikan berat badan, atau komposisi tubuh berdasarkan data antropometri dan variabel kesehatan lain. Penelitian [10] menunjukkan bahwa data antropometri dan laboratorium dapat digunakan untuk memprediksi lemak tubuh, tetapi penggunaan variabel laboratorium membuat pendekatan tersebut tidak sepenuhnya merepresentasikan skenario estimasi berbasis antropometri sederhana. Penelitian [11] mengembangkan model prediksi tingkat obesitas dengan dukungan *explainable artificial intelligence*, tetapi keluaran yang dihasilkan masih berupa klasifikasi tingkat obesitas, bukan estimasi regresi terhadap indikator komposisi tubuh berbasis DXA. Sementara itu, penelitian [12] menggunakan pendekatan *machine learning* untuk memprediksi kenaikan berat badan pada populasi dewasa, sehingga target yang dianalisis berbeda dari persentase lemak tubuh dan indeks massa otot apendikular. Dengan demikian, penelitian terdahulu telah menunjukkan potensi data kesehatan untuk pemodelan prediktif, tetapi belum secara langsung menjawab kebutuhan perbandingan antara model linier terregularisasi dan model nonlinier berbasis *boosting* dalam estimasi komposisi tubuh berbasis antropometri dengan target acuan DXA. Perbandingan tersebut penting karena data antropometri memiliki dua persoalan utama, yaitu keterkaitan antarfitur dan kemungkinan hubungan nonlinier terhadap target. Fitur seperti berat badan, tinggi badan, BMI, lingkaran pinggang, dan lingkaran lengan cenderung saling berkorelasi, sehingga Elastic Net dipilih sebagai model regresi linier terregularisasi untuk menjaga kestabilan koefisien pada kondisi multikolinearitas. Sebaliknya, hubungan antara fitur antropometri dan komposisi tubuh tidak selalu linier karena dapat dipengaruhi oleh jenis kelamin, usia, dan distribusi ukuran tubuh, sehingga CatBoost dipilih sebagai model nonlinier berbasis *boosting* yang mampu mempelajari interaksi antarfitur pada data tabular.

Oleh karena itu, perbandingan CatBoost dan Elastic Net dalam penelitian ini tidak hanya diarahkan untuk mencari model dengan nilai kesalahan prediksi paling rendah, tetapi juga untuk menilai apakah pendekatan nonlinier memberikan peningkatan performa dibandingkan model linier yang lebih sederhana dan stabil. Kebaruan penelitian ini terletak pada perbandingan langsung antara CatBoost dan Elastic Net dalam konteks regresi untuk estimasi komposisi tubuh berbasis antropometri dengan target acuan DXA, menggunakan fitur input yang sama pada kedua model. Penelitian ini mengestimasi dua target komposisi tubuh, yaitu persentase lemak tubuh dan indeks massa otot apendikular, menggunakan data National Health and Nutrition Examination Survey (NHANES) 2017–2018 dengan delapan fitur input antropometri. Performa model dievaluasi menggunakan *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), dan koefisien determinasi ( $R^2$ ). Hasil model kemudian diarahkan ke dalam prototipe sederhana berbasis Streamlit pada *Hugging Face Spaces* untuk memperlihatkan proses inferensi terhadap data antropometri baru.

## 2. METODOLOGI PENELITIAN

Penelitian ini menerapkan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM) untuk membandingkan performa CatBoost dan Elastic Net dalam estimasi komposisi tubuh berbasis antropometri. Pendekatan ini digunakan karena proses pemodelan regresi pada data ukuran tubuh memiliki tantangan, terutama adanya hubungan antarvariabel yang dapat menyebabkan multikolinearitas dan memengaruhi kestabilan model. Elastic Net digunakan sebagai model regresi linier terregularisasi yang mampu menangani fitur saling berkorelasi, sedangkan CatBoost digunakan sebagai model nonlinier berbasis *gradient boosting* yang mampu mempelajari pola

kompleks pada data tabular. CRISP-DM dipilih karena menyediakan tahapan analisis data yang terstruktur, mulai dari pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi, hingga penyebaran sederhana. Pendekatan CRISP-DM juga telah digunakan dalam pengembangan model *machine learning* pada data kesehatan karena alurnya membantu menjaga proses penelitian tetap sistematis, transparan, dan dapat dievaluasi pada setiap tahap [13]. Rangkaian prosedur penelitian ini ditampilkan pada Gambar 1.



**Gambar 1.** Alur penelitian berdasarkan tahapan CRISP-DM.

### 2.1 Business Understanding (Pemahaman Bisnis)

Tahap *business understanding* digunakan untuk merumuskan permasalahan penelitian ke dalam tujuan analisis data yang jelas dan terukur. Permasalahan utama dalam penelitian ini adalah perlunya pendekatan estimasi komposisi tubuh yang lebih praktis dibandingkan pemeriksaan berbasis alat khusus, tetapi tetap lebih informatif dibandingkan penggunaan Indeks Massa Tubuh (BMI) saja. Penelitian ini diarahkan untuk membandingkan performa algoritma regresi CatBoost dan Elastic Net dalam mengestimasi komposisi tubuh berdasarkan data antropometri sederhana. Variabel prediktor yang digunakan berupa ukuran fisik tubuh, sedangkan target yang diestimasi berupa komposisi tubuh berbasis DXA. Perencanaan analisis data dan pengujian kelayakan model pada tahap awal penting dilakukan agar proses pengembangan model berjalan sesuai dengan tujuan penelitian [14]. Hasil dari tahap ini menjadi dasar untuk menentukan kebutuhan data, strategi pemodelan, metrik evaluasi, dan bentuk implementasi sederhana. Model yang telah dievaluasi selanjutnya digunakan sebagai dasar pembuatan prototipe berbasis Streamlit untuk memperlihatkan proses inferensi terhadap data antropometri baru.

### 2.2 Data Understanding (Pemahaman Data)

Tahap *data understanding* dilakukan untuk memahami sumber, struktur, dan karakteristik data yang digunakan dalam penelitian. Data penelitian berasal dari National Health and Nutrition Examination Survey (NHANES) siklus 2017–2018 yang dikelola oleh National Center for Health Statistics (NCHS), bagian dari Centers for Disease Control and Prevention (CDC) [15]. NHANES merupakan dataset kesehatan berskala besar yang memuat berbagai informasi responden, sehingga data yang dibutuhkan untuk penelitian ini tidak tersedia dalam satu file, tetapi tersebar pada beberapa file sesuai jenis pemeriksaannya. Penelitian ini menggunakan tiga kelompok data utama, yaitu data demografi, data antropometri, dan data pengukuran komposisi tubuh berbasis *Dual-Energy X-ray Absorptiometry* (DXA). Pemilihan NHANES dilakukan karena dataset ini menyediakan informasi antropometri dan pengukuran DXA yang relevan sebagai dasar pemodelan estimasi komposisi tubuh. Eksplorasi analitik menggunakan data NHANES juga relevan karena dataset kesehatan berskala besar dapat digunakan untuk melatih algoritma *machine learning* dalam mengekstraksi hubungan antara metrik antropometri luar dengan komposisi tubuh internal [16].

Informasi yang dibutuhkan dalam penelitian ini diperoleh dari tiga file data, yaitu DEMO\_J, BMX\_J, dan DXX\_J. File DEMO\_J berasal dari kelompok *Demographics Data* dan digunakan untuk memperoleh informasi usia serta jenis kelamin responden. Sementara itu, file BMX\_J dan DXX\_J berasal dari kelompok *Examination Data*, dengan BMX\_J memuat hasil pemeriksaan antropometri dan DXX\_J memuat hasil pemeriksaan komposisi tubuh berbasis DXA. File DEMO\_J berisi data demografi dengan 9.254 baris dan 46 kolom, BMX\_J berisi data antropometri dengan 8.704 baris dan 21 kolom, sedangkan DXX\_J berisi data DXA atau *body composition* dengan 5.114 baris dan 93 kolom. Ketiga file tersebut digabungkan menggunakan kolom SEQN sebagai identitas responden, sehingga menghasilkan data gabungan sebanyak 5.114 baris dan 158 kolom. Jumlah data gabungan mengikuti file DXX\_J karena tidak semua responden NHANES memiliki data pemeriksaan DXA. Setelah proses penggabungan, dilakukan seleksi variabel dan penyesuaian nama kode teknis NHANES agar lebih mudah digunakan pada tahap *data preparation*, *modeling*, dan *evaluation*. Ringkasan variabel yang digunakan dalam penelitian ini ditampilkan pada Tabel 1.

**Tabel 1.** Variabel yang Digunakan dalam Penelitian

Jenis	Sumber Data	Nama Asli	Nama Normalisasi	Keterangan
Fitur	DEMO_J	RIDAGEYR	AGE_YEARS	Usia responden
Fitur	DEMO_J	RIAGENDR	GENDER	Jenis kelamin

Jenis	Sumber Data	Nama Asli	Nama Normalisasi	Keterangan
Fitur	BMX_J	BMXWT	WEIGHT_KG	Berat badan
Fitur	BMX_J	BMXHT	HEIGHT_CM	Tinggi badan
Fitur	BMX_J	BMXBMI	BMI	Indeks massa tubuh
Fitur	BMX_J	BMXWAIST	WAIST_CM	Lingkar pinggang
Fitur	BMX_J	BMXARMC	ARM_CIRCUMFERE NCE_CM	Lingkar lengan
Fitur	BMX_J	BMXLEG	UPPER_LEG LENG TH_CM	Panjang paha atas
Target	DXX_J	DXDTPF	BODY_FAT_PERCE NT	Persentase lemak tubuh (%)
Target	DXX_J	DXDLALE,DXDR ALE,DXDLLLE,D XDRLLE	APPENDICULAR_L EAN_MASS_INDEX	Indeks massa otot apendikular (kg/m <sup>2</sup> )

Berdasarkan Tabel 1, penelitian ini menggunakan delapan fitur input dan dua target. Fitur input berasal dari data demografi dan antropometri, sedangkan target berasal dari data DXA. Fitur yang digunakan meliputi usia, jenis kelamin, berat badan, tinggi badan, BMI, lingkar pinggang, lingkar lengan, dan panjang paha atas. Target yang digunakan adalah BODY\_FAT\_PERCENT (BFP) untuk merepresentasikan persentase lemak tubuh dan APPENDICULAR\_LEAN\_MASS\_INDEX (ALMI) untuk merepresentasikan indeks massa otot apendikular. Normalisasi nama variabel dilakukan agar kode teknis NHANES lebih mudah dipahami, seperti RIDAGEYR menjadi AGE\_YEARS, BMXWT menjadi WEIGHT\_KG, dan BMXBMI menjadi BMI. Kolom SEQN hanya digunakan sebagai kunci penggabungan antarfile data dan tidak digunakan sebagai fitur pemodelan karena berfungsi sebagai identitas responden, bukan karakteristik antropometri. Pemahaman terhadap karakteristik data ini menjadi dasar untuk menentukan strategi seleksi variabel, pembersihan data, dan persiapan data pada tahap berikutnya. BFP dinyatakan dalam satuan persen (%) dan diperoleh dari variabel DXDTPF pada data DXA. Sementara itu, ALMI dinyatakan dalam satuan kg/m<sup>2</sup> dan dihitung dari penjumlahan massa lean lengan kiri, lengan kanan, tungkai kiri, dan tungkai kanan yang dibagi dengan kuadrat tinggi badan dalam meter. Karena variabel DXDLALE, DXDRAL, DXDLLLE, dan DXDRLLE pada NHANES dinyatakan dalam gram, nilai tersebut terlebih dahulu dikonversi ke kilogram. Rumus perhitungan ALMI ditunjukkan pada Persamaan (1).

$$ALMI = \frac{(DXDLALE + DXDRAL + DXDLLLE + DXDRLLE)}{(1000 \times (HEIGHT_{CM} / 100)^2)} \quad (1)$$

Keterangan:

ALMI = appendicular lean mass index dalam kg/m<sup>2</sup>

DXDLALE = lean mass lengan kiri dalam gram

DXDRAL = lean mass lengan kanan dalam gram

DXDLLLE = lean mass tungkai kiri dalam gram

DXDRLLE = lean mass tungkai kanan dalam gram

HEIGHT\_CM = tinggi badan responden dalam sentimeter

### 2.3 Data Preparation (Persiapan Data)

Tahap *data preparation* dilakukan untuk menyiapkan dataset agar sesuai dengan kebutuhan pelatihan dan evaluasi model. Data demografi, antropometri, dan DXA terlebih dahulu digabungkan menggunakan ID responden sebagai kunci penghubung antarfile. Kriteria inklusi dalam penelitian ini adalah responden NHANES 2017–2018 yang memiliki data demografi, antropometri, dan DXA yang dapat digabungkan melalui SEQN, serta memiliki nilai target BFP dan ALMI secara lengkap. Responden dikeluarkan dari data pemodelan apabila tidak memiliki data DXA, memiliki nilai kosong pada salah satu target, atau tidak memiliki fitur antropometri yang digunakan dalam penelitian. Setelah kriteria tersebut diterapkan, data akhir yang digunakan untuk pemodelan berjumlah 3.551 baris dengan delapan fitur input dan dua target. Setelah proses penggabungan selesai, ID responden tidak digunakan sebagai fitur pemodelan karena hanya berfungsi sebagai penanda data dan tidak merepresentasikan karakteristik antropometri. Tahapan pembersihan dan normalisasi data memiliki peran penting karena kualitas data dapat memengaruhi performa estimasi algoritma kecerdasan buatan pada data kesehatan [17]. Ekstraksi dan standarisasi fitur antropometri secara terstruktur juga dapat meningkatkan kemampuan model dalam memprediksi persentase lemak tubuh dibandingkan hanya menggunakan nilai observasi mentah [10].

Langkah awal pembersihan data dilakukan dengan menghapus baris yang memiliki nilai kosong pada target BFP dan ALMI. Penghapusan target kosong dilakukan karena target merupakan variabel keluaran yang harus tersedia dalam proses *supervised learning*. Setelah proses tersebut, data yang digunakan untuk pemodelan berjumlah 3.551 baris dengan delapan fitur input dan dua target. Data kemudian dibagi dengan rasio 80:20 untuk data latih dan data uji, sehingga diperoleh 2.840 data latih dan 711 data uji. Penanganan *missing value* pada fitur dilakukan setelah proses pembagian data latih dan data uji untuk mencegah kebocoran informasi dari data uji ke

data latih. Pada jalur Elastic Net, fitur numerik diimputasi menggunakan median, kemudian distandardisasi menggunakan *StandardScaler*. Fitur kategorikal GENDER diimputasi menggunakan nilai yang paling sering muncul, lalu dikodekan menggunakan *OrdinalEncoder* agar dapat diproses oleh model regresi linier. Pada jalur CatBoost, fitur numerik diimputasi menggunakan median dari data latih, sedangkan fitur kategorikal diisi dengan nilai "Unknown" dan tetap dipertahankan sebagai fitur kategorikal. Perbedaan *preprocessing* pada kedua model ditampilkan pada Tabel 2.

**Tabel 2.** Perbedaan *Preprocessing* Elastic Net dan CatBoost

Tahap	Elastic Net	CatBoost
<i>Missing value</i> numerik	Imputasi <i>median</i>	Imputasi <i>median</i> dari data latih
<i>Missing value</i> kategorikal	Nilai paling sering muncul	Diisi dengan " <i>Unknown</i> "
<i>Encoding</i> kategorikal	<i>OrdinalEncoder</i>	Tidak dilakukan manual
<i>Scaling</i>	<i>StandardScaler</i>	Tidak dilakukan
Jumlah fitur akhir	8	8

Berdasarkan Tabel 2, kedua model menggunakan delapan fitur input yang sama, tetapi dengan perlakuan *preprocessing* yang disesuaikan dengan karakteristik masing-masing algoritma. Elastic Net memerlukan data dalam bentuk numerik dan sensitif terhadap perbedaan skala fitur, sehingga fitur numerik distandardisasi dan fitur kategorikal dikodekan secara numerik. CatBoost tidak memerlukan standardisasi dan dapat memproses fitur kategorikal secara langsung, sehingga fitur GENDER tetap dipertahankan sebagai fitur kategorikal. Standardisasi pada Elastic Net dilakukan dengan menghitung skor-z sebagaimana ditunjukkan pada Persamaan (2).

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Keterangan:

$z$  = Nilai standar hasil transformasi

$x$  = Nilai observasi asli dari sebuah fitur

$\mu$  = Nilai rata-rata fitur pada data latih

$\sigma$  = Standar deviasi fitur pada data latih

Pemeriksaan *outlier* dilakukan pada fitur numerik menggunakan metode *Interquartile Range* (IQR). Metode ini digunakan untuk mendeteksi nilai ekstrem berdasarkan jarak antara kuartil pertama dan kuartil ketiga. Rumus IQR ditunjukkan pada Persamaan (3), sedangkan batas bawah dan batas atas *outlier* ditunjukkan pada Persamaan (4) dan Persamaan (5).

$$IQR = Q_3 - Q_1 \tag{3}$$

$$Lower\ Bound = Q_1 - 1,5 \times IQR \tag{4}$$

$$Upper\ Bound = Q_3 + 1,5 \times IQR \tag{5}$$

Keterangan:

$Q_1$  = kuartil pertama

$Q_3$  = kuartil ketiga

IQR = rentang interkuartil

*Lower Bound* = batas bawah untuk mendeteksi outlier

*Upper Bound* = batas atas untuk mendeteksi outlier

Nilai ekstrem yang terdeteksi melalui metode IQR tidak langsung dihapus karena data antropometri dapat memiliki variasi biologis antarresponden. *Outlier* dipertahankan selama masih berada dalam rentang yang masuk akal dan tidak menunjukkan kesalahan pengukuran yang jelas. Dengan demikian, tahap *data preparation* menghasilkan data yang telah bersih dari *missing value* pada fitur dan target, tidak menggunakan kolom identitas sebagai fitur, serta siap digunakan dalam proses pemodelan.

## 2.4 Modeling (Pemodelan)

Data yang telah melewati proses pembersihan dan persiapan selanjutnya digunakan pada tahap pelatihan model. Penelitian ini menerapkan pendekatan komparatif antara dua algoritma regresi yang memiliki karakteristik berbeda, yaitu Elastic Net dan CatBoost. Pemilihan kedua algoritma tersebut didasarkan pada pendekatan perbandingan antara model regresi linier terregularisasi dan model regresi berbasis pohon keputusan pada data terstruktur [18]. Elastic Net digunakan untuk merepresentasikan pendekatan linier yang mampu menangani multikolinearitas antarfitur, sedangkan CatBoost digunakan untuk merepresentasikan pendekatan nonlinier berbasis *ensemble* yang mampu mempelajari interaksi kompleks pada data tabular. Implementasi *ensemble learning* juga banyak digunakan dalam pemodelan data kesehatan karena mampu membantu menangani pola data yang kompleks [19]. Dengan demikian, tahap pemodelan dilakukan untuk melihat perbedaan kemampuan kedua algoritma dalam mengestimasi komposisi tubuh berbasis antropometri.

Pada tahap pemodelan, kedua target diprediksi menggunakan pendekatan *multi-output regression*. Elastic Net diterapkan menggunakan *MultiOutputRegressor* dengan model dasar *ElasticNetCV*, sehingga BFP dan ALMI diprediksi melalui model regresi yang dibangun untuk masing-masing target. Pada Elastic Net, parameter *l1\_ratio*

dicari pada nilai 0,1; 0,3; 0,5; 0,7; dan 0,9, sedangkan  $\alpha$  dicari menggunakan 50 nilai logaritmik pada rentang  $10^{-3}$  sampai  $10^3$ . Proses pemilihan parameter Elastic Net dilakukan menggunakan *5-fold cross-validation*, dengan *random\_state* sebesar 42 dan *max\_iter* sebesar 10000. Sementara itu, CatBoost diterapkan sebagai model regresi *multi-output* dengan *loss\_function* dan *eval\_metric* berupa MultiRMSE, sehingga kedua target dapat dipelajari dalam satu proses pelatihan. Parameter CatBoost yang digunakan meliputi *iterations* sebesar 3000, *learning\_rate* sebesar 0,03, *depth* sebesar 4, *l2\_leaf\_reg* sebesar 10, *random\_strength* sebesar 1, *bootstrap\_type* Bernoulli, *subsample* sebesar 0,8, *rsm* sebesar 0,8, dan *random\_seed* sebesar 42. Pada proses pelatihan CatBoost, sebesar 15% dari data latih digunakan sebagai data validasi internal dengan *early\_stopping\_rounds* sebesar 100 dan *use\_best\_model* bernilai True.

### 2.4.1 Elastic Net Regression

Pemodelan Elastic Net mewakili kelompok regresi linier terregularisasi. Model linier dapat mengalami penurunan kestabilan ketika dihadapkan pada fitur yang saling berkorelasi kuat. Elastic Net digunakan sebagai solusi untuk mengurangi dampak multikolinearitas tanpa harus membuang dimensi data secara manual [20]. Cara kerja algoritma ini bertumpu pada penggabungan dua penalti matematis, yaitu Lasso dan Ridge, selama proses pelatihan. Penalti Lasso membantu menyusutkan koefisien fitur yang kurang relevan, sedangkan penalti Ridge membantu menstabilkan koefisien pada fitur yang saling berkorelasi. Persamaan fungsi objektif Elastic Net ditunjukkan pada Persamaan (6).

$$\hat{\beta} = \arg \frac{\min}{\beta} \left( \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right) \quad (6)$$

Keterangan:

$y$  = Vektor target (komposisi lemak aktual)

$X$  = Matriks multidimensi dari fitur antropometri

$\beta$  = Estimator koefisien bobot regresi

$\lambda_1$  = Parameter penalti regularisasi  $L_1$  (Lasso) untuk penyusutan fitur non-esensial

$\lambda_2$  = Parameter penalti regularisasi  $L_2$  (Ridge) untuk meredam efek multikolinearitas

Mekanisme penggabungan kedua penalti tersebut dikendalikan oleh parameter pencampuran yang menyeimbangkan pengaruh *Lasso* dan *Ridge* dalam proses regularisasi [21]. Dalam penelitian ini, Elastic Net digunakan sebagai model pembandingan yang stabil dan lebih mudah diinterpretasikan pada data antropometri yang memiliki hubungan antarfitur cukup kuat.

### 2.4.2 CatBoost

CatBoost merupakan algoritma *Gradient Boosting Decision Tree* (GBDT) yang membangun model secara sekuensial melalui kumpulan pohon keputusan. Berbeda dengan pendekatan linier, CatBoost mempelajari pola data dengan memperbaiki kesalahan prediksi secara bertahap pada setiap iterasi. Model ini dapat menangani data tabular dan fitur kategorikal tanpa memerlukan proses penyandian manual yang berlebihan. Karakteristik tersebut membuat CatBoost sesuai digunakan pada data kesehatan yang memiliki struktur kompleks dan interaksi antarvariabel [22]. Selain itu, CatBoost juga mendukung interpretasi model melalui pengukuran kontribusi fitur atau *feature importance*, sehingga peran setiap variabel antropometri terhadap hasil estimasi dapat dianalisis [23]. Prediksi aditif pada CatBoost ditunjukkan pada Persamaan (7).

$$F_M(x) = \sum_{m=1}^M \gamma_m h_m(x) \quad (7)$$

Keterangan:

$F_M(x)$  = Hasil estimasi prediksi tingkat lemak akhir

$M$  = Total iterasi atau jumlah maksimal pembangunan pohon dasar

$\gamma_m$  = Nilai laju pembelajaran (learning rate)

$h_m(x)$  = Skor estimasi pohon pemisah dasar ke- $m$

Proses pelatihan CatBoost berfokus pada perbaikan galat secara berulang agar model mampu menangkap pola hubungan antara fitur antropometri dan target komposisi tubuh. Model berbasis pohon keputusan ini juga digunakan untuk melihat kontribusi fitur melalui *feature importance*, sehingga hasil pemodelan tidak hanya dinilai dari performa prediksi, tetapi juga dari variabel yang berperan dalam proses estimasi. Analisis *feature importance* dapat membantu memahami pergeseran dan kontribusi variabel dalam pemodelan *machine learning*, sehingga interpretasi model menjadi lebih mudah dilakukan [24].

## 2.5 Evaluation (Evaluasi)

Tahap *evaluation* dilakukan untuk mengukur dan membandingkan performa prediksi dari model Elastic Net dan CatBoost. Keandalan fungsi prediktif diukur menggunakan tiga metrik evaluasi regresi, yaitu *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi ( $R^2$ ). Ketiga metrik tersebut digunakan untuk melihat seberapa jauh hasil estimasi model berbeda dari nilai aktual komposisi tubuh [25]. RMSE digunakan

untuk mengukur besar galat prediksi dengan memberikan penalti lebih besar pada kesalahan yang tinggi, sedangkan MAE digunakan untuk menghitung rata-rata kesalahan absolut antara nilai aktual dan nilai prediksi.  $R^2$  digunakan untuk mengetahui kemampuan model dalam menjelaskan variasi data target. Rumus RMSE ditunjukkan pada Persamaan (8), MAE pada Persamaan (9), dan  $R^2$  pada Persamaan (10).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

Keterangan:

$n$  merupakan jumlah total data observasi

$y_i$  merepresentasikan nilai komposisi lemak aktual

$\hat{y}_i$  merupakan tebakan prediksi model

$\bar{y}$  adalah nilai rata-rata target aktual

Evaluasi dilakukan pada data latih dan data uji untuk melihat kemampuan model dalam mempelajari pola data serta menguji kemampuan generalisasinya terhadap data yang belum pernah digunakan pada proses pelatihan. Perbandingan hasil evaluasi antara data latih dan data uji juga digunakan untuk mengidentifikasi kemungkinan *overfitting*, yaitu kondisi ketika model memiliki performa sangat baik pada data latih, tetapi mengalami penurunan pada data uji. Evaluasi utama model dilakukan menggunakan skema *train-test split*, dengan pembagian data latih dan data uji sebesar 80:20. Pada Elastic Net, *5-fold cross-validation* digunakan dalam proses pemilihan parameter regularisasi melalui ElasticNetCV. Sementara itu, pada CatBoost, sebagian data latih digunakan sebagai *validation split* untuk proses *early stopping*. *Validation split* tersebut hanya berasal dari data latih dan tidak mengubah data uji yang digunakan pada evaluasi akhir. Data uji tidak digunakan selama proses pelatihan dan hanya digunakan pada tahap evaluasi akhir model. Selain evaluasi menggunakan RMSE, MAE, dan  $R^2$ , penelitian ini juga melakukan analisis *feature importance* sebagai pendukung interpretasi model. Analisis ini digunakan untuk melihat kontribusi masing-masing variabel antropometri terhadap hasil estimasi komposisi tubuh. Model dengan nilai RMSE dan MAE yang lebih rendah serta nilai  $R^2$  yang lebih tinggi pada data uji dinilai memiliki performa estimasi yang lebih baik. Hasil evaluasi tersebut digunakan sebagai dasar untuk menentukan model yang lebih sesuai dalam estimasi komposisi tubuh berbasis antropometri.

## 2.6 Deployment

Tahap *deployment* dilakukan sebagai penerapan sederhana dari model hasil penelitian ke dalam prototipe berbasis web. Model yang telah dievaluasi disimpan dalam format Joblib agar dapat dipanggil kembali tanpa menjalankan proses pelatihan ulang. Integrasi sistem dilakukan dengan memasukkan model ke dalam antarmuka berbasis Streamlit, kemudian diarahkan untuk dijalankan melalui Hugging Face Spaces sebagai media hosting prototipe. Implementasi model ke dalam antarmuka berbasis web dapat membantu proses inferensi menjadi lebih mudah diuji dan digunakan dalam bentuk prototipe [26]. Dalam penelitian ini, *deployment* digunakan untuk memperlihatkan proses inferensi terhadap data antropometri baru, seperti usia, jenis kelamin, berat badan, tinggi badan, BMI, lingkar pinggang, lingkar lengan atas, dan panjang paha atas. Prototipe ini digunakan sebagai media simulasi inferensi, sehingga model dapat diuji kembali pada data baru tanpa melakukan pelatihan ulang. Penggunaan prototipe masih terbatas pada implementasi hasil penelitian dan belum diarahkan untuk penggunaan klinis langsung.

## 3. HASIL DAN PEMBAHASAN

Hasil dan pembahasan menyajikan proses pelatihan, evaluasi, serta analisis perbandingan model regresi Elastic Net dan CatBoost dalam mengestimasi komposisi tubuh berbasis antropometri. Evaluasi dilakukan terhadap dua target, yaitu BFP dan ALMI. Hasil yang diperoleh dianalisis menggunakan metrik MAE, RMSE, dan  $R^2$ , kemudian didukung dengan analisis *feature importance* untuk melihat kontribusi fitur antropometri terhadap hasil estimasi model.

### 3.1 Hasil Modeling

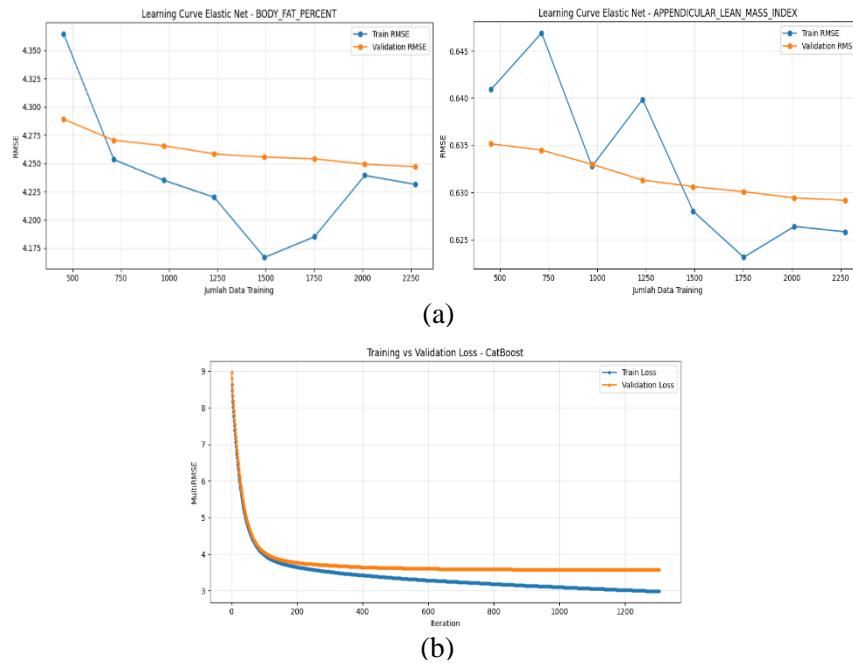
Tahap *training* dilakukan menggunakan data hasil *data preparation* yang telah dibagi menjadi data latih dan data uji. Data latih berjumlah 2.840 data, sedangkan data uji berjumlah 711 data. Elastic Net dan CatBoost menggunakan delapan fitur input yang sama, yaitu AGE\_YEARS, GENDER, WEIGHT\_KG, HEIGHT\_CM, BMI, WAIST\_CM, ARM\_CIRCUMFERENCE\_CM, dan UPPER\_LEG\_LENGTH\_CM. Selain kedua model

utama tersebut, penelitian ini juga menambahkan BMI-only Linear Regression sebagai baseline sederhana. Baseline ini hanya menggunakan fitur BMI untuk melihat sejauh mana model berbasis beberapa fitur antropometri mampu memberikan peningkatan performa dibanding pendekatan berbasis BMI saja. Penyamaan jumlah fitur pada Elastic Net dan CatBoost dilakukan agar perbandingan performa antara kedua model utama dapat dilakukan secara lebih adil. Meskipun menggunakan fitur yang sama, kedua model memiliki perlakuan *training* yang berbeda sesuai karakteristik algoritmanya. Elastic Net digunakan sebagai model regresi linier terregularisasi yang membutuhkan data numerik dengan skala sebanding, sedangkan CatBoost digunakan sebagai model *gradient boosting* yang dapat memproses fitur kategorikal dan mempelajari pola data secara iteratif melalui pembentukan pohon keputusan. Ringkasan data yang digunakan pada tahap *training* ditampilkan pada Tabel 3.

**Tabel 3.** Data Masukan pada Tahap Modeling

Model	Data Latih	Data Uji	Jumlah Fitur	Jumlah Target
BMI-only Linear Regression	2840	711	1	2
Elastic Net	2840	711	8	2
CatBoost	2840	711	8	2

Berdasarkan Tabel 3, seluruh model menggunakan data latih dan data uji yang sama. BMI-only Linear Regression digunakan sebagai *baseline* sederhana dengan satu fitur input, yaitu BMI, sedangkan Elastic Net dan CatBoost menggunakan delapan fitur input yang sama. Target yang digunakan terdiri dari dua variabel, yaitu BFP dan ALMI, sehingga proses *training* dapat digunakan untuk melihat kemampuan masing-masing model pada kondisi pembagian data yang setara. Visualisasi proses *training* pada kedua model ditunjukkan pada Gambar 2. Gambar 2(a) menunjukkan *learning curve* Elastic Net pada target BFP, sedangkan Gambar 2(b) menunjukkan grafik *training* CatBoost berdasarkan nilai *MultiRMSE* pada data latih dan validasi. Visualisasi ini digunakan untuk memberikan gambaran awal mengenai kestabilan proses pelatihan dan kemungkinan perbedaan performa antara data latih dan data validasi.



**Gambar 2.** Visualisasi proses training model: (a) learning curve Elastic Net dan (b) grafik training CatBoost berdasarkan MultiRMSE

Pada Gambar 2(a), nilai RMSE data latih dan validasi Elastic Net terlihat berada pada rentang yang relatif berdekatan, sehingga menunjukkan proses *training* yang stabil dan tidak menunjukkan perbedaan ekstrem antara data latih dan validasi. Pola tersebut berkaitan dengan karakteristik Elastic Net yang menggunakan regularisasi untuk menjaga kestabilan model pada fitur yang saling berkorelasi. Pada Gambar 2(b), nilai *train loss* dan *validation loss* CatBoost mengalami penurunan tajam pada iterasi awal, kemudian mulai stabil pada iterasi berikutnya. Penurunan tersebut menunjukkan bahwa CatBoost mampu memperbaiki galat prediksi secara bertahap selama proses *boosting*. Selisih antara *train loss* dan *validation loss* tetap perlu diperhatikan karena dapat menunjukkan kecenderungan model dalam mempelajari data latih lebih kuat dibandingkan data validasi. Namun, *validation loss* yang tidak mengalami kenaikan tajam menunjukkan bahwa performa pada data validasi masih terkendali.

### 3.2 Hasil Evaluasi Model

Evaluasi model dilakukan untuk mengetahui kemampuan BMI-only Linear Regression sebagai *baseline*, Elastic Net, dan CatBoost dalam mengestimasi dua target komposisi tubuh, yaitu BFP dan ALMI. Evaluasi dilakukan pada data latih dan data uji menggunakan metrik MAE, RMSE, dan R<sup>2</sup>. MAE dan RMSE digunakan untuk mengukur besar kesalahan prediksi model, sedangkan R<sup>2</sup> digunakan untuk mengetahui kemampuan model dalam menjelaskan variasi data target. Berdasarkan hasil evaluasi, CatBoost memperoleh performa lebih baik dibandingkan Elastic Net pada kedua target data uji. Pada target BFP, CatBoost menghasilkan MAE sebesar 2,725529, RMSE sebesar 3,507958, dan R<sup>2</sup> sebesar 0,836537, sedangkan Elastic Net memperoleh MAE sebesar 3,280431, RMSE sebesar 4,196854, dan R<sup>2</sup> sebesar 0,766031. Hasil tersebut menunjukkan bahwa CatBoost memiliki kesalahan prediksi yang lebih rendah dan kemampuan menjelaskan variasi persentase lemak tubuh yang lebih tinggi. Pada target ALMI, CatBoost juga lebih unggul dengan MAE sebesar 0,452253, RMSE sebesar 0,581853, dan R<sup>2</sup> sebesar 0,889122, sedangkan Elastic Net memperoleh MAE sebesar 0,495618, RMSE sebesar 0,620910, dan R<sup>2</sup> sebesar 0,873737. Selisih performa pada target ini lebih kecil dibandingkan target persentase lemak tubuh, tetapi CatBoost tetap menunjukkan hasil yang lebih baik.

Secara rata-rata pada data uji, CatBoost memperoleh Mean MAE sebesar 1,588891, Mean RMSE sebesar 2,044905, dan Mean R<sup>2</sup> sebesar 0,862830. Sementara itu, Elastic Net memperoleh Mean MAE sebesar 1,888025, Mean RMSE sebesar 2,408882, dan Mean R<sup>2</sup> sebesar 0,819884. Perbandingan data latih dan data uji menunjukkan bahwa Elastic Net memiliki performa yang relatif stabil karena nilai R<sup>2</sup> pada data latih dan data uji tidak berbeda jauh, tetapi nilai kesalahannya masih tinggi dibandingkan CatBoost. Pada CatBoost, nilai R<sup>2</sup> data latih lebih tinggi dibandingkan data uji, tetapi penurunannya tidak ekstrem karena nilai R<sup>2</sup> data uji tetap lebih tinggi dibandingkan Elastic Net pada kedua target. Kondisi tersebut menunjukkan bahwa CatBoost masih mampu melakukan generalisasi terhadap data uji, meskipun model mempelajari pola data latih dengan lebih kuat. Perbedaan nilai evaluasi ini juga menunjukkan bahwa model terbaik tidak hanya ditentukan dari kemampuan mempelajari data latih, tetapi dari kemampuannya mempertahankan performa pada data yang belum pernah digunakan selama proses *training*. Dengan demikian, CatBoost ditetapkan sebagai model terbaik dalam penelitian ini, sedangkan Elastic Net tetap menunjukkan performa yang stabil sebagai model regresi linier pembanding. Perbedaan performa tersebut menunjukkan bahwa CatBoost lebih konsisten menghasilkan kesalahan prediksi yang lebih rendah pada kedua target, terutama pada BFP yang memiliki selisih evaluasi lebih besar dibandingkan ALMI. Sementara itu, Elastic Net tetap menunjukkan hasil yang stabil karena selisih performa antara data latih dan data uji tidak terlalu besar. Kondisi ini memperlihatkan bahwa CatBoost lebih unggul dari sisi akurasi estimasi, sedangkan Elastic Net masih relevan sebagai model pembanding yang lebih sederhana dan stabil. Hasil evaluasi model secara lengkap ditampilkan pada Tabel 4.

**Tabel 4.** Hasil Evaluasi Model pada Hasil Data Latih dan Data Uji

Model	Data	Target	MAE	RMSE	R <sup>2</sup>
BMI-only Linear Regression	Latih	BFP	5,842604	6,883673	0,370575
BMI-only Linear Regression	Latih	ALMI	0,932540	1,122645	0,609516
BMI-only Linear Regression	Uji	BFP	6,016857	7,041404	0,341387
BMI-only Linear Regression	Uji	ALMI	0,952384	1,125390	0,585215
Elastic Net	Latih	BFP	3,358008	4,233127	0,761973
Elastic Net	Latih	ALMI	0,491307	0,626202	0,878508
Elastic Net	Uji	BFP	3,280431	4,196854	0,766031
Elastic Net	Uji	ALMI	0,495618	0,620910	0,873737
CatBoost	Latih	BFP	2,409168	3,065679	0,875159
CatBoost	Latih	ALMI	0,401477	0,525734	0,914365
CatBoost	Uji	BFP	2,725529	3,507958	0,836537
CatBoost	Uji	ALMI	0,452253	0,581853	0,889122

Sebagai pembanding sederhana, BMI-only Linear Regression digunakan sebagai *baseline* karena BMI merupakan indikator antropometri yang umum digunakan, tetapi memiliki keterbatasan dalam membedakan persentase lemak tubuh dan massa otot. Berdasarkan hasil evaluasi rata-rata pada data uji, BMI-only Linear Regression memperoleh Mean MAE sebesar 3,484621, Mean RMSE sebesar 4,083397, dan Mean R<sup>2</sup> sebesar 0,463301. Nilai tersebut lebih rendah dibandingkan Elastic Net dengan Mean R<sup>2</sup> sebesar 0,819884 dan CatBoost dengan Mean R<sup>2</sup> sebesar 0,862830. Hasil ini menunjukkan bahwa penggunaan beberapa fitur antropometri memberikan informasi prediktif yang lebih baik dibandingkan pendekatan berbasis BMI saja.

### 3.3 Pembahasan Perbandingan Model

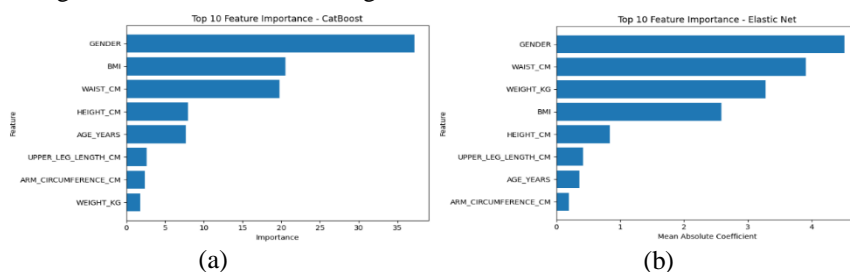
Berdasarkan hasil evaluasi, CatBoost menunjukkan performa lebih baik dibandingkan Elastic Net pada kedua target estimasi komposisi tubuh. Keunggulan tersebut terlihat dari nilai MAE dan RMSE yang lebih rendah serta nilai R<sup>2</sup> yang lebih tinggi pada data uji. Hasil ini menunjukkan bahwa CatBoost lebih mampu menangkap pola hubungan antara fitur antropometri dan target komposisi tubuh, terutama karena karakteristiknya sebagai model *gradient boosting* berbasis pohon keputusan yang dapat mempelajari hubungan nonlinier dan interaksi antarfitur. Pada data antropometri, hubungan antara GENDER, BMI, WAIST\_CM, WEIGHT\_KG, dan target komposisi

tubuh tidak selalu membentuk pola linier sederhana. Pengaruh BMI terhadap persentase lemak tubuh, misalnya, dapat berbeda berdasarkan jenis kelamin, usia, dan ukuran lingkaran pinggang. Pola seperti ini lebih mudah dipelajari oleh CatBoost karena model membangun pohon keputusan secara bertahap untuk memperbaiki galat prediksi pada setiap iterasi. Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa CatBoost mampu menghasilkan performa tinggi pada pemodelan data kesehatan terstruktur [27]. Sementara itu, Elastic Net tetap menunjukkan performa yang baik dan stabil, meskipun belum mampu melampaui CatBoost sebagai model terbaik. Stabilitas Elastic Net terlihat dari selisih nilai  $R^2$  antara data latih dan data uji yang relatif kecil pada kedua target. Kondisi tersebut sesuai dengan karakteristik Elastic Net sebagai model regresi linier terregularisasi yang menggabungkan penalti Lasso dan Ridge untuk menjaga kestabilan koefisien ketika fitur *input* saling berkorelasi. Pada data antropometri, fitur seperti berat badan, BMI, tinggi badan, dan lingkaran pinggang cenderung memiliki keterkaitan satu sama lain, sehingga regularisasi pada Elastic Net membantu menjaga kestabilan model. Penggunaan Elastic Net pada pemodelan prediktif data kesehatan juga didukung oleh penelitian sebelumnya yang menunjukkan bahwa model ini dapat mempertahankan kemampuan prediksi melalui proses validasi [28]. Keberadaan BMI-only Linear Regression sebagai *baseline* memperkuat temuan bahwa BMI saja belum cukup untuk mengestimasi komposisi tubuh secara akurat. Elastic Net dan CatBoost menghasilkan performa lebih baik karena menggunakan beberapa fitur antropometri, seperti usia, jenis kelamin, lingkaran pinggang, lingkaran lengan, dan panjang paha atas, selain BMI. Dengan demikian, peningkatan performa kedua model utama dibandingkan *baseline* menunjukkan bahwa kombinasi fitur antropometri sederhana dapat memberikan estimasi komposisi tubuh yang lebih informatif daripada penggunaan BMI saja.

Perbedaan performa antara kedua model menunjukkan bahwa hubungan antara fitur antropometri dan target komposisi tubuh tidak sepenuhnya bersifat linier. Elastic Net mampu memberikan hasil yang stabil sebagai model linier pembanding, tetapi pendekatan linier memiliki keterbatasan dalam menangkap pola interaksi yang lebih kompleks. Sebaliknya, CatBoost lebih fleksibel dalam mempelajari pola nonlinier pada data tabular, sehingga menghasilkan kesalahan prediksi yang lebih rendah pada kedua target. Dengan demikian, hasil penelitian ini menunjukkan bahwa CatBoost lebih sesuai digunakan untuk estimasi komposisi tubuh berbasis antropometri, sedangkan Elastic Net tetap memiliki nilai penting sebagai model pembanding yang stabil dan lebih mudah diinterpretasikan. Hasil perbandingan ini juga menunjukkan bahwa karakteristik target memengaruhi besar kecilnya selisih performa antar model. Pada target BFP, perbedaan nilai MAE, RMSE, dan  $R^2$  antara CatBoost dan Elastic Net terlihat lebih besar dibandingkan target ALMI. Kondisi tersebut mengindikasikan bahwa persentase lemak tubuh memiliki pola hubungan yang lebih kompleks terhadap fitur antropometri, sehingga lebih diuntungkan oleh kemampuan CatBoost dalam menangkap interaksi nonlinier. Sebaliknya, selisih performa pada target ALMI relatif lebih kecil, yang menunjukkan bahwa hubungan antara fitur antropometri dan indeks massa otot apendikular masih dapat dipelajari cukup baik oleh pendekatan linier terregularisasi. Dengan demikian, pemilihan algoritma tidak hanya ditentukan oleh nilai rata-rata evaluasi, tetapi juga perlu memperhatikan karakteristik target yang diestimasi.

### 3.4 Feature Importance

Analisis *feature importance* dilakukan untuk mengetahui fitur antropometri yang paling berkontribusi terhadap hasil estimasi model. Visualisasi *feature importance* ditampilkan pada Gambar 3. Gambar 3(a) menunjukkan urutan fitur penting pada CatBoost, sedangkan Gambar 3(b) menunjukkan urutan fitur penting pada Elastic Net. Pada CatBoost, *feature importance* diperoleh dari kontribusi fitur dalam proses pembentukan pohon keputusan dan boosting. Sementara itu, pada Elastic Net, *feature importance* dilihat berdasarkan nilai koefisien absolut rata-rata setelah proses *preprocessing*. Perbedaan dasar perhitungan tersebut menyebabkan urutan fitur penting pada kedua model tidak sepenuhnya sama. Berdasarkan Gambar 3, fitur GENDER menjadi fitur paling berpengaruh pada kedua model. Tingginya kontribusi fitur ini dapat dikaitkan dengan perbedaan biologis antara laki-laki dan perempuan dalam komposisi tubuh, terutama pada distribusi lemak dan massa otot. Perbedaan tersebut membuat GENDER menjadi informasi penting bagi model dalam membedakan pola target BFP dan ALMI. Pada CatBoost, fitur penting berikutnya adalah BMI dan WAIST\_CM. Fitur BMI merepresentasikan hubungan antara berat badan dan tinggi badan, sehingga dapat memberi gambaran umum mengenai ukuran tubuh. Fitur WAIST\_CM merepresentasikan lingkaran pinggang yang berkaitan dengan distribusi lemak tubuh, sehingga relevan untuk membantu estimasi persentase lemak tubuh. Pada Elastic Net, fitur yang paling berpengaruh setelah GENDER adalah WAIST\_CM, WEIGHT\_KG, dan BMI. Urutan ini dapat terjadi karena Elastic Net menilai kontribusi fitur berdasarkan hubungan linier antara fitur dan target melalui besar koefisien model.



**Gambar 3.** *Feature importance* model: (a) CatBoost dan (b) Elastic Net.

Perbedaan urutan fitur penting antara CatBoost dan Elastic Net terjadi karena mekanisme pembelajaran kedua model berbeda. CatBoost merupakan model berbasis pohon keputusan, sehingga suatu fitur dapat dianggap penting ketika sering digunakan dalam proses pemisahan data dan membantu menurunkan kesalahan prediksi selama boosting. Karena itu, CatBoost mampu menangkap interaksi nonlinier antarfitur, misalnya hubungan antara GENDER, BMI, dan WAIST\_CM terhadap komposisi tubuh. Sementara itu, Elastic Net merupakan model regresi linier terregularisasi, sehingga kontribusi fitur dihitung berdasarkan besar koefisien linier setelah proses *preprocessing*. Fitur dengan hubungan linier yang lebih kuat terhadap target akan memiliki nilai *importance* lebih tinggi. Dengan demikian, meskipun fitur utama seperti GENDER, BMI, WAIST\_CM, dan WEIGHT\_KG sama-sama muncul sebagai fitur penting, urutan kontribusinya dapat berbeda karena cara model mempelajari pola data tidak sama. Analisis ini digunakan sebagai interpretasi kontribusi fitur pada model, bukan sebagai bukti hubungan sebab-akibat secara medis. Nilai *feature importance* hanya menunjukkan peran fitur dalam membantu model menghasilkan prediksi berdasarkan pola data, sehingga tidak dapat digunakan sebagai dasar diagnosis, rekomendasi klinis, atau kesimpulan kausal mengenai perubahan komposisi tubuh.

### 3.5 Deployment

Tahap *deployment* menghasilkan prototipe berbasis Streamlit yang diarahkan untuk dijalankan melalui Hugging Face Spaces sebagai media simulasi inferensi. Prototipe ini menerima *input* data antropometri, seperti usia, jenis kelamin, berat badan, tinggi badan, BMI, lingkar pinggang, lingkar lengan atas, dan panjang paha atas. Berdasarkan Gambar 4, aplikasi menampilkan form *input* pada bagian kiri dan hasil estimasi pada bagian kanan untuk dua target, yaitu persentase lemak tubuh dan ALMI. Prototipe ini digunakan untuk menunjukkan bahwa model hasil penelitian dapat dipanggil kembali dan digunakan pada data baru tanpa melakukan proses *training* ulang. Prototipe ini masih terbatas sebagai media simulasi inferensi hasil penelitian dan tidak ditujukan sebagai alat diagnosis atau dasar rekomendasi klinis.



**Gambar 4.** Tampilan prototipe *deployment* untuk input data antropometri dan hasil prediksi komposisi tubuh.

## 4. KESIMPULAN

Berdasarkan hasil penelitian, CatBoost menunjukkan performa yang lebih baik dibandingkan Elastic Net dalam estimasi komposisi tubuh berbasis antropometri pada dua target yang digunakan. Keunggulan tersebut terlihat dari nilai kesalahan prediksi yang lebih rendah dan nilai  $R^2$  yang lebih tinggi pada data uji. Hasil ini menunjukkan bahwa hubungan antara fitur antropometri dan komposisi tubuh cenderung tidak sepenuhnya linier, sehingga model berbasis *gradient boosting* lebih mampu menangkap pola nonlinier dan interaksi antarfitur. Meskipun demikian, Elastic Net tetap memberikan hasil yang stabil dan dapat digunakan sebagai model pembanding karena mampu menjaga kestabilan regresi pada fitur antropometri yang saling berkorelasi. Selain itu, perbandingan dengan *baseline* BMI-only Linear Regression menunjukkan bahwa penggunaan beberapa fitur antropometri memberikan informasi prediktif yang lebih baik dibandingkan penggunaan BMI saja. Analisis *feature importance* menunjukkan bahwa GENDER, BMI, WAIST\_CM, dan WEIGHT\_KG menjadi fitur yang berperan penting dalam proses estimasi, tetapi interpretasi tersebut hanya menunjukkan kontribusi prediktif dan bukan hubungan sebab-akibat medis. Prototipe berbasis Streamlit juga memperlihatkan bahwa model hasil penelitian dapat digunakan untuk simulasi inferensi terhadap data antropometri baru, tetapi tidak ditujukan sebagai alat diagnosis atau rekomendasi klinis. Penelitian ini memiliki beberapa keterbatasan. Dataset yang digunakan hanya berasal dari NHANES 2017–2018, sehingga kemampuan generalisasi model terhadap populasi lain masih perlu diuji lebih lanjut. Model juga belum divalidasi menggunakan dataset eksternal dan belum melalui validasi klinis, sehingga hasil estimasi masih terbatas sebagai keluaran penelitian berbasis data. Penelitian selanjutnya dapat melakukan validasi eksternal menggunakan data antropometri nyata dari populasi lokal atau daerah yang berbeda dengan tetap mempertahankan variabel input yang sama. Pengujian tersebut diperlukan untuk mengetahui apakah model yang dikembangkan dari dataset NHANES 2017–2018 tetap mampu menghasilkan estimasi yang stabil pada karakteristik populasi yang berbeda.

## REFERENCES

- [1] A. M. González-Martin *et al.*, “Anthropometric Measurements for Predicting Low Appendicular Lean Mass Index for the Diagnosis of Sarcopenia : A Machine Learning Model,” *J. Funct. Morphol. Kinesiol.*, vol. 10, no. 3, pp. 1–22, 2025, doi: 10.3390/jfmk10030276.
- [2] K. Yasmin, “Cardiometabolic Risk Factors in South Asians: An Epidemiological and Anthropological Study in an Urban Populace of Eastern India.,” *medRxiv*, 2024, doi: 10.1101/2024.12.06.24318633.
- [3] L. Zeng, X. Guo, H. Wu, and C. Huang, “Machine learning-based estimation of trunk fat percentage and its association with cardiometabolic risk leveraging two large national cohorts,” *Front. Nutr.*, vol. 13, no. January, pp. 1–10, 2026, doi: 10.3389/fnut.2026.1715570.
- [4] N. Birk, B. Kulkarni, S. Bhogadi, A. Aggarwal, and G. K. Walia, “Machine learning-based equations for improved body composition estimation in Indian adults,” *PLOS Digit. Heal.*, vol. 4, no. 6, pp. 1–14, 2025, doi: 10.1371/journal.pdig.0000671.
- [5] T. Yousefi and Ö. Varlıklar, “New hybrid multi-objective feature selection : Boruta-XGBoost,” *PeerJ Comput. Sci.*, vol. 12, p. e3463, 2026, doi: 10.7717/peerj-cs.3463.
- [6] Z. F. Ghareeb and S. A. S. Al-temimi, “A new shrinkage method for higher dimensions regression model to remedy of multicollinearity problem,” *Period. Eng. Nat. Sci.*, vol. 11, no. 3, pp. 18–28, 2023, [Online]. Available: <https://pen.ius.edu.ba/index.php/pen/article/view/3550>
- [7] A. R. Nur, A. K. Jaya, and S. Siswanto, “Comparative Analysis of Ridge , LASSO , and Elastic Net Regularization Approaches in Handling Multicollinearity for Infant Mortality Data in South Sulawesi,” *J. Mat. Stat. dan Komputasi*, vol. 20, no. 2, pp. 311–319, 2024, doi: 10.20956/j.v20i2.31632.
- [8] K. Harefa and J. Priambodo, “Implementation of the LightGBM – CatBoost Ensemble Method for Obesity Risk Classification in Productive Age,” *Bull. Comput. Sci. Res.*, vol. 6, no. 1, pp. 531–538, 2025, doi: 10.47065/bulletincsr.v6i1.930.
- [9] A. Maulana, R. Perucha, F. Afidh, N. B. Maulydia, and G. M. Idroes, “Predicting Obesity Levels with High Accuracy : Insights from a CatBoost Machine Learning Model,” *Infolitiika J. Data Sci.*, vol. 2, no. 1, pp. 17–27, 2024, doi: 10.60084/ijds.v2i1.195.
- [10] Z. Fan, R. C. Id, Z. Hu, F. K. Id, and F. Chiong, “Body fat prediction through feature extraction based on anthropometric and laboratory measurements,” *PLoS One*, vol. 17, no. 2, pp. 1–24, 2022, doi: 10.1371/journal.pone.0263333.
- [11] Y. Görmez, F. H. Yagin, B. Yagin, and Y. Aygun, “Prediction of obesity levels based on physical activity and eating habits with a machine learning model integrated with explainable artificial intelligence,” *Front. Physiol.*, vol. 16, no. July, pp. 1–15, 2025, doi: 10.3389/fphys.2025.1549306.
- [12] D. Jawara *et al.*, “Using Machine Learning to Predict Weight Gain in Adults: an Observational Analysis from the All of Us Research Program,” *J. Surg. Res.*, vol. 306, pp. 43–53, 2025, doi: 10.1016/j.jss.2024.11.042.
- [13] B. Mazunze, L. M. Vicky, K. N. Franck, and M. M. Pierre-stéphane, “Optimization of Malaria Diagnosis by Machine Learning According to the CRISP-DM Model Applied to the University Teaching Hospital Clinics of Lubumbashi ( DRC ),” *Open Access Libr. J.*, vol. 12, no. 10, pp. 1–23, 2025, doi: 10.4236/oalib.1114143.
- [14] N. N. Muhamad and H. Thamrin, “Analyzing and Forecasting Admission Data using Time Series Model,” *JOIN J. Online Inform.*, vol. 5, no. 1, pp. 35–44, 2020, doi: 10.15575/join.v5i1.546.
- [15] National Center for Health Statistics, “NHANES 2017–2018 Questionnaires, Datasets, and Related Documentation,” Centers for Disease Control and Prevention. [Online]. Available: <https://www.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017>
- [16] Y. Hong, X. Chen, L. Wang, and F. Zhang, “Machine learning prediction of metabolic dysfunction-associated fatty liver disease risk in American adults using body composition : explainable analysis based on SHapley Additive exPlanations,” *Front. Nutr.*, vol. 12, no. June, pp. 1–13, 2025, doi: 10.3389/fnut.2025.1616229.
- [17] F. A. Kusumaningrum and D. Gunawan, “Analisis kinerja algoritma machine learning dalam deteksi kanker otak,” *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 4, pp. 4172–4180, 2025, doi: 10.29100/jupi.v10i4.7080.
- [18] A. A. Shaban and O. S. Kareem, “Comparative Evaluation of Linear Regression , Tree-Based Regression , and Neural Network Models for Structured Car Price Prediction,” *Polaris Glob. J. Sch. Res. Trends*, vol. 4, no. 1, pp. 1–21, 2025, doi: 10.58429/pgjsrt.v4n1a216.
- [19] N. T. R. Adiningrum and N. H. Harani, “Analisis Perbandingan Ensemble Machine Learning Dengan Teknik Smote Untuk Prediksi Diabetes,” *JEIS J. Elektro dan Inform. Swadharma*, vol. 5, no. 1, 2025, doi: 10.56486/jeis.vol5no1.681.
- [20] M. Ali, M. Bandpei, R. Osqueizadeh, H. Goudarzi, and N. Rahmani, “Predictors of Static Postural Loading in Primary-School-Aged Children : Comparing Elastic Net and Multiple Regression Methods,” *Children*, vol. 12, no. 6, pp. 1–15, 2025, doi: 10.3390/children12060744.
- [21] J. Candia and J. S. Tsang, “eNetXplorer : an R package for the quantitative exploration of elastic net families for generalized linear models,” *BMC Bioinformatics*, vol. 20, pp. 1–11, 2019, doi: 10.1186/s12859-019-2778-5.
- [22] S. DeepaLakshmi, “Leveraging Ensemble Methods For Robust Multiclass Classification In Obesity Prediction,” *IAPQR Trans.*, vol. 49, no. 02, pp. 155–160, 2024, [Online]. Available: [https://www.researchgate.net/publication/388842199\\_LEVERAGING\\_ENSEMBLE\\_METHOD](https://www.researchgate.net/publication/388842199_LEVERAGING_ENSEMBLE_METHOD)
- [23] W. Fhaldian and A. Fahmi, “Explainable Machine Learning for Poverty Prediction in Central Java Regencies and Cities,” *Sink. J. dan Penelit. Tek. Inform.*, vol. 9, no. 4, pp. 2080–2097, 2025, doi: 10.33395/sinkron.v9i4.15312.
- [24] D. R. Buntoro, “Pergeseran Feature Importance pada Prediksi Pasar Saham Teknologi Menggunakan Machine Learning: Studi Komparatif Pra dan Pasca Pandemi COVID-19,” *JAMI J. Ahli Muda Indones.*, vol. 6, no. December, pp. 114–119, 2025, doi: 10.46510/jami.v6i2.380.
- [25] J. Liu, D. Guan, and X. Liu, “Comparative Performance Analysis of Machine Learning Models for Compressive Strength Prediction in Concrete Mix Design,” *Math. Comput. Appl.*, vol. 30, no. 6, pp. 1–29, 2025, doi: 10.3390/mca30060128.
- [26] M. D. H. Fahri and D. Gunawan, “Analisis Sentimen Pengguna X terhadap Perempuan di Lingkungan Kerja



- Menggunakan Algoritma Machine Learning,” *J. Technol. Informatics*, vol. 7, no. 2, pp. 134–146, 2025, doi: 10.37802/joti.v7i2.1087.
- [27] M. Hamid, F. Hajjej, A. S. Alluhaidan, and N. Waleed, “Fine tuned CatBoost machine learning approach for early detection of cardiovascular disease through predictive modeling,” *Sci. Rep.*, no. 1, pp. 1–15, 2025, doi: 10.1038/s41598-025-13790-x.
- [28] W. Zou *et al.*, “An elastic net regression model for predicting the risk of ICU admission and death for hospitalized patients with COVID - 19,” *Sci. Rep.*, no. 0123456789, pp. 1–12, 2024, doi: 10.1038/s41598-024-64776-0.