IMPLEMENTASI MODEL REGRESI DALAM PREDIKSI KADAR LEMAK TUBUH

Disusun untuk Memenuhi Tugas Mata kuliah Machine learning

Dosen Pengampu:

Erika Maulidiya, S.Kom., M.Kom.

Disusun Oleh:

Kelompok 5

Dessy Nurulita	2310817220024
Firda Khoirunisa	2310817220025
Galih Aji Sabdaraya	2310817210005

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS TEKNIK

UNIVERSITAS LAMBUNG MANGKURAT

BANJARMASIN

2026

1. Latar Belakang dan Tujuan

Persentase lemak tubuh merupakan parameter krusial dalam dunia kesehatan dan kebugaran, namun pengukurannya sering kali memerlukan alat laboratorium yang mahal atau tenaga ahli. Dalam bidang Teknologi Informasi, khususnya *Data Science*, kita dapat memanfaatkan data performa untuk membangun model prediktif. Dengan memanfaatkan dataset *Body Performance Data* dari platform Kaggle, penelitian ini mengeksplorasi hubungan antara berbagai variabel kemampuan fisik dengan komposisi lemak tubuh. Melalui pendekatan regresi linier, diharapkan tercipta solusi alternatif yang lebih efisien untuk memprediksi profil kesehatan seseorang berdasarkan performa atletiknya.

2. Deskripsi Dataset

2.1. Sumber Dataset

Dataset yang digunakan dalam penelitian ini adalah *Body Performance Data* yang tersedia secara publik melalui platform Kaggle dengan sumber asli berasal dari *Korea Sports Promotion Foundation*. Dataset dapat diakses melalui tautan berikut: https://www.kaggle.com/datasets/kukuroo3/body-performance-data. Dataset ini berisi data performa fisik individu yang dikaitkan dengan usia serta berbagai indikator kebugaran tubuh.

Proyek ini berfokus pada variabel *body fat (%)* sebagai target prediksi. Persentase lemak tubuh merupakan indikator penting dalam evaluasi kesehatan dan kebugaran fisik karena berkaitan dengan komposisi tubuh serta risiko berbagai kondisi metabolik. Nilai ini bersifat kontinu, sehingga pendekatan regresi linier dapat diterapkan untuk memodelkan hubungan antara variabel input dan target.

Fitur-fitur dalam dataset mencakup informasi usia, jenis kelamin, tinggi badan, berat badan, tekanan darah (sistolik dan diastolik), kekuatan genggaman tangan (*grip force*), fleksibilitas (*sit and bend forward*), jumlah *sit-up*, serta jarak *broad jump*. Kombinasi fitur tersebut memungkinkan analisis hubungan antara kondisi fisik dan komposisi tubuh secara kuantitatif.

2.2. Karakteristik Dataset

Dataset memiliki total 13.393 baris dan 12 kolom. Seluruh baris merepresentasikan individu dengan rentang usia 20–64 tahun. Berikut adalah daftar variabel yang tersedia:

Tabel 1. Karakteristik Dataset

Nama Fitur	Deskripsi	Tipe Data
age	Usia individu	Numerik
gender	Jenis kelamin (F/M)	Kategorikal
height_cm	Tinggi badan (cm)	Numerik
weight_kg	Berat badan (kg)	Numerik
body fat_%	Persentase lemak tubuh	**Numerik (Target)**
diastolic	Tekanan darah diastolik	Numerik
systolic	Tekanan darah sistolik	Numerik
gripForce	Kekuatan genggaman tangan	Numerik
sit and bend forward_cm	Jarak fleksibilitas tubuh (cm)	Numerik
sit-ups counts	Jumlah sit-up	Numerik
broad jump_cm	Jarak lompat jauh (cm)	Numerik
class	Kelas performa (A–D)	Kategorikal

3. Data Preprocessing Pipeline

3.1 Data Understanding dan Eksplorasi Awal

Tahap awal penelitian dimulai dengan memuat dataset *Body Performance* dan melakukan inspeksi awal untuk memahami struktur data. Proses ini meliputi pemeriksaan jumlah observasi dan variabel, identifikasi tipe data, serta pengecekan keberadaan nilai hilang. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung missing value sehingga tidak diperlukan proses imputasi.

Selanjutnya dilakukan analisis eksploratif awal (Exploratory Data Analysis) untuk memahami distribusi variabel target (*body fat_%*), mendeteksi potensi outlier melalui boxplot, serta menganalisis hubungan antar fitur menggunakan matriks korelasi. Tahapan ini bertujuan untuk memperoleh gambaran karakteristik data sebelum dilakukan pemodelan.

3.2 Transformasi dan Persiapan Fitur

Setelah memahami struktur data, dilakukan proses persiapan fitur yang meliputi beberapa tahapan berikut:

1. Pemisahan variable:

   Dataset dipisahkan menjadi variabel independen (X) dan variabel dependen (y), dengan *body fat_%* sebagai target prediksi.

2. Encoding variabel kategorikal:

   Variabel kategorikal dikonversi menggunakan metode One-Hot Encoding. Parameter drop_first=True digunakan untuk menghindari dummy variable trap dan mengurangi risiko multikolinearitas.

3. Konversi tipe data:

   Seluruh fitur dikonversi menjadi tipe numerik (*float*) untuk memastikan kompatibilitas dengan algoritma regresi.

4. Analisis multikolinearitas:

   Multikolinearitas antar fitur independen dianalisis menggunakan Variance Inflation Factor (VIF). Langkah ini penting untuk menjaga stabilitas koefisien pada model regresi linear.

Tahapan ini memastikan bahwa seluruh fitur berada dalam format numerik yang sesuai dan siap digunakan dalam proses pelatihan model.

3.3 Penanganan Outlier dan Feature Scaling

Penanganan outlier untuk meningkatkan kualitas data pelatihan menggunakan metode *Interquartile Range* (IQR).

- $Q1 - 1.5 \times IQR$
- $Q3 + 1.5 \times IQR$

Data yang berada di luar batas dianggap sebagai *outlier* dan dihapus. Proses ini hanya diterapkan pada data pelatihan dalam setiap *fold cross-validation* guna mencegah data leakage. Selanjutnya, dilakukan standarisasi fitur menggunakan *StandardScaler*, yang mengubah setiap variabel sehingga memiliki rata-rata 0 dan standar deviasi 1. Standarisasi dilakukan di dalam *pipeline* agar parameter *scaling* dihitung hanya dari data pelatihan dan kemudian diterapkan pada data pengujian.

3.4 Validasi Menggunakan K-Fold Cross Validation

Metode *5-Fold Cross Validation* digunakan untuk memperoleh estimasi performa model yang lebih stabil. Data dibagi menjadi lima subset, di mana setiap subset secara bergantian digunakan sebagai data pengujian, sementara sisanya digunakan sebagai data pelatihan. Pendekatan ini membantu mengurangi risiko overfitting serta memberikan evaluasi model yang lebih representatif dibandingkan pembagian data tunggal.

4. Metodologi dan Model

Dalam penelitian ini, dilakukan perbandingan antara model regresi linier konvensional (*baseline*) dengan dua model pengembangan regresi linier modern, yaitu Ridge Regression dan Lasso Regression. Ketiganya termasuk dalam keluarga model linier, namun memiliki karakteristik dan pendekatan yang berbeda dalam menyelesaikan permasalahan regresi.

4.1 Linear Regression

Linear Regression atau OLS (*Ordinary Least Squares*) merupakan fondasi dari semua model regresi linier. Model ini bekerja dengan meminimalkan jumlah kuadrat residual (RSS) antara nilai aktual dan nilai prediksi.

4.2 Ridge Regression

Ridge Regression merupakan pengembangan dari linear regression yang diperkenalkan untuk mengatasi kelemahan OLS, terutama dalam menangani multikolinearitas. Model ini menambahkan penalti L2 ke dalam fungsi loss.

4.3 Lasso Regression

Lasso Regression (*Least Absolute Shrinkage and Selection Operator*) merupakan inovasi lebih lanjut dari ridge regression dengan menggunakan penalti L1. Keunikan lasso adalah kemampuannya melakukan seleksi fitur secara otomatis.

4.4 Perbandingan Karakteristik Model

Secara keseluruhan, ketiga model ini menawarkan pendekatan yang berbeda dalam menangani data *Body Performance*. Linear Regression berperan sebagai model dasar yang paling sederhana namun sangat rentan terhadap *outlier* dan multikolinearitas dibandingkan dua model lainnya. Sebagai pengembangan, Ridge Regression memberikan stabilitas lebih tinggi dengan menangani hubungan antar fitur yang terlalu kuat tanpa menghapusnya, menjadikannya lebih tangguh daripada model linier biasa namun tetap mempertahankan kompleksitas fitur. Di sisi lain, Lasso Regression unggul dalam menyederhanakan model melalui seleksi fitur otomatis yang mampu mengnolkan koefisien fitur tidak relevan, yakni sebuah kemampuan yang tidak dimiliki baik oleh Linear Regression maupun Ridge. Pemilihan di antara ketiganya sangat bergantung pada apakah tujuan penelitian lebih mengutamakan stabilitas seluruh fitur (Ridge) atau kesederhanaan model (Lasso).

5. Experiment Log

Tabel 2. Experiment Log

Model	Fitur	Parameter	MAE	RMSE	R^2	MAPE (%)	Catatan
Linear Regression	Semua fitur + OHE	Default	2.9154	3.7566	0.7319	14.51	Baseline model tanpa regularisasi
Ridge	Semua fitur + OHE	alpha = 0.01	2.9154	3.7566	0.7319	14.51	Tidak ada perubahan signifikan
Ridge	Semua fitur + OHE	alpha = 0.1	2.9154	3.7566	0.7319	14.51	Regularisasi sangat kecil
Ridge	Semua fitur + OHE	alpha = 1.0	2.9155	3.7566	0.7319	14.51	Masih stabil
Ridge	Semua fitur + OHE	alpha = 10	2.9163	3.7569	0.7319	14.51	Performa mulai sedikit turun
Ridge	Semua fitur + OHE	alpha = 100	2.9308	3.7686	0.7302	14.57	Regularisasi mulai berdampak
Ridge	Semua fitur + OHE	alpha = 500	3.0547	3.9072	0.7098	15.10	Regularisasi terlalu kuat
Lasso	Semua fitur + OHE	alpha = 0.01	2.9174	3.7572	0.7318	14.51	Hampir sama dengan baseline
Lasso	Semua fitur + OHE	alpha = 0.1	2.9708	3.8121	0.7239	14.73	Mulai kehilangan informasi

Lasso	Semua fitur + OHE	alpha = 1.0	4.1260	5.2321	0.4765	20.24	Banyak koefisien menjadi nol
Lasso	Semua fitur + OHE	alpha = 10	5.8637	7.3504	-0.0264	29.06	Model gagal belajar (underfitting)
Lasso	Semua fitur + OHE	alpha = 100	5.8637	7.3504	-0.0264	29.06	Koefisien hampir semua nol
Lasso	Semua fitur + OHE	alpha = 500	5.8637	7.3504	-0.0264	29.06	Underfitting ekstrem

6. Hasil Evaluasi

Bagian ini menyajikan performa dari tiga model yang diuji menggunakan metode 5-Fold Cross Validation. Metrik yang digunakan adalah MAE, RMSE, R^2, dan MAPE untuk memberikan gambaran menyeluruh mengenai akurasi prediksi kadar lemak tubuh. Berdasarkan hasil eksperimen, berikut adalah ringkasan skor performa untuk masing-masing model:

Metrik Evaluasi	**Linear Regression (Baseline)**	**Ridge Regression**	**Lasso Regression**
MAE	2.9154	2.9155	2.9708
RMSE	3.7565	3.7565	3.8120
R^2 Score	0.7319	0.7319	0.7239
MAPE (%)	14.51%	14.51%	14.73%

1) Akurasi Model

Ketiga model menunjukkan performa yang sangat kompetitif dengan nilai R^2 di kisaran 0.72 - 0.73. Ini berarti model mampu menjelaskan sekitar 73% variasi data lemak tubuh berdasarkan fitur fisik yang diberikan. Nilai MAE sebesar 2.91 menunjukkan bahwa rata-rata kesalahan prediksi hanya sekitar 2.91% dari nilai asli.

2) Perbandingan Performa

- Linear Regression dan Ridge Regression: Kedua model ini memiliki hasil yang hampir identik. Hal ini terjadi karena pada dataset ini, hubungan antar variabel

cenderung linier dan bobot penalti (alpha) pada Ridge tidak memberikan perubahan drastis pada koefisien. Namun, Ridge tetap unggul dalam stabilitas numerik untuk mencegah multikolinearitas.

- Lasso Regression: Menunjukkan penurunan performa yang sangat kecil (RMSE naik menjadi 3.812). Hal ini mengindikasikan bahwa fitur-fitur fisik dalam dataset ini hampir semuanya memiliki kontribusi penting; Lasso yang cenderung melakukan seleksi fitur (menghilangkan fitur) sedikit kurang optimal dibandingkan Ridge.

3) Stabilitas Model

Sesuai dengan metodologi 5-Fold Cross Validation yang dijelaskan pada Bab 3, hasil evaluasi ini merupakan nilai rata-rata dari 5 kali pengujian berbeda. Konsistensi nilai di setiap *fold* membuktikan bahwa model memiliki kemampuan generalisasi yang tinggi dan tidak mengalami gejala *overfitting*.

7. Log Prompt dan Modifikasi Kode

Tabel 3. Log Prompt

Tujuan Prompt	Prompt (Asli)	Ringkasan Output AI	Perubahan dan Alasan
Menentukan batas outlier pada stress test	"sekarang outlier stress gimana?"	Menggunakan batas berbasis mean dan standar deviasi (`mean + 5 * std`).	Diubah menjadi `mean + 2 * std` karena batas 5 std menyebabkan penurunan performa model yang sangat signifikan saat pengujian menjadi R^2 negatif saat 5%.
Menentukan tingkat noise untuk uji robustness	"untuk noise stress gimana"	Menyarankan noise sebesar 0.05.	Diperluas menjadi `[0, 0.05, 0.10, 0.20, 0.30]` agar pengujian lebih komprehensif.

Menentukan model yang diuji pada stress test	"untuk stress test apakah perlu semua model yang diperlukan atau model yang terbaik?"	Menguji seluruh model dan tuning-nya.	Dipilih model terbaik saja dari tiga model, yaitu regresi linear , regresi Ridge Alpha (0.01), dan Lasso Regression (0.01) agar analisis fokus pada performa optimal dan tidak redundan.

8. Stress Test

Melakukan dua skenario uji stres untuk menguji ketangguhan (*robustness*) model yang telah dibangun:

1) Noise Stress Test

Skenario ini mensimulasikan pergeseran distribusi pada fitur input (misalnya kesalahan alat timbang) tanpa mengubah hubungan dasar antara fitur dan target.

a. Ketahanan Rendah (0.00 - 0.10): Ketiga model sangat stabil dengan RMSE terjaga di kisaran 3.78 - 3.79.
b. Degradasi Tinggi (0.40): Terjadi lonjakan error yang signifikan di mana RMSE mencapai 4.25 - 4.27.

Hal tersebut menunjukkan bahwa model cukup *robust* terhadap gangguan kecil, namun akurasi akan menurun drastis jika terjadi penyimpangan alat ukur yang ekstrem.

2) Outlier Stress Test

Skenario ini menyisipkan data menyimpang yang mengubah statistik data secara keseluruhan, menciptakan fenomena pergeseran distribusi yang ekstrem.

a. Titik Fatal: Pada tingkat kontaminasi *outlier* sebesar 0%, nilai R^2 anjlok hingga angka negatif (≈ -0.07).
b. Analisis: Nilai negatif menunjukkan model kehilangan kemampuan generalisasi dan berperforma lebih buruk daripada tebakan rata-rata.

Hal tersebut menunjukkan bahwa model regresi linier (termasuk Ridge/Lasso) sangat rentan terhadap *outlier* yang masif, sehingga tahap pembersihan data (*preprocessing*) di awal sangatlah krusial.

3) *Covariate Shift*

Skenario ini mensimulasikan perubahan distribusi pada fitur input (X_test) secara konsisten, misalnya karena perubahan demografi populasi atau kalibrasi alat yang bergeser secara sistematis. Berbeda dengan *noise* yang acak, *covariate shift* menggeser seluruh data uji ke arah yang sama.

a. Pergeseran kecil (faktor 0,5): Model masih cukup robust dengan penurunan R^2 minimal (dari 0,73 menjadi 0,59).
b. Pergeseran sedang (faktor 1,0): Kinerja model mulai tidak dapat diandalkan ($R^2 = 0,18$).
c. Pergeseran besar (faktor ≥ 1,5): Model gagal total dengan nilai R^2 negatif.

Hal ini menunjukkan bahwa model regresi linier sangat sensitif terhadap perubahan distribusi fitur. Jika karakteristik populasi atau kondisi pengukuran berubah secara signifikan, model harus dilatih ulang dengan data baru yang relevan.

9. Kesimpulan

Berdasarkan keseluruhan rangkaian eksperimen dan evaluasi, dapat disimpulkan bahwa model regresi linier mampu memprediksi persentase lemak tubuh dengan performa yang cukup baik, ditunjukkan oleh nilai R^2 sekitar 0,73 serta MAE rata-rata ±2,91 yang menandakan kesalahan prediksi relatif kecil. Linear Regression sebagai baseline dan Ridge Regression menghasilkan performa yang hampir identik, menunjukkan bahwa hubungan antar fitur dalam dataset cenderung linier dan tidak mengalami multikolinearitas berat, sementara Lasso Regression sedikit menurun performanya karena kecenderungan menghilangkan fitur yang ternyata masih relevan. Hasil 5-Fold Cross Validation memperlihatkan model cukup stabil dan memiliki generalisasi yang baik pada data dengan distribusi serupa. Namun, melalui stress test (noise, outlier, dan covariate shift), terlihat bahwa model regresi linier cukup robust terhadap gangguan kecil tetapi sangat sensitif terhadap outlier ekstrem dan pergeseran distribusi data, sehingga tahap preprocessing dan pemantauan distribusi data menjadi faktor krusial dalam implementasi nyata

10. Link Youtube

https://youtu.be/4QE2xkgeawM

11. Link GitHub

https://github.com/Kylorts/regression-linear-body-perfomance