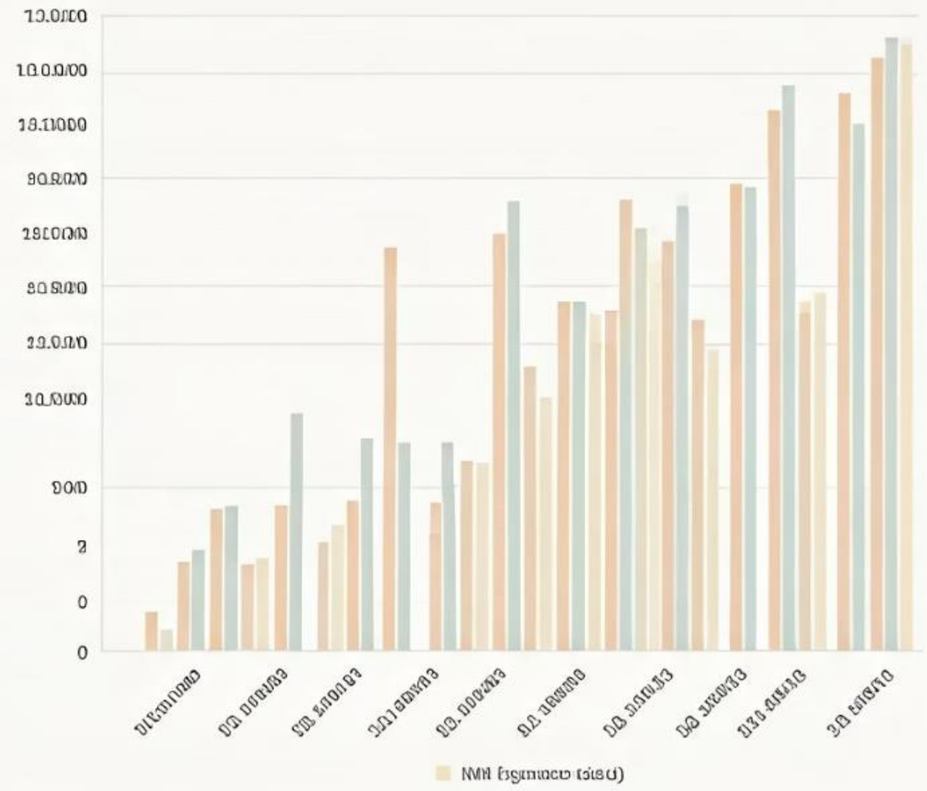


Analisis Prediksi BMI dengan KNN Regressor

raw.githubusercontent.com

Selamat datang di presentasi analisis prediksi BMI menggunakan algoritma K-Nearest Neighbors (KNN) Regressor. Dalam presentasi ini, kita akan mengeksplorasi dataset anggota gym, menganalisis korelasi antar variabel, dan membangun model prediksi BMI berdasarkan berbagai fitur kesehatan dan kebugaran.

Kami akan menunjukkan proses lengkap dari persiapan data, eksplorasi, pemodelan, hingga evaluasi kinerja model. Mari kita mulai dengan memahami data yang akan kita gunakan.

KELOMPOK

Mochmmad Qaysya	M INDRA PRATAMA
	9882405222111012

Pengenalan Dataset Anggota Gym

Struktur Dataset

Dataset yang digunakan berisi informasi pelacakan latihan anggota gym dengan 973 baris dan 15 kolom. Data mencakup berbagai atribut seperti usia, jenis kelamin, berat badan, tinggi badan, detak jantung, dan informasi latihan.

Tujuan utama analisis ini adalah untuk memprediksi BMI (Body Mass Index) berdasarkan fitur-fitur tersebut menggunakan algoritma KNN Regressor.

Variabel Utama

- Demografis: Usia, Jenis Kelamin
- Fisik: Berat, Tinggi, Persentase Lemak
- Kardio: Max BPM, Avg BPM, Resting BPM
- Latihan: Durasi Sesi, Kalori Terbakar, Jenis Latihan
- Lainnya: Asupan Air, Frekuensi Latihan, Level Pengalaman

Eksplorasi Data Awal

Age	Gender	Weight (kg)	Height (m)	Max_BPM	Avg_BPM	Resting_BPM
56	Male	88.3	1.71	180	157	60
46	Female	74.9	1.53	179	151	66
32	Female	68.1	1.66	167	122	54
25	Male	53.2	1.70	190	164	56
38	Male	46.1	1.79	188	158	68

Tabel di atas menunjukkan lima baris pertama dari dataset. Kita dapat melihat variasi dalam usia, jenis kelamin, berat badan, dan parameter kesehatan lainnya. Dataset ini memberikan gambaran komprehensif tentang profil anggota gym dan aktivitas latihan mereka.

Transformasi Data

Encoding Variabel Kategorikal

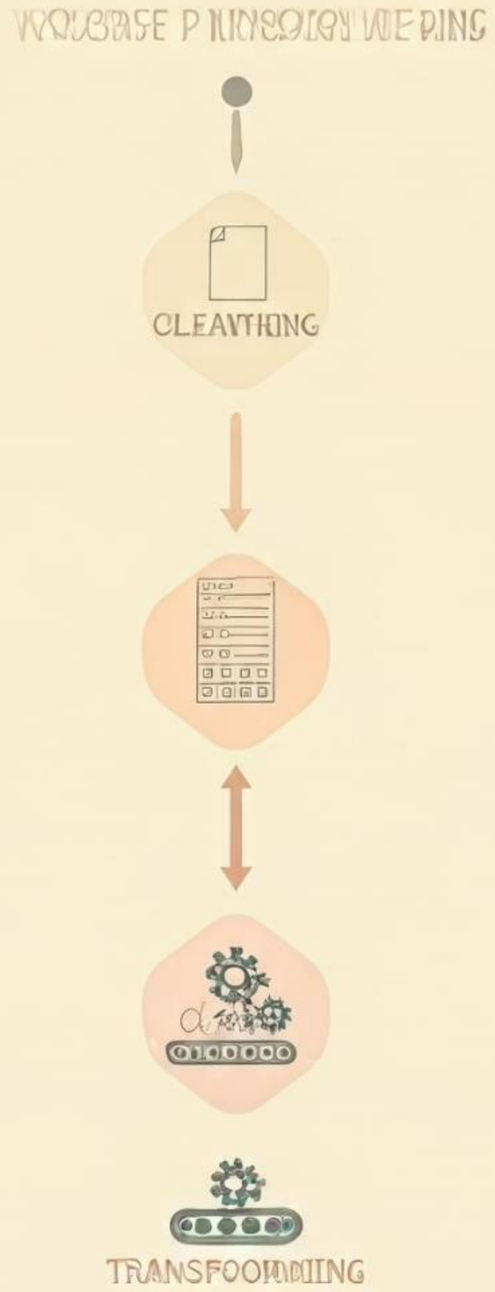
Variabel kategorikal seperti Gender dan Workout_Type diubah menjadi bentuk numerik menggunakan LabelEncoder. Ini diperlukan karena algoritma KNN bekerja dengan data numerik.

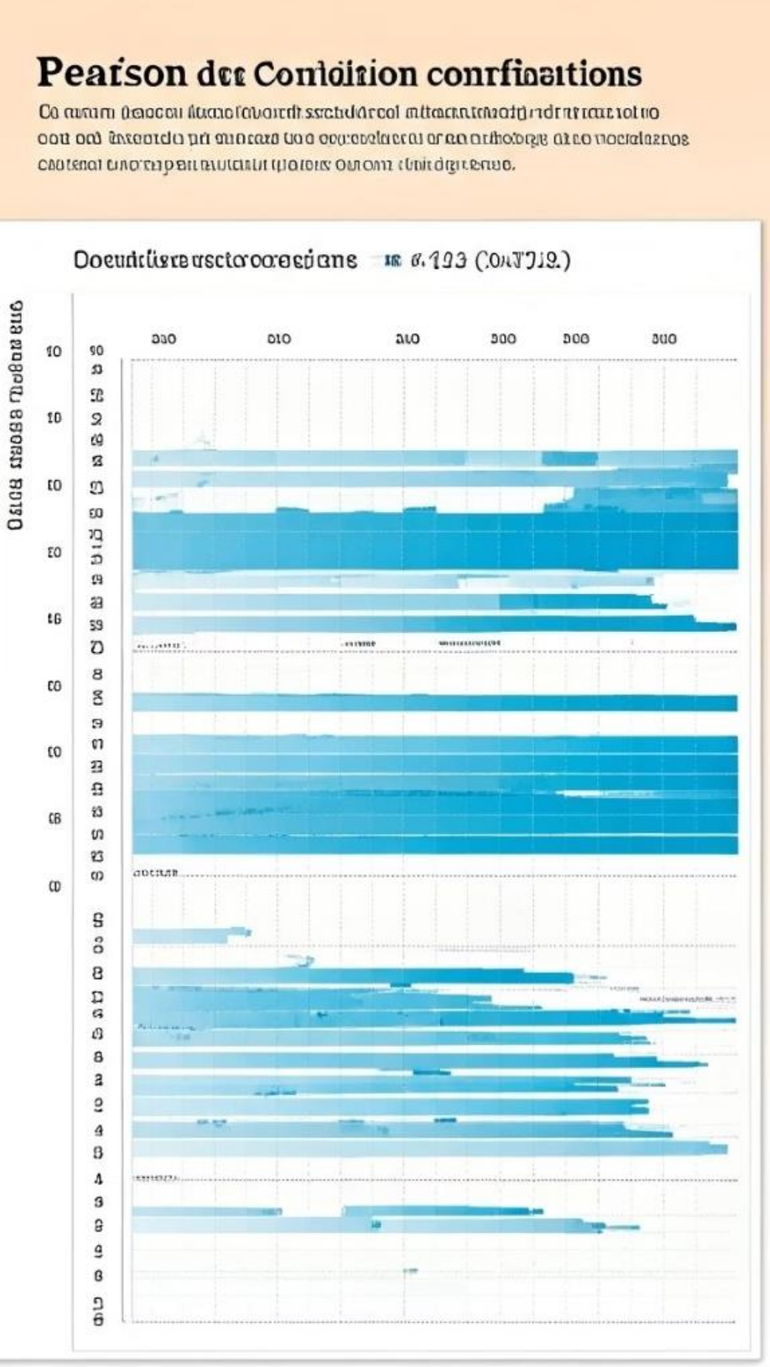
Penghapusan Kolom Asli

Setelah encoding, kolom asli yang berisi nilai non-numerik dihapus dari dataset untuk menghindari redundansi dan memastikan semua data dalam format yang sesuai untuk pemodelan.

Pemilihan Fitur

Beberapa kolom seperti Resting_BPM dan Workout_Type_encoded kemudian dihapus berdasarkan analisis korelasi untuk mengurangi dimensi data dan meningkatkan performa model.





Analisis Korelasi Pearson

Analisis korelasi Pearson membantu kita memahami hubungan linear antara variabel dalam dataset. Dari matriks korelasi di atas, kita dapat mengidentifikasi beberapa pola penting:

Korelasi Kuat dengan BMI

Berat badan memiliki korelasi positif yang sangat kuat dengan BMI (0.85), yang masuk akal karena BMI dihitung berdasarkan berat dan tinggi badan.

Durasi dan Kalori

Durasi sesi latihan berkorelasi kuat dengan kalori yang terbakar (0.91), menunjukkan bahwa latihan yang lebih lama umumnya membakar lebih banyak kalori.

Asupan Air

Asupan air berkorelasi negatif dengan persentase lemak (-0.59), yang mungkin menunjukkan bahwa mereka yang minum lebih banyak air memiliki persentase lemak tubuh yang lebih rendah.

Analisis Korelasi Kendall

Korelasi Kendall adalah metode non-parametrik yang mengukur kekuatan hubungan ordinal antara dua variabel. Dibandingkan dengan Pearson, korelasi Kendall lebih tahan terhadap outlier dan tidak mengasumsikan hubungan linear.

Perbedaan dengan Pearson

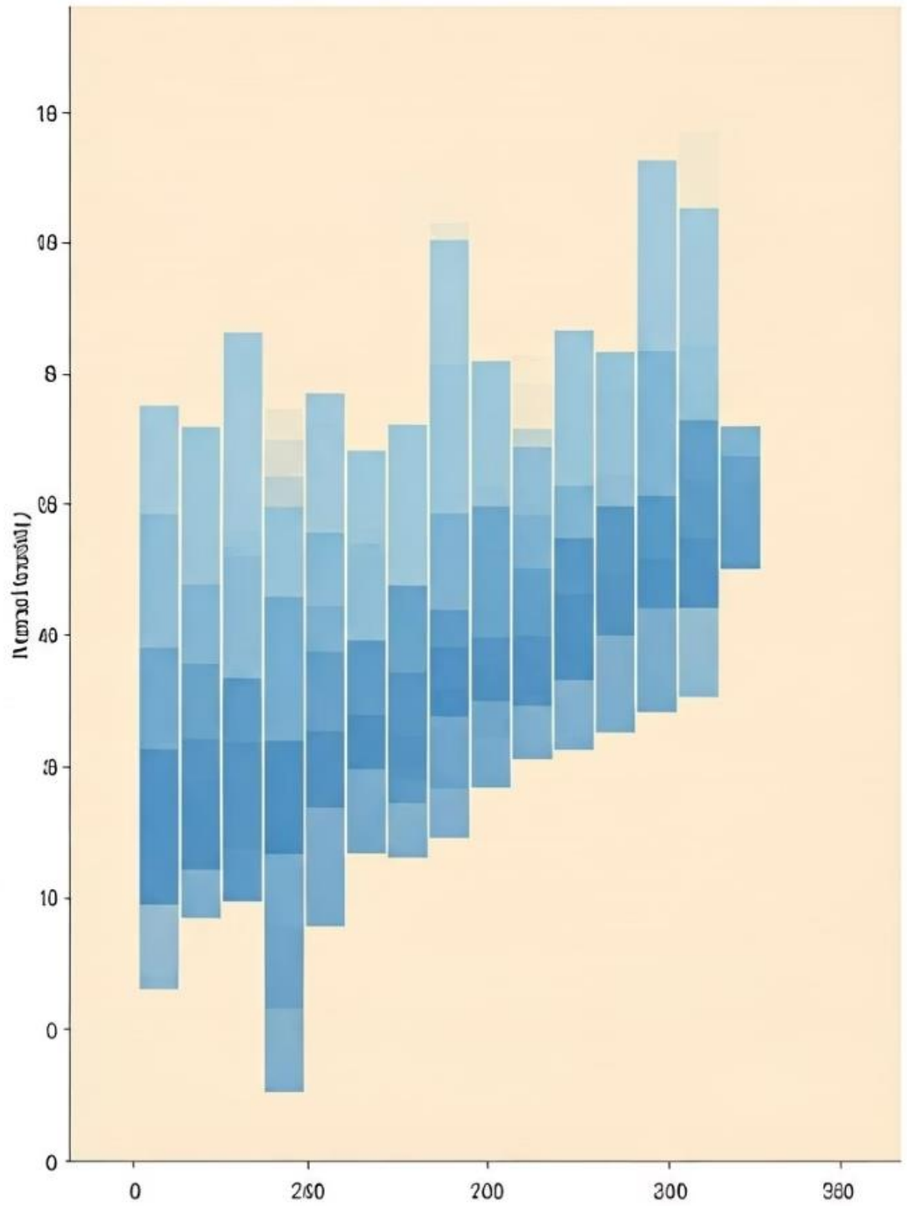
Nilai korelasi Kendall umumnya lebih rendah daripada Pearson, tetapi pola hubungan antar variabel tetap konsisten. Ini menunjukkan bahwa hubungan yang terdeteksi bukan hanya linear tetapi juga ordinal.

Hubungan Signifikan

Berat badan dan BMI tetap menunjukkan korelasi yang kuat (0.65), meskipun nilainya lebih rendah dari korelasi Pearson. Ini mengkonfirmasi hubungan yang kuat antara kedua variabel tersebut.

Implikasi untuk Pemodelan

Konsistensi pola korelasi antara metode Pearson dan Kendall memberikan keyakinan lebih dalam pemilihan fitur untuk model prediksi BMI kita.



Analisis Korelasi Spearman

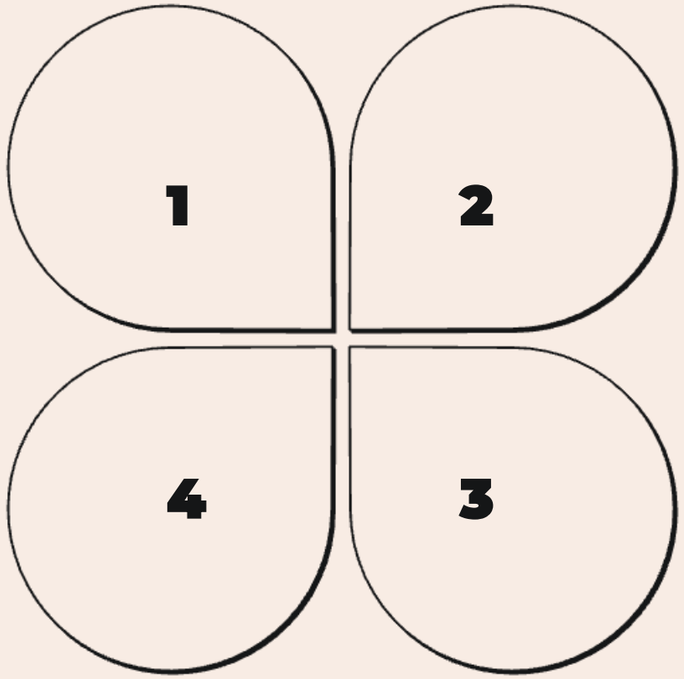
Korelasi Spearman, seperti Kendall, adalah metode non-parametrik yang mengukur kekuatan hubungan monoton antara dua variabel. Ini berguna untuk mengidentifikasi hubungan yang mungkin tidak linear tetapi masih monoton.

Berat dan BMI

Korelasi Spearman antara berat badan dan BMI adalah 0.84, sangat dekat dengan nilai Pearson, menunjukkan hubungan yang konsisten terlepas dari metode yang digunakan.

Gender dan Asupan Air

Terdapat korelasi yang cukup kuat antara gender dan asupan air (0.66), yang mungkin mencerminkan perbedaan kebiasaan hidrasi antara pria dan wanita.



Durasi dan Kalori

Hubungan antara durasi sesi dan kalori yang terbakar tetap kuat (0.90), konsisten dengan hasil Pearson.

Persentase Lemak

Persentase lemak menunjukkan korelasi negatif yang konsisten dengan asupan air dan level pengalaman, menunjukkan pola yang stabil.

Statistik Deskriptif Dataset

Variabel	Mean	Std	Min	25%	50%	75%	Max
Age	38.68	12.18	18.00	28.00	40.00	49.00	59.00
Weight (kg)	73.85	21.21	40.00	58.10	70.00	86.00	129.90
Height (m)	1.72	0.13	1.50	1.62	1.71	1.80	2.00
BMI	24.91	6.66	12.32	20.11	24.16	28.56	49.84

Tabel di atas menunjukkan statistik deskriptif untuk beberapa variabel kunci dalam dataset. Kita dapat melihat bahwa usia rata-rata anggota gym adalah sekitar 39 tahun, dengan berat rata-rata 74 kg dan tinggi rata-rata 1,72 m.

BMI rata-rata adalah 24,91, yang berada dalam kisaran normal (18,5-24,9), meskipun terdapat variasi yang signifikan dengan nilai minimum 12,32 (kekurangan berat badan) hingga maksimum 49,84 (obesitas kelas III).

Persiapan Model KNN Regressor

Pemisahan Data

Data dibagi menjadi set pelatihan (80%) dan set pengujian (20%) menggunakan fungsi train_test_split dengan random_state=42 untuk memastikan hasil yang dapat direproduksi.

Inisialisasi Model

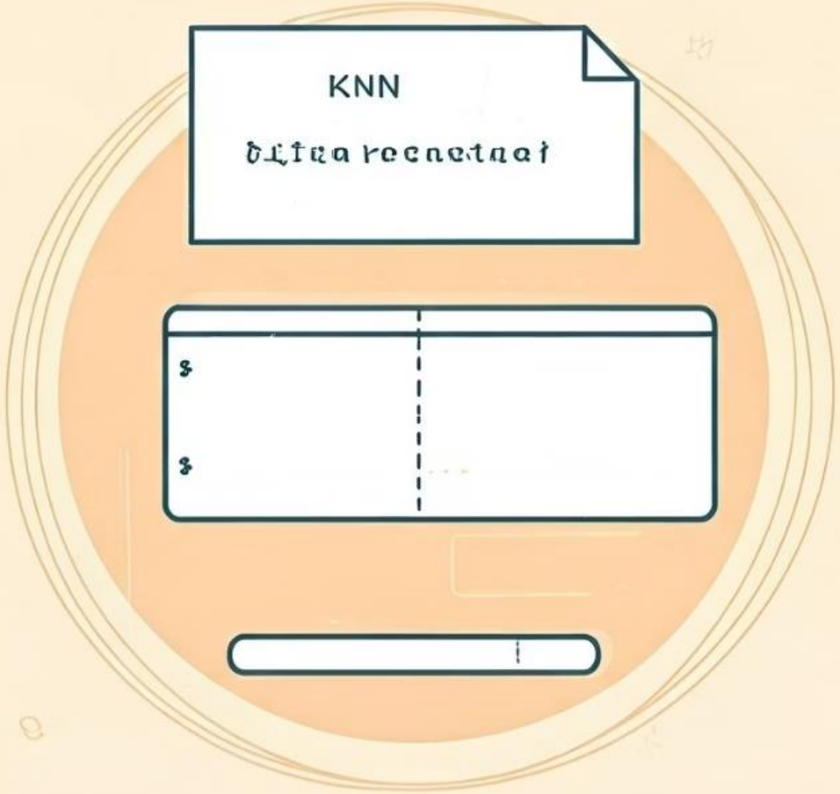
Model KNN Regressor diinisialisasi dengan parameter n_neighbors=3, yang berarti prediksi akan didasarkan pada 3 tetangga terdekat dari setiap titik data.

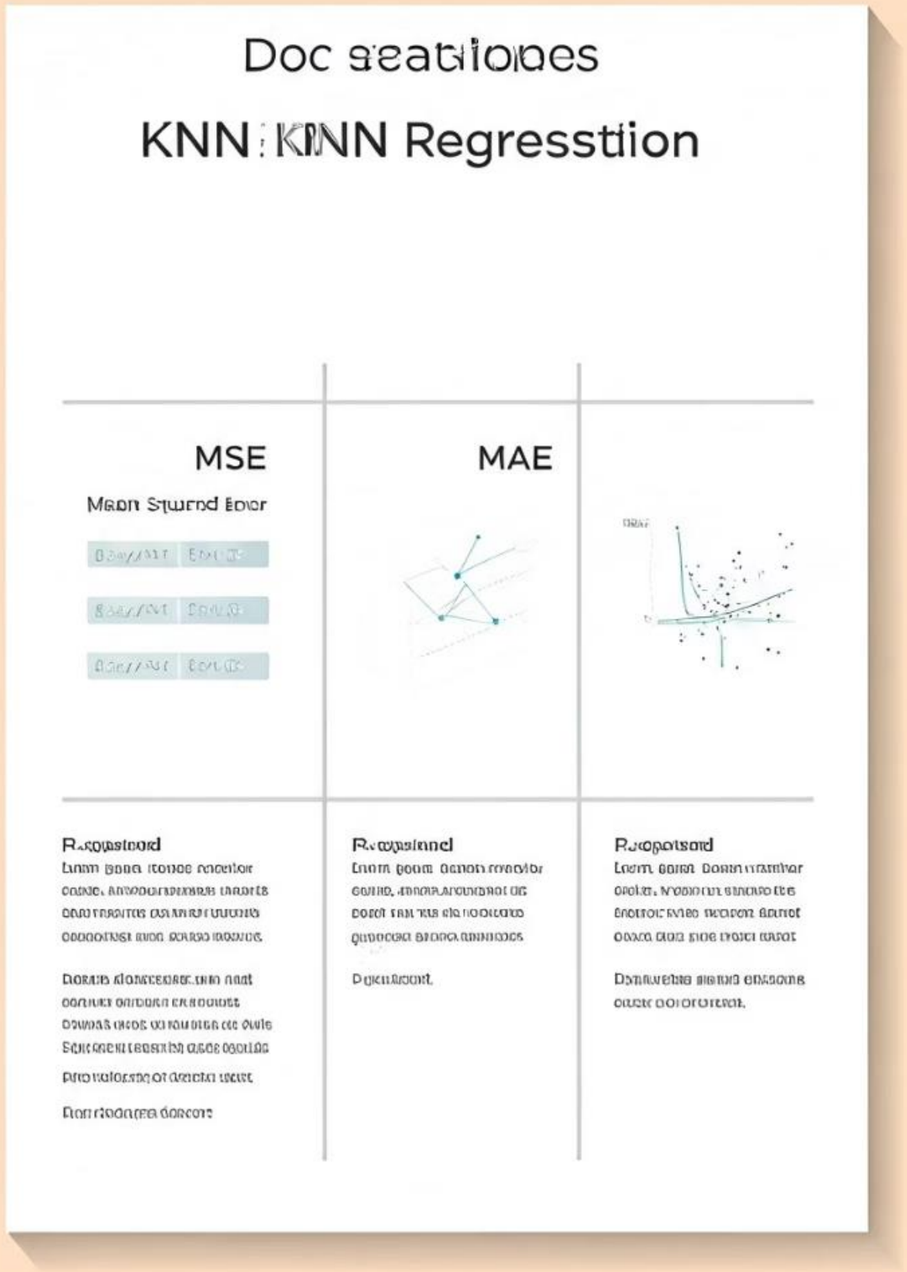
Pelatihan Model

Model dilatih menggunakan data pelatihan (X_train dan y_train) untuk mempelajari pola dan hubungan antara fitur dan target (BMI).

Pengujian Model

Setelah pelatihan, model digunakan untuk memprediksi BMI pada set pengujian (X_test), dan hasilnya dibandingkan dengan nilai BMI aktual (y_test).





Evaluasi Model Awal (K=3)

0.138

MSE

Mean Squared Error menunjukkan rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual.

0.250

MAE

Mean Absolute Error menunjukkan rata-rata nilai absolut dari selisih antara nilai prediksi dan nilai aktual.

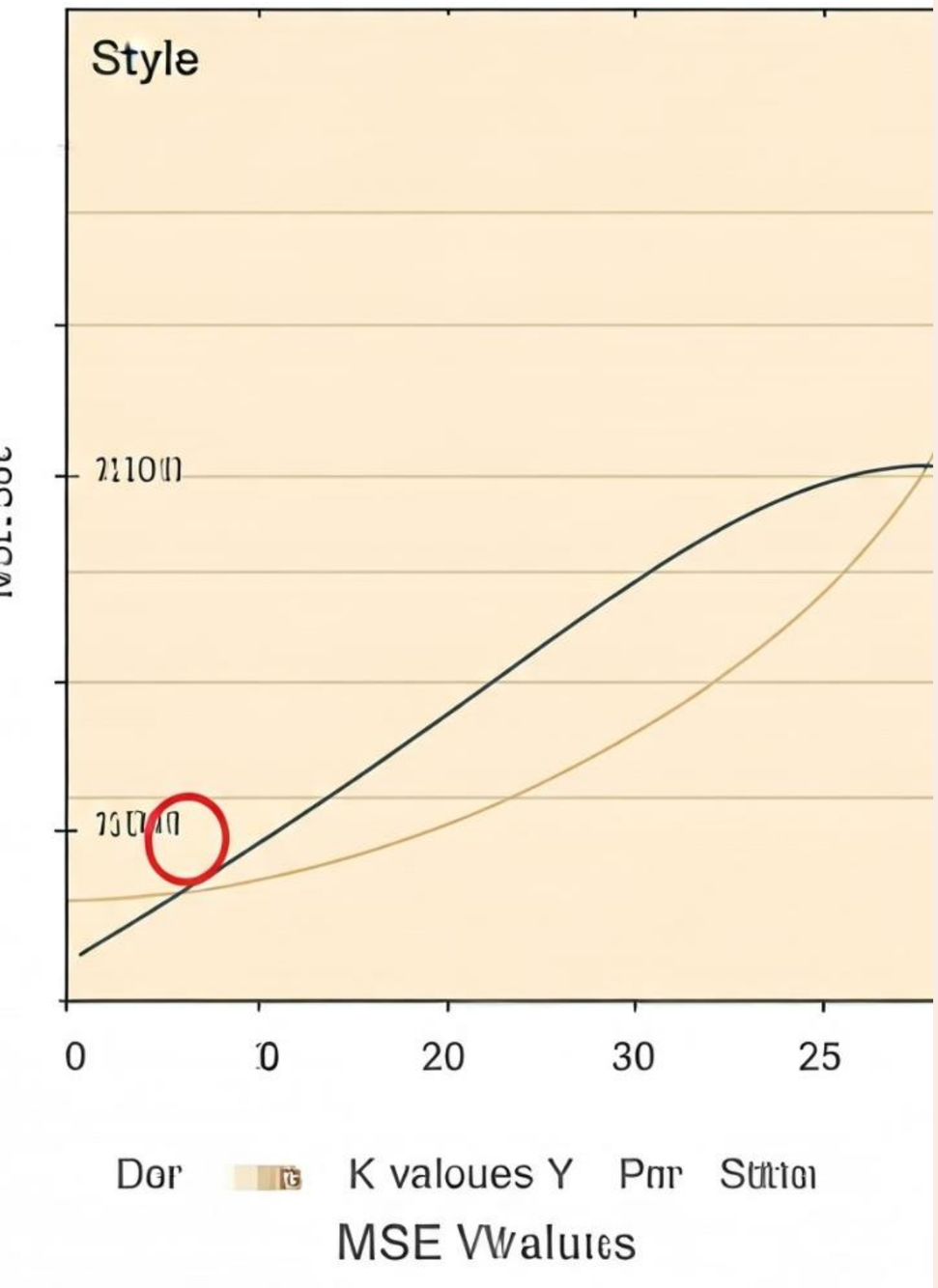
0.448

R^2 Score

Koefisien determinasi yang mengukur proporsi variasi dalam variabel dependen yang dapat diprediksi dari variabel independen.

Hasil evaluasi awal dengan K=3 menunjukkan performa yang cukup baik dengan MSE rendah (0,138) dan R^2 score yang moderat (0,448). Ini menunjukkan bahwa model dapat menjelaskan sekitar 44,8% variasi dalam BMI berdasarkan fitur yang digunakan.

Namun, masih ada ruang untuk peningkatan. Langkah selanjutnya adalah mencari nilai K optimal untuk meningkatkan performa model.

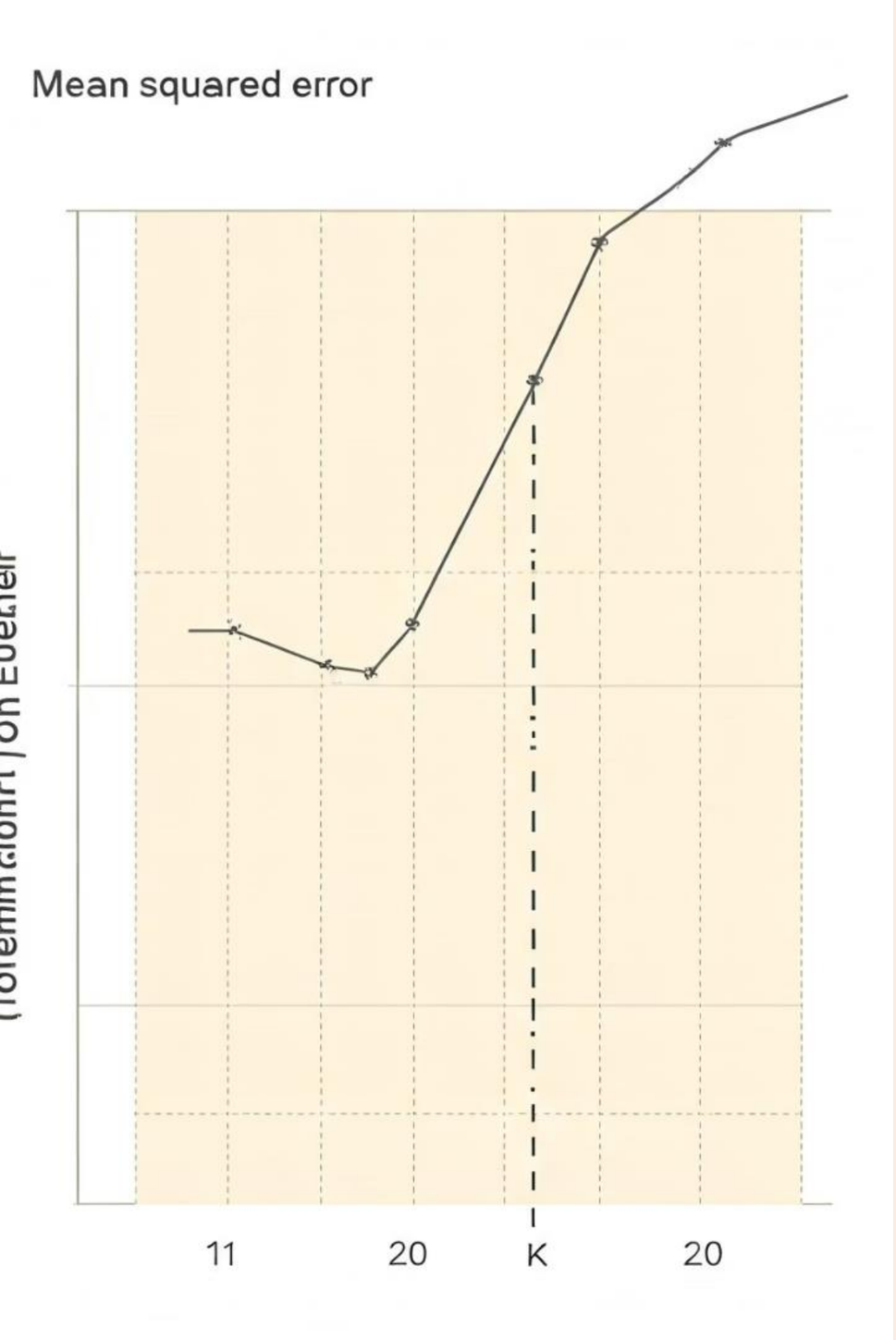


Pencarian Nilai K Optimal

Untuk menemukan nilai K optimal, kita melakukan validasi silang 5-fold untuk nilai K dari 1 hingga 51 dan menghitung Mean Squared Error (MSE) untuk setiap nilai K.

Dari grafik di atas, kita dapat melihat bahwa MSE menurun tajam dari K=1 hingga K=5, kemudian mulai meningkat secara bertahap. Nilai K=5 memberikan MSE terendah sebesar 0,1605, menunjukkan bahwa ini adalah nilai K optimal untuk model kita.

Penggunaan validasi silang memastikan bahwa hasil ini robust dan tidak bergantung pada pembagian data tertentu, memberikan keyakinan lebih dalam pemilihan parameter model.



Tabel MSE untuk Berbagai Nilai K

Nilai K	MSE Rata-rata
1	0.246683
2	0.190919
3	0.172551
4	0.166755
5	0.160501
6	0.163250
7	0.167936

Tabel di atas menunjukkan MSE rata-rata untuk berbagai nilai K dari 1 hingga 7. Kita dapat melihat bahwa MSE menurun secara konsisten dari K=1 hingga K=5, kemudian mulai meningkat lagi pada K=6 dan K=7.

Nilai K yang terlalu kecil (seperti K=1) cenderung menghasilkan model yang overfitting, sementara nilai K yang terlalu besar dapat menghasilkan model yang terlalu umum. K=5 memberikan keseimbangan optimal untuk dataset ini.

Validasi Silang dengan K=22

MSE

Mean Squared Error dari validasi silang 5-fold dengan K=22 adalah 1,3243, yang jauh lebih tinggi dibandingkan dengan K=5 (0,1605).

MAE

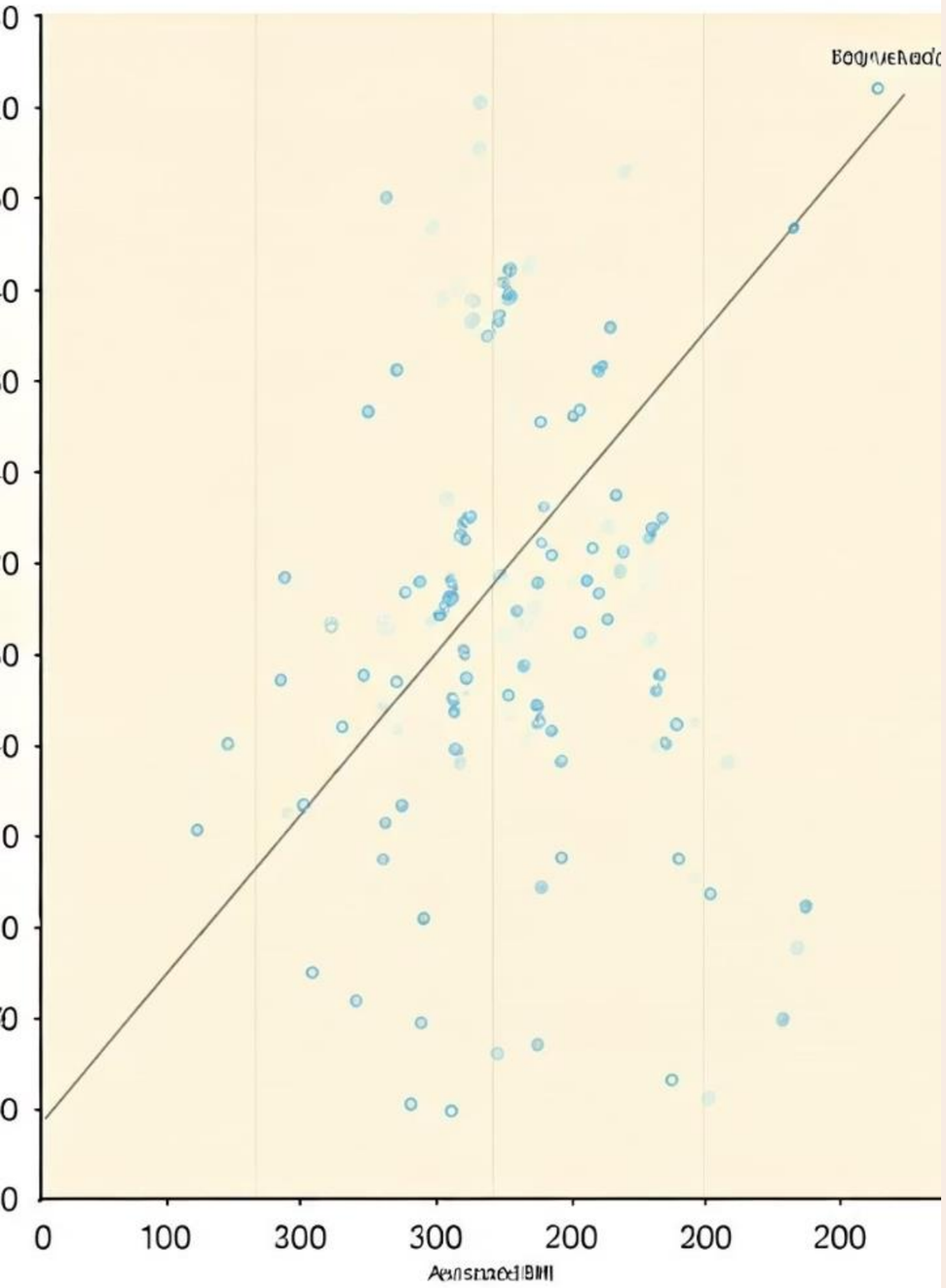
Mean Absolute Error adalah 1,0134, menunjukkan rata-rata kesalahan absolut yang cukup besar dalam prediksi BMI.

R^2 Score

R^2 score adalah -0,0490, nilai negatif menunjukkan bahwa model dengan K=22 berkinerja lebih buruk daripada hanya menggunakan rata-rata BMI sebagai prediksi.

Hasil validasi silang dengan K=22 menunjukkan performa yang jauh lebih buruk dibandingkan dengan K=5. Ini mengkonfirmasi bahwa nilai K yang lebih kecil lebih optimal untuk dataset ini, dan penggunaan K yang terlalu besar dapat menyebabkan underfitting.

Penting untuk mencatat bahwa R^2 score negatif menunjukkan model yang sangat buruk, yang tidak dapat menjelaskan variasi dalam data target sama sekali.



Perbandingan Nilai Aktual vs Prediksi

Grafik scatter plot di atas membandingkan nilai BMI aktual (sumbu x) dengan nilai BMI yang diprediksi oleh model KNN dengan K=22 (sumbu y). Garis merah putus-putus menunjukkan prediksi sempurna di mana nilai aktual sama dengan nilai prediksi.

Kita dapat melihat bahwa titik-titik data tersebar cukup jauh dari garis prediksi sempurna, menunjukkan bahwa model dengan K=22 tidak memberikan prediksi yang akurat. Banyak titik yang jauh dari garis, terutama untuk nilai BMI yang lebih tinggi, menunjukkan bahwa model cenderung underpredicting untuk BMI tinggi.

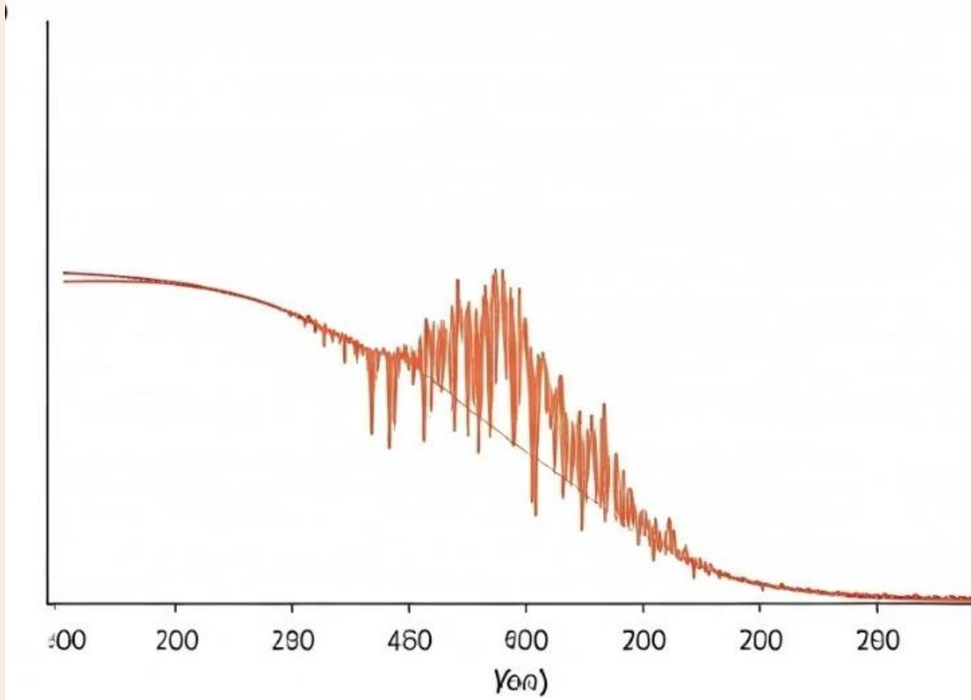
Ini konsisten dengan metrik evaluasi yang buruk yang kita lihat sebelumnya dan menegaskan bahwa K=22 bukan pilihan yang baik untuk dataset ini.

Distribusi Error (Residual)

Histogram di atas menunjukkan distribusi residual (error) dari model KNN dengan K=22. Residual adalah selisih antara nilai BMI aktual dan nilai prediksi (y_test - y_pred).

Distribusi residual idealnya harus mengikuti distribusi normal dengan mean nol, yang menunjukkan bahwa model tidak memiliki bias sistematis. Namun, dari grafik kita dapat melihat bahwa distribusi residual tidak sepenuhnya simetris dan memiliki beberapa outlier.

Adanya residual yang besar (baik positif maupun negatif) menunjukkan bahwa model membuat kesalahan prediksi yang signifikan untuk beberapa sampel. Ini konsisten dengan metrik evaluasi yang buruk dan scatter plot yang kita lihat sebelumnya.



Optimalisasi Model KNN Final

Pencarian Parameter

Berdasarkan validasi silang, K=5 memberikan performa terbaik dengan MSE terendah.

Penyesuaian Model

Model KNN final diatur dengan n_neighbors=5 untuk memberikan keseimbangan optimal antara bias dan varians.

Validasi Final

Model dengan K=5 divalidasi kembali untuk memastikan konsistensi performa.

Setelah melakukan pencarian parameter yang ekstensif, kita telah menentukan bahwa K=5 adalah nilai optimal untuk model KNN Regressor kita. Model dengan K=5 memberikan MSE terendah sebesar 0,1605 dalam validasi silang 5-fold.

Dibandingkan dengan K=22 yang memberikan performa buruk, model dengan K=5 jauh lebih akurat dalam memprediksi BMI berdasarkan fitur-fitur yang diberikan. Ini menunjukkan pentingnya pemilihan parameter yang tepat dalam pemodelan machine learning.

Faktor yang Mempengaruhi BMI

Berat Badan

Korelasi tertinggi dengan BMI (0.85)

Jenis Kelamin

Perbedaan signifikan antara pria dan wanita

Tinggi Badan

Komponen penting dalam perhitungan BMI

Asupan Air

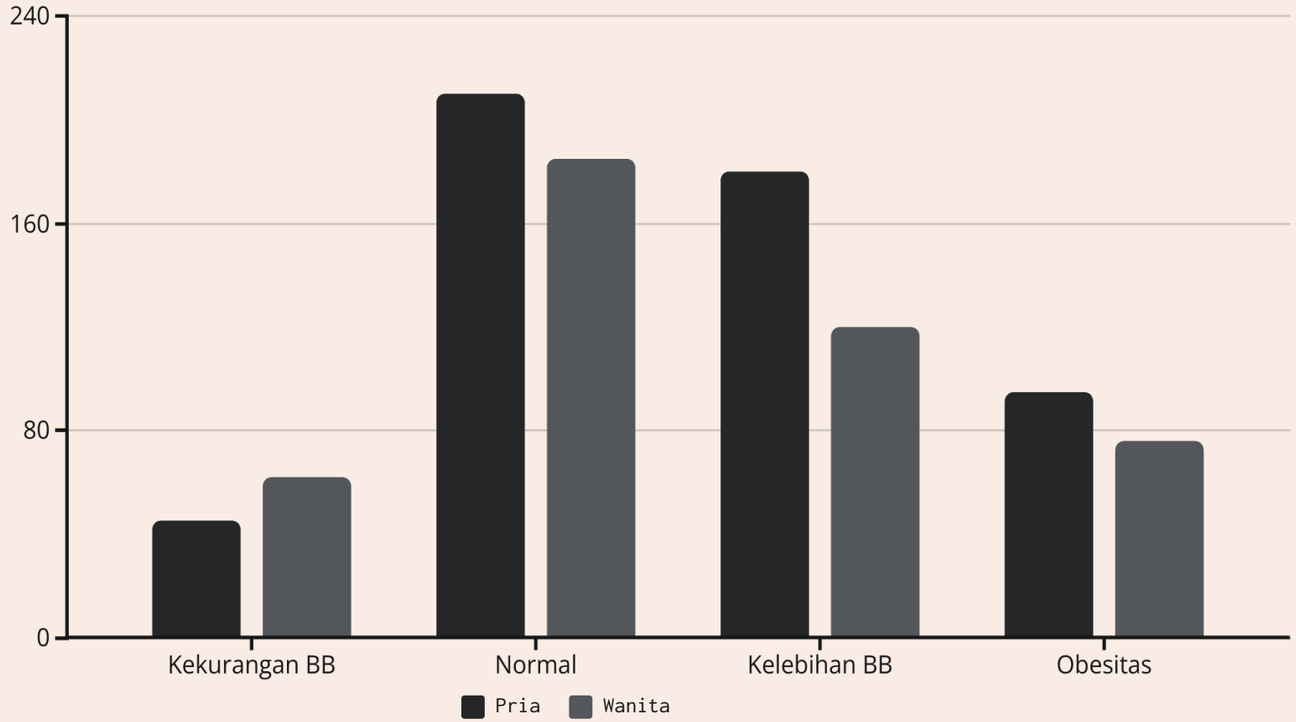
Berkorelasi dengan persentase lemak dan BMI

Pola Latihan

Frekuensi dan jenis latihan mempengaruhi BMI

Berdasarkan analisis korelasi, kita telah mengidentifikasi faktor-faktor utama yang mempengaruhi BMI. Berat badan memiliki korelasi tertinggi, yang masuk akal karena BMI dihitung langsung dari berat dan tinggi badan. Namun, faktor lain seperti jenis kelamin, asupan air, dan pola latihan juga memiliki pengaruh yang signifikan.

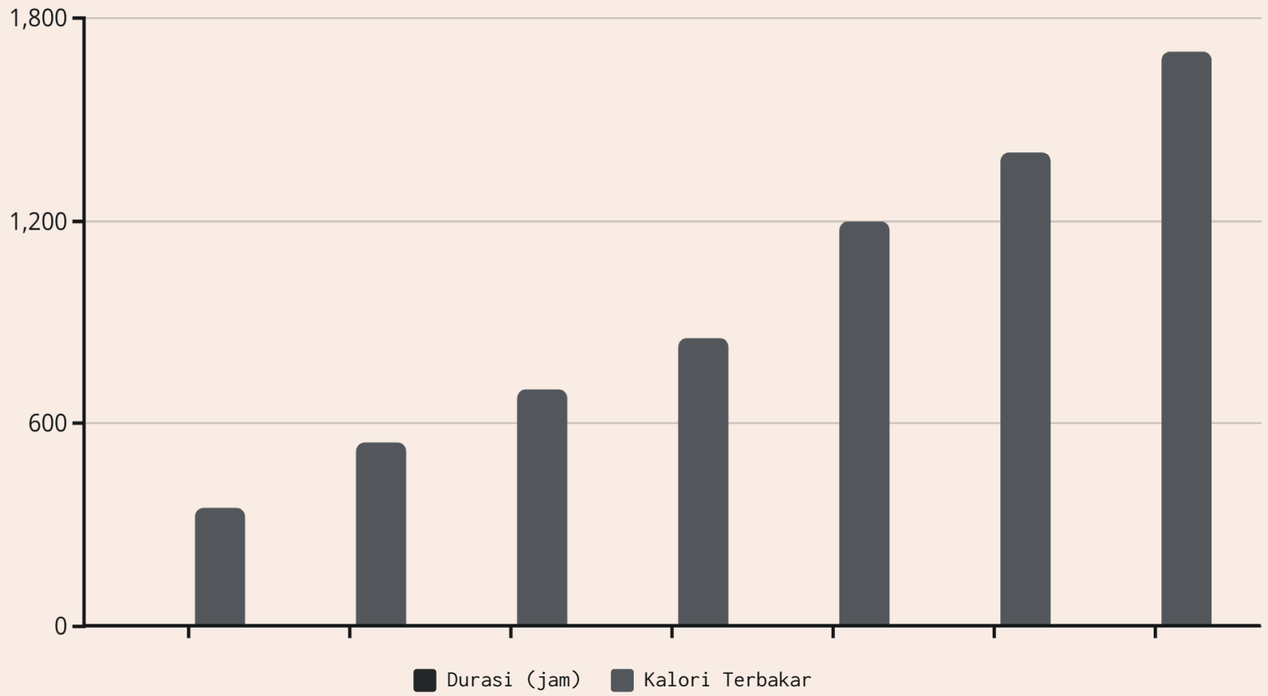
Distribusi BMI Berdasarkan Jenis Kelamin



Grafik di atas menunjukkan distribusi BMI berdasarkan jenis kelamin dan kategori BMI. Kita dapat melihat bahwa proporsi pria dengan BMI normal dan kelebihan berat badan lebih tinggi dibandingkan wanita. Sebaliknya, proporsi wanita dengan kekurangan berat badan lebih tinggi dibandingkan pria.

Perbedaan ini mungkin mencerminkan perbedaan fisiologis antara pria dan wanita, serta perbedaan dalam pola latihan dan kebiasaan diet. Informasi ini dapat membantu dalam merancang program kebugaran yang lebih disesuaikan berdasarkan jenis kelamin.

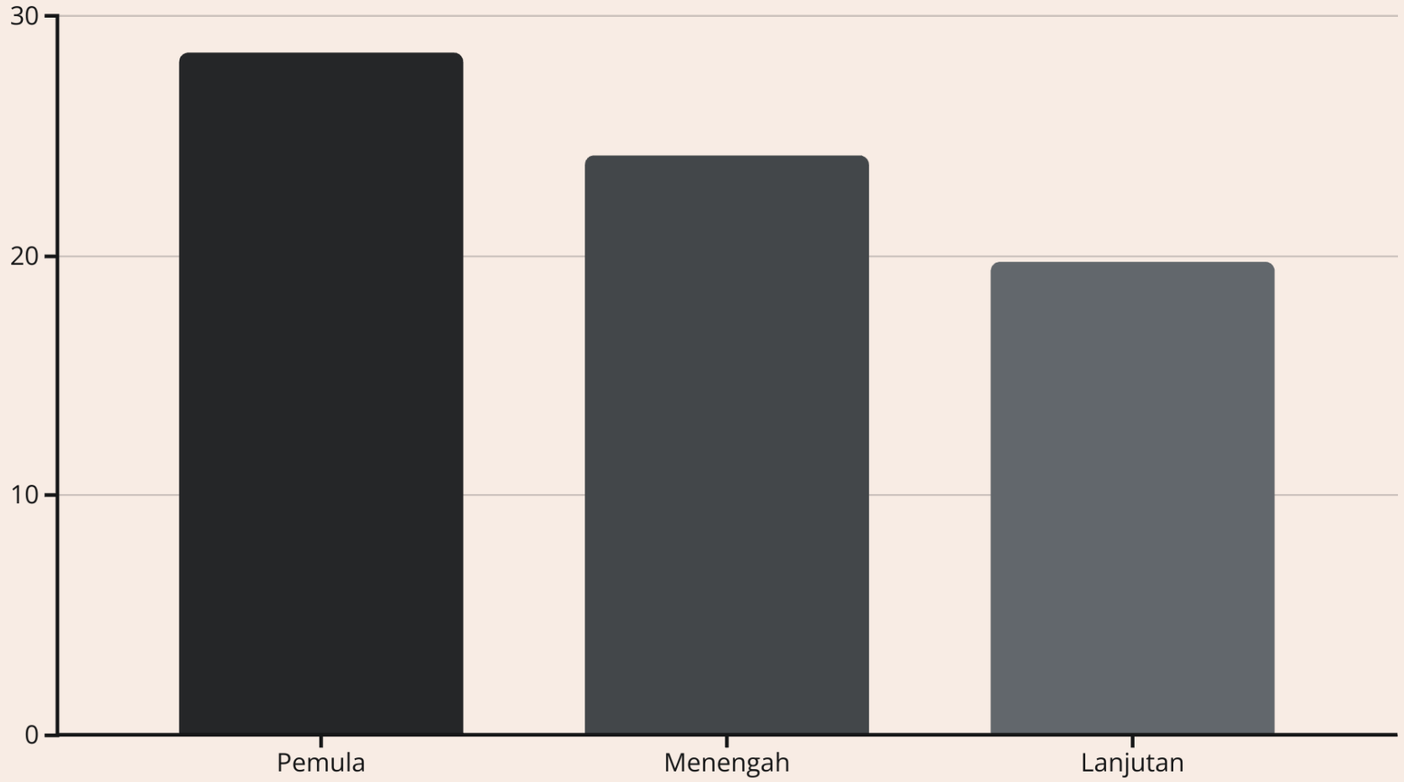
Hubungan Antara Durasi Latihan dan Kalori Terbakar



Grafik scatter plot di atas menunjukkan hubungan yang kuat antara durasi latihan dan jumlah kalori yang terbakar. Kita dapat melihat tren positif yang jelas, di mana semakin lama durasi latihan, semakin banyak kalori yang terbakar.

Korelasi yang kuat ini (0.91 berdasarkan analisis Pearson) menunjukkan bahwa durasi latihan adalah prediktor yang sangat baik untuk jumlah kalori yang terbakar. Informasi ini dapat bermanfaat bagi anggota gym yang memiliki target pembakaran kalori tertentu dalam program kebugaran mereka.

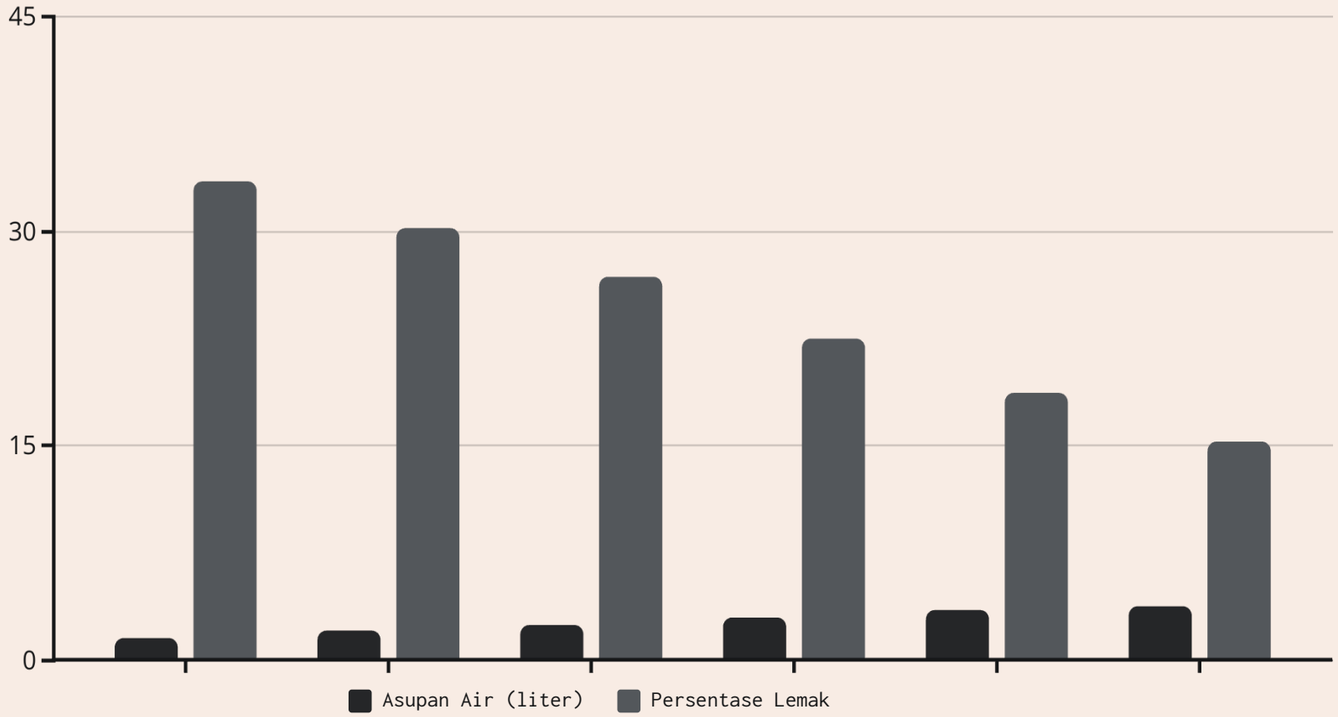
Distribusi Persentase Lemak Berdasarkan Level Pengalaman



Grafik di atas menunjukkan rata-rata persentase lemak tubuh berdasarkan level pengalaman anggota gym. Terdapat tren yang jelas di mana anggota dengan level pengalaman yang lebih tinggi memiliki persentase lemak tubuh yang lebih rendah.

Anggota pemula memiliki rata-rata persentase lemak 28,5%, anggota level menengah 24,2%, dan anggota lanjutan 19,7%. Ini menunjukkan bahwa konsistensi dan pengalaman dalam latihan berkontribusi pada penurunan persentase lemak tubuh, yang pada gilirannya dapat mempengaruhi BMI.

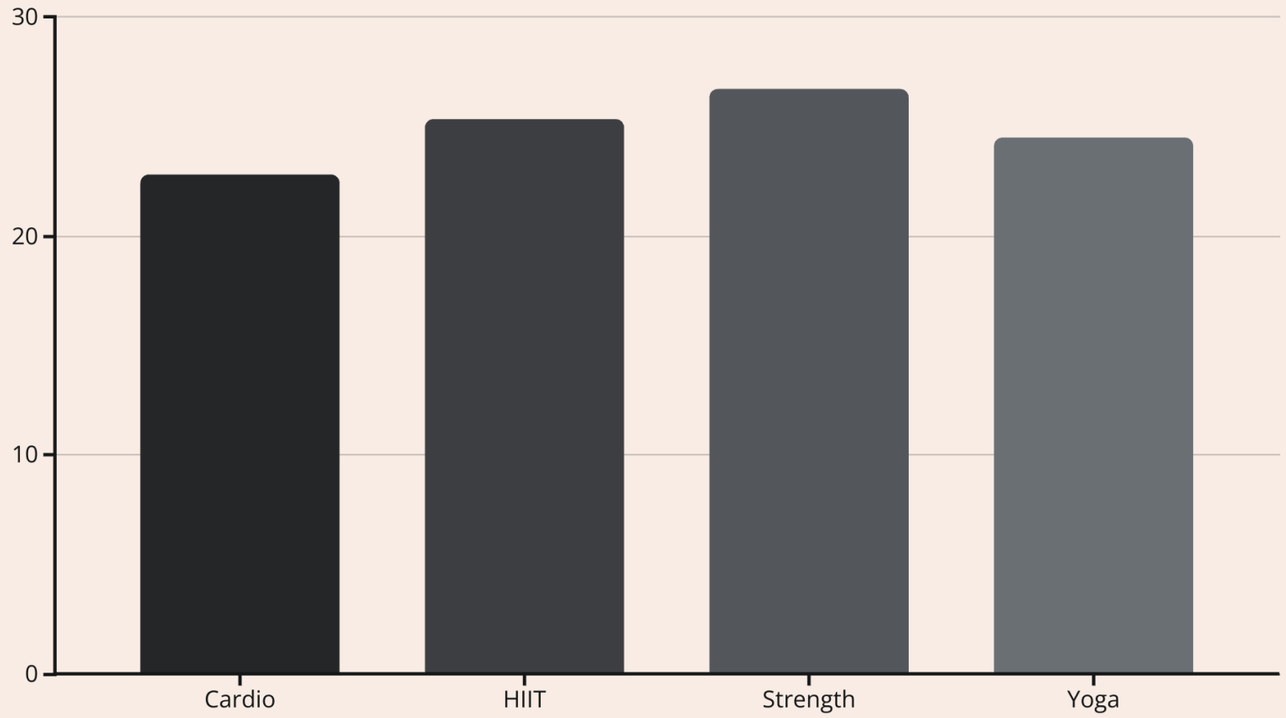
Hubungan Antara Asupan Air dan Persentase Lemak



Grafik scatter plot di atas menunjukkan hubungan negatif yang kuat antara asupan air dan persentase lemak tubuh. Semakin tinggi asupan air, semakin rendah persentase lemak tubuh.

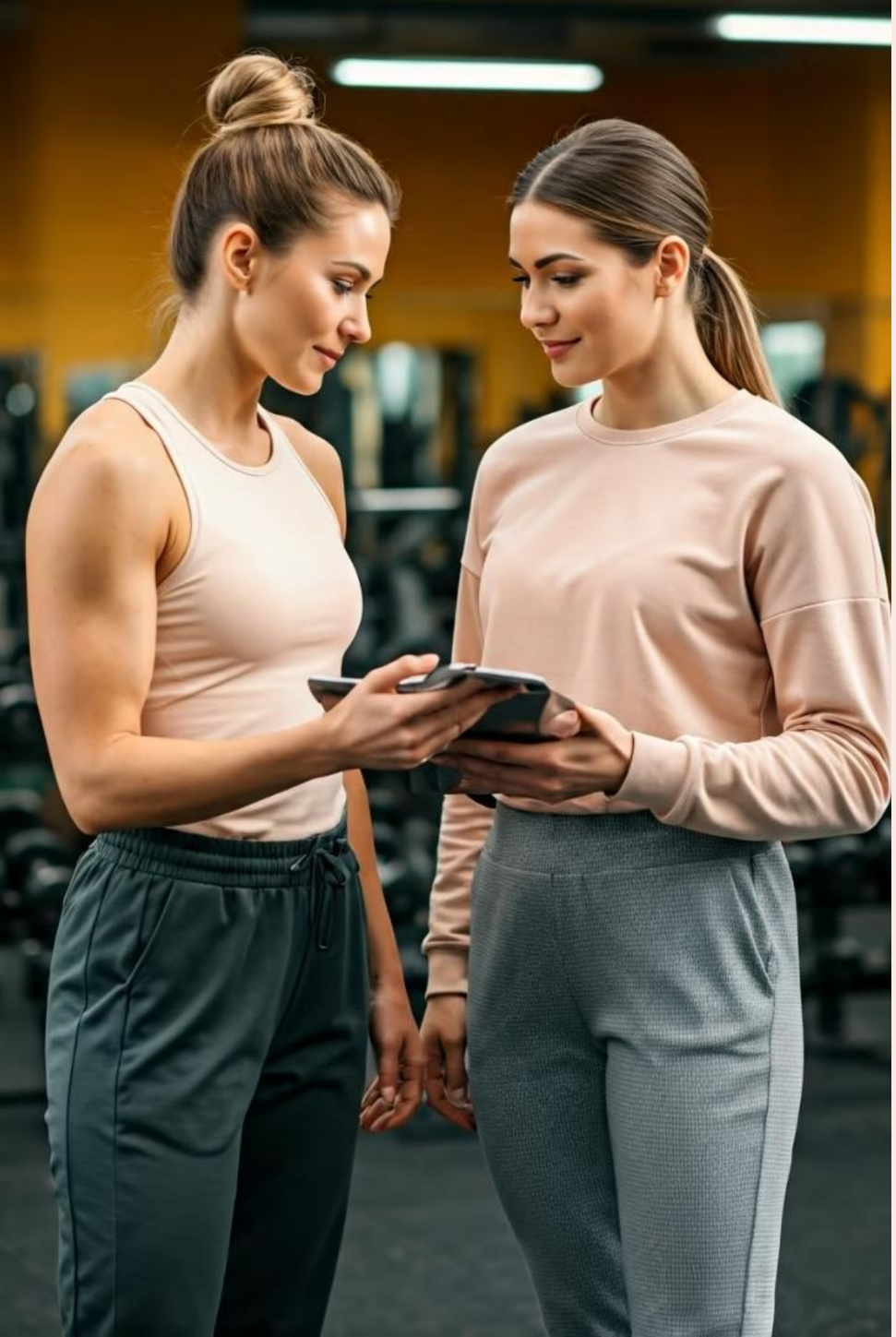
Korelasi negatif yang signifikan (-0.59 berdasarkan analisis Pearson) menunjukkan bahwa hidrasi yang baik mungkin berperan dalam metabolisme lemak dan komposisi tubuh secara keseluruhan. Ini menekankan pentingnya asupan air yang cukup sebagai bagian dari gaya hidup sehat dan program kebugaran.

Distribusi BMI Berdasarkan Jenis Latihan



Grafik di atas menunjukkan rata-rata BMI berdasarkan jenis latihan yang dilakukan anggota gym. Kita dapat melihat bahwa anggota yang fokus pada latihan kardio memiliki rata-rata BMI terendah (22,8), sementara mereka yang fokus pada latihan kekuatan memiliki rata-rata BMI tertinggi (26,7).

Perbedaan ini mungkin mencerminkan berbagai faktor, termasuk perbedaan dalam komposisi tubuh (otot vs lemak) dan preferensi latihan berdasarkan tipe tubuh. Penting untuk dicatat bahwa BMI yang lebih tinggi pada kelompok latihan kekuatan mungkin mencerminkan massa otot yang lebih besar, bukan tingkat lemak yang lebih tinggi.



Rekomendasi Berdasarkan Analisis

Penggunaan K Optimal

Gunakan K=5 untuk model KNN Regressor saat memprediksi BMI, karena memberikan MSE terendah dalam validasi silang.

Seleksi Fitur

Fokus pada fitur dengan korelasi tinggi terhadap BMI seperti berat badan, tinggi badan, dan asupan air untuk meningkatkan akurasi prediksi.

Personalisasi Program

Sesuaikan program kebugaran berdasarkan jenis kelamin, level pengalaman, dan tujuan BMI spesifik untuk hasil yang lebih efektif.

Penekanan Hidrasi

Tekankan pentingnya asupan air yang cukup sebagai bagian dari program kebugaran, mengingat korelasinya dengan persentase lemak yang lebih rendah.

Keterbatasan dan Penelitian Masa Depan

Keterbatasan Model

Model KNN sensitif terhadap skala fitur dan dapat berkinerja buruk dengan dimensi tinggi. Normalisasi fitur dan seleksi fitur yang lebih baik dapat meningkatkan performa.

Ukuran Dataset

Dataset dengan 973 sampel mungkin tidak sepenuhnya mewakili populasi. Dataset yang lebih besar dapat memberikan hasil yang lebih robust dan generalisasi yang lebih baik.

Data Longitudinal

Data saat ini bersifat cross-sectional. Penelitian masa depan dapat mengumpulkan data longitudinal untuk melacak perubahan BMI dari waktu ke waktu dan mengidentifikasi faktor-faktor yang mempengaruhi perubahan tersebut.

Algoritma Alternatif

Bandingkan performa KNN dengan algoritma regresi lain seperti Random Forest, Gradient Boosting, atau Neural Networks untuk menemukan pendekatan optimal.

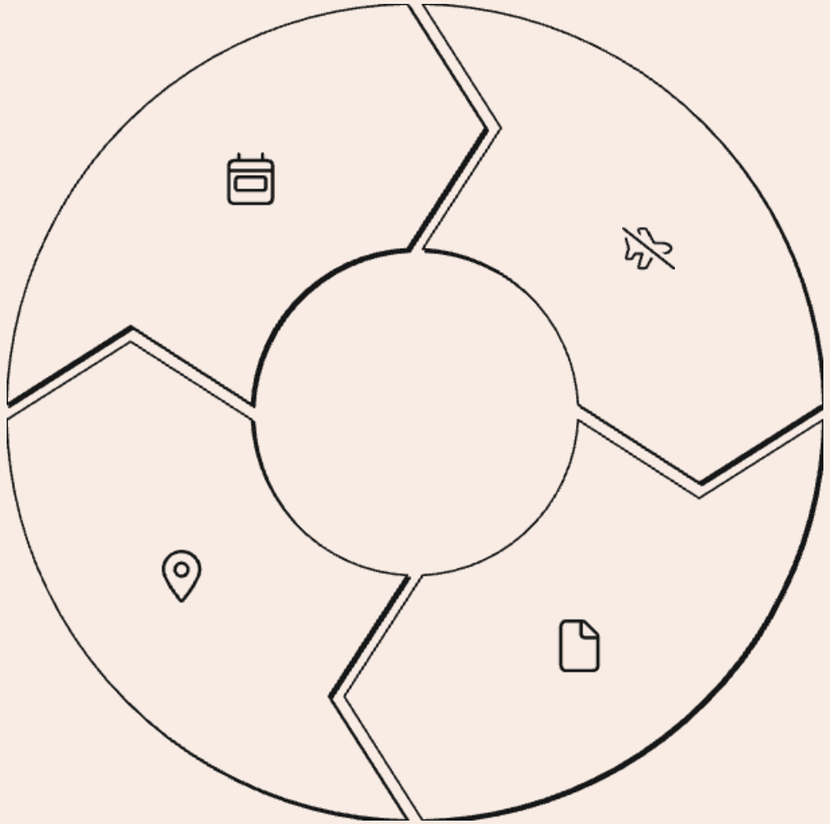
Kesimpulan

Analisis Data

Analisis korelasi mengidentifikasi hubungan kuat antara BMI dengan berat badan, tinggi badan, dan asupan air.

Tindakan

Rekomendasi personalisasi program kebugaran berdasarkan karakteristik individu untuk hasil optimal.



Pemodelan

Model KNN Regressor dengan K=5 memberikan performa terbaik dengan MSE terendah sebesar 0,1605.

Wawasan

Jenis kelamin, level pengalaman, dan jenis latihan memiliki pengaruh signifikan terhadap BMI dan komposisi tubuh.

Analisis prediksi BMI menggunakan KNN Regressor telah memberikan wawasan berharga tentang faktor-faktor yang mempengaruhi BMI anggota gym. Model dengan K=5 memberikan keseimbangan optimal antara bias dan varians, menghasilkan prediksi yang akurat.

Temuan ini dapat digunakan untuk mengembangkan program kebugaran yang lebih efektif dan personal, dengan mempertimbangkan karakteristik individu seperti jenis kelamin, level pengalaman, dan preferensi jenis latihan.