PREDIKSI TINGKAT OBESITAS MENGGUNAKAN RANDOM FOREST

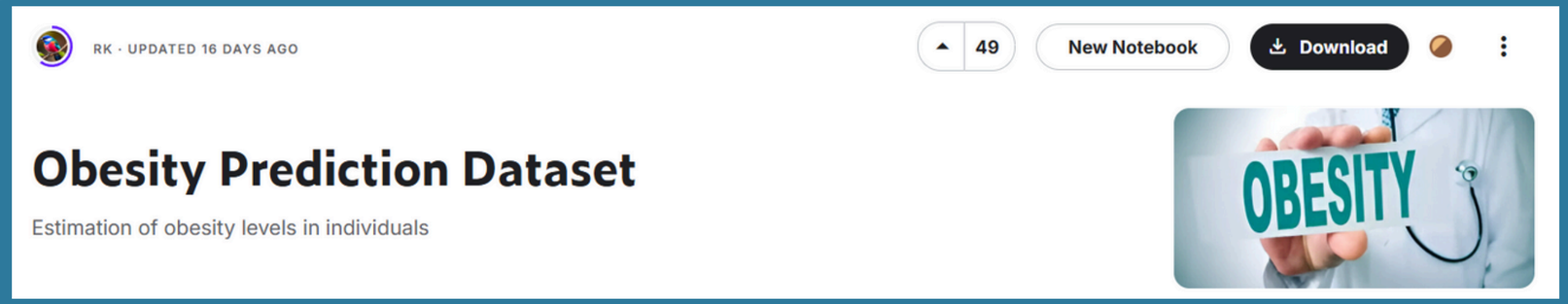
by: Dian Pandu Syahfitra

MASALAH KESEHATAN GLOBAL

Menurut data terbaru dari Organisasi Kesehatan Dunia (WHO) pada tahun 2022, **sekitar 2,5 miliar orang** dewasa berusia 18 tahun ke atas mengalami kelebihan berat badan, dengan 890 juta di antaranya terklasifikasi sebagai obesitas. Obesitas dapat menyebabkan berbagai masalah kesehatan, seperti Penyakit jantung (penyebab utama kematian di dunia), Diabetes tipe 2, Hipertensi, Gangguan tidur hingga Kanker tertentu

Dengan latar belakang masalah obesitas yang serius, proyek ini bertujuan untuk membuat prediksi tingkat obesitas pada individu berdasarkan faktor-faktor gaya hidup dan kesehatan dari dataset kaggle “Obesity Prediction”, menggunakan **algoritma Random Forest.**

DESKRIPSI DATASET

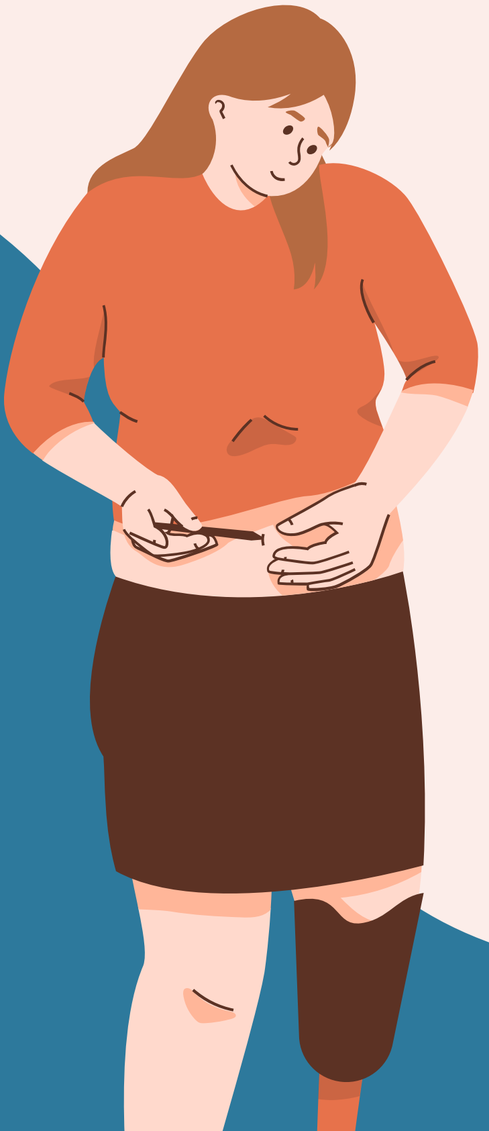


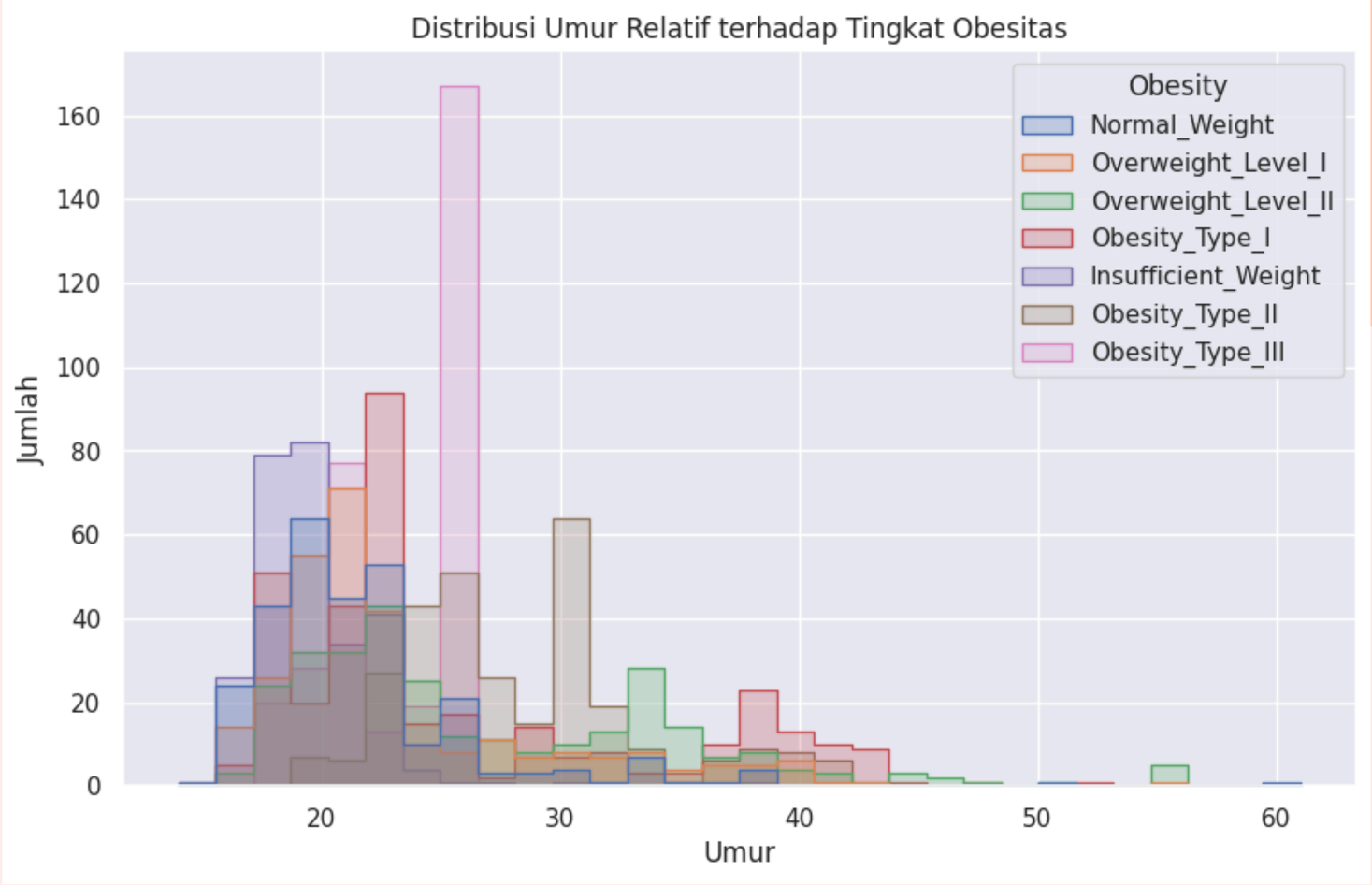
Sumber Data: **Dataset Kaggle Obesity Prediction**

Jumlah Data ; **2111 entri, 17 kolom**

Dataset ini mencakup informasi demografi (Jenis Kelamin, Usia), pengukuran fisik (Tinggi, Berat Badan), serta faktor gaya hidup seperti riwayat obesitas keluarga, pola makan (FAVC, FCVC, NCP, CAEC), dan kebiasaan merokok (SMOKE). Fitur perilaku meliputi konsumsi air (CH2O), pemantauan kalori (SCC), aktivitas fisik (FAF), penggunaan perangkat elektronik (TUE), dan konsumsi alkohol (CALC). Moda transportasi (MTRANS) juga disertakan, dengan variabel target Obesity (Tingkat Obesitas: Normal, Overweight, Obesity Level 1, 2, 3).

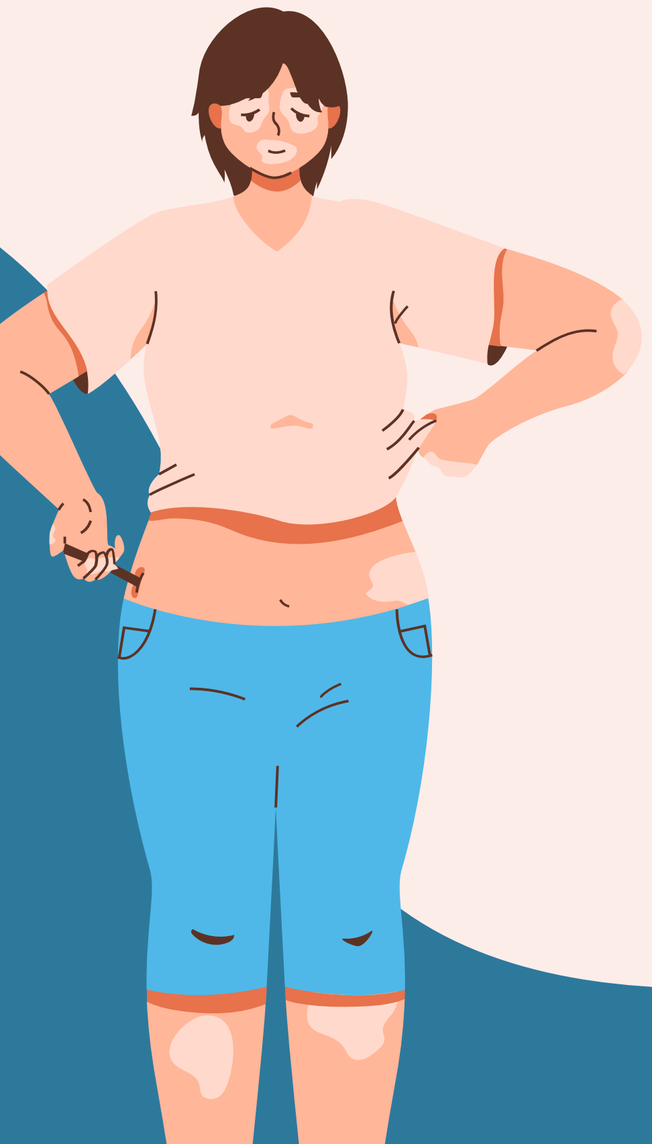
EDA

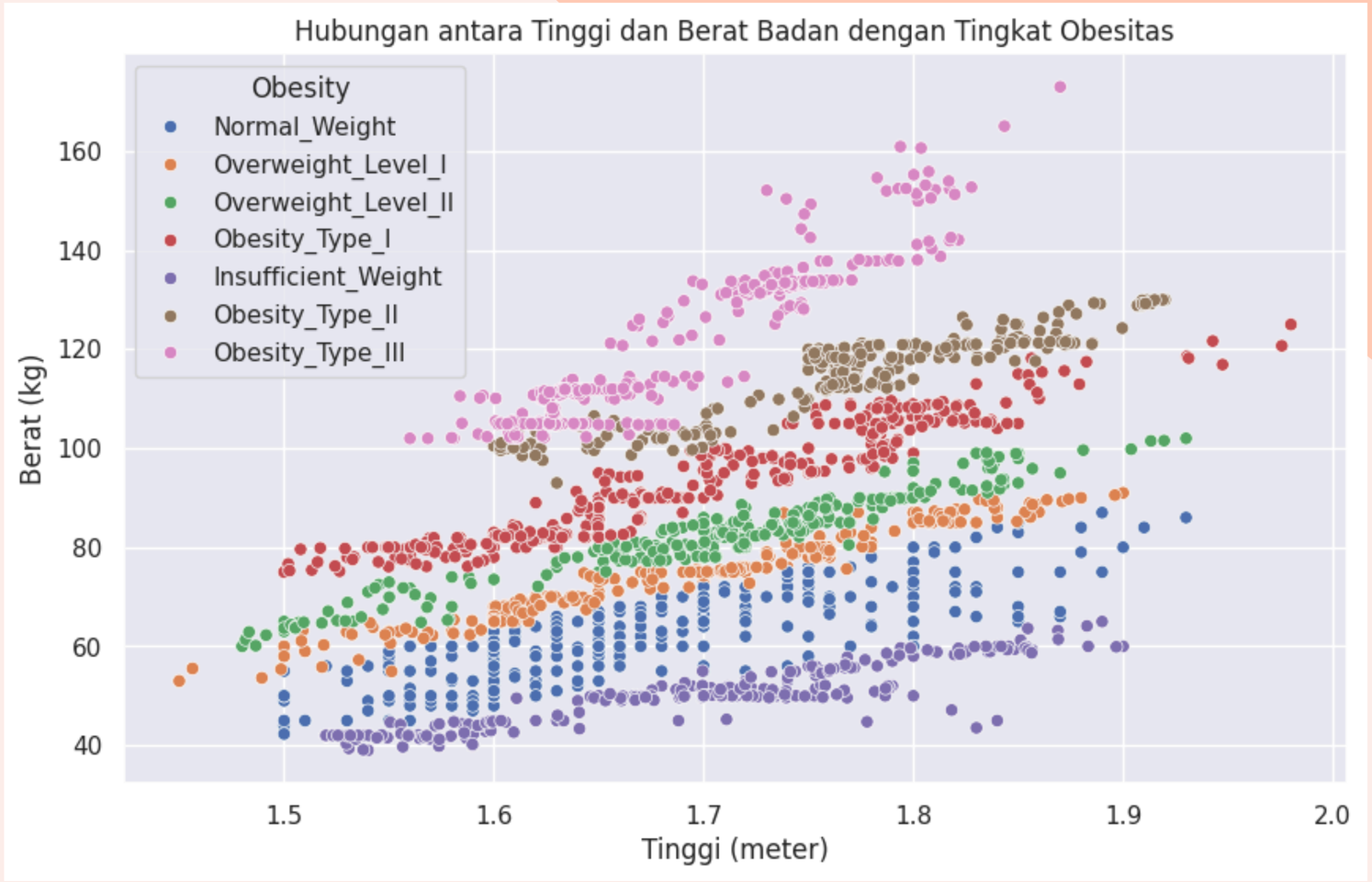




Grafik ini menunjukkan distribusi usia terhadap tingkat obesitas. Usia 20-an mendominasi hampir semua kategori, terutama Overweight dan Obesity Type I, menunjukkan variasi obesitas yang lebih besar di usia muda.

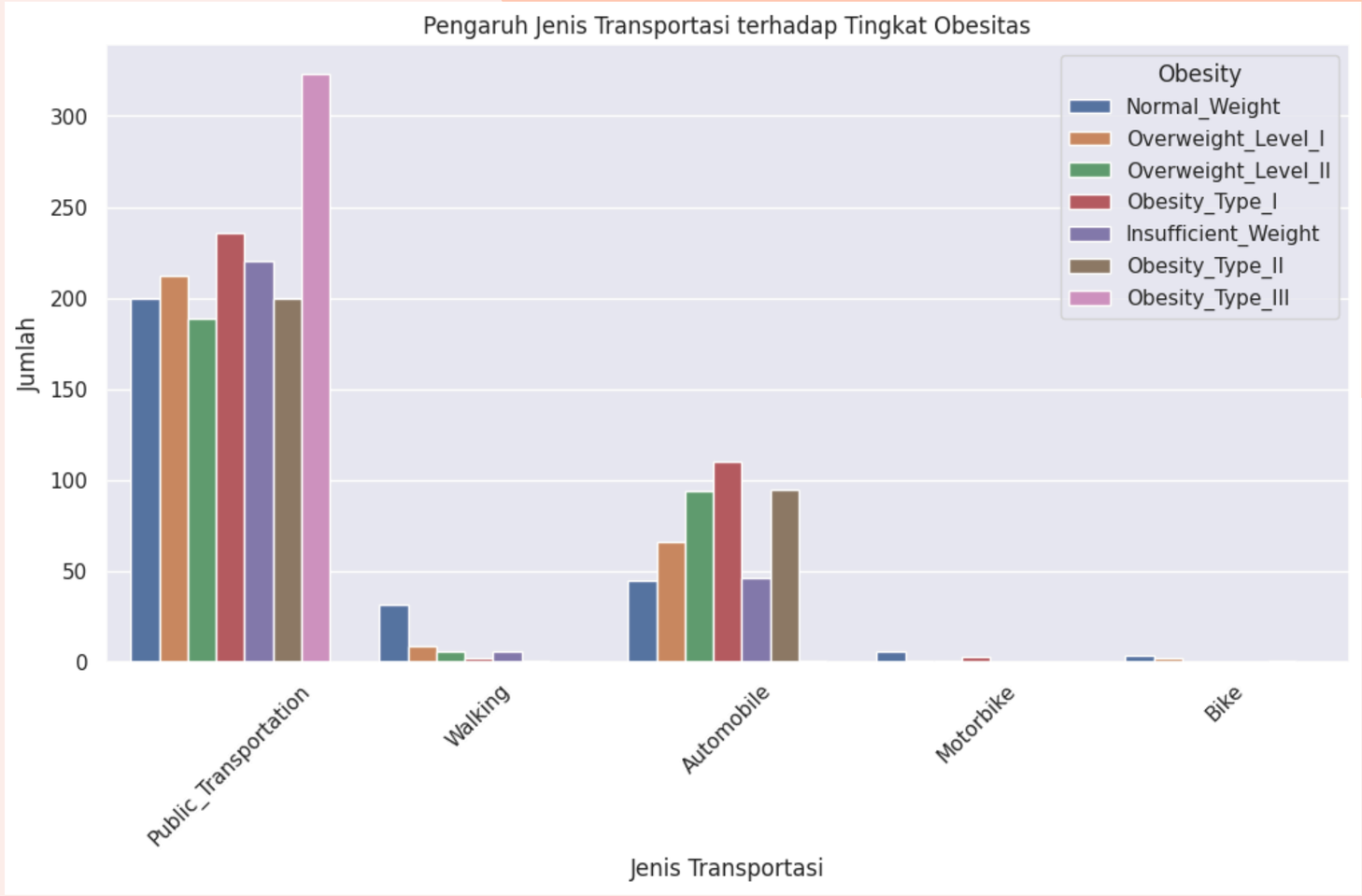
EDA





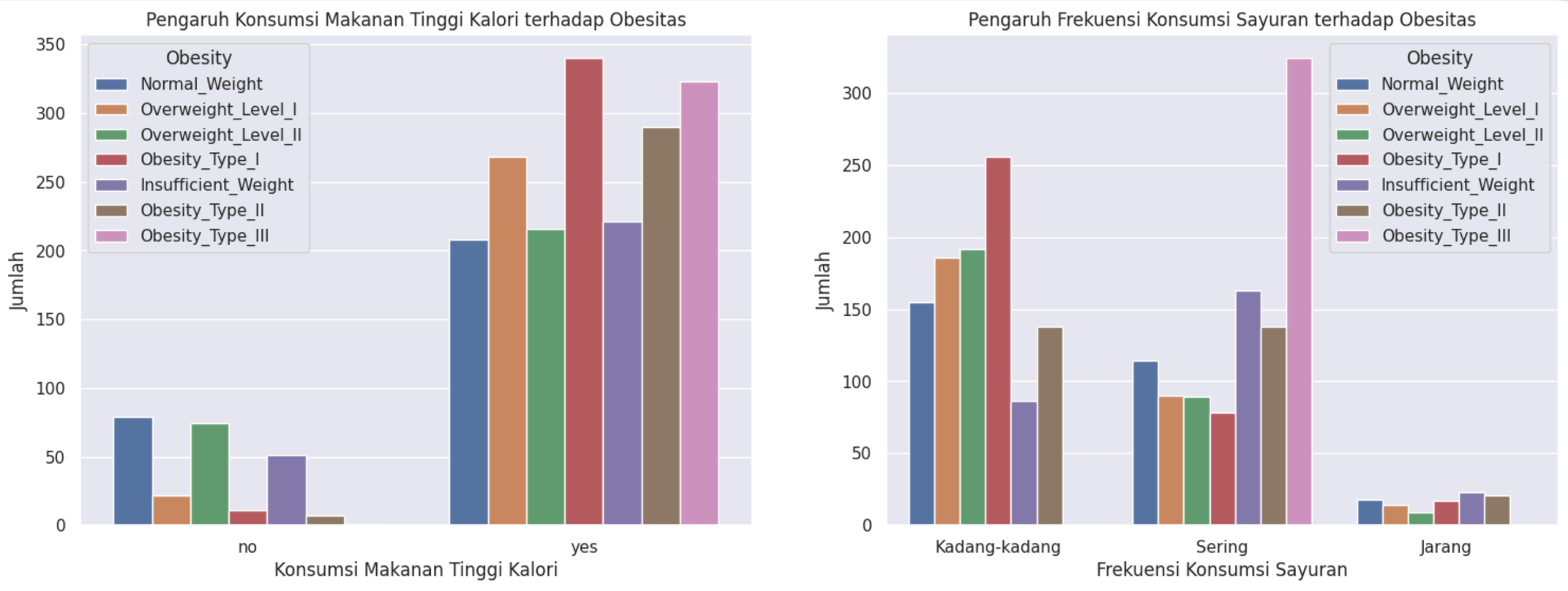
Visualisasi ini menunjukkan hubungan antara tinggi dan berat badan, di mana semakin tinggi berat badan relatif terhadap tinggi, semakin tinggi tingkat obesitas. Hal ini menegaskan hubungan positif antara berat badan dan obesitas.

EDA



Penggunaan transportasi umum cenderung dikaitkan dengan tingkat obesitas yang lebih rendah dibandingkan mobil atau motor, kemungkinan karena lebih banyak aktivitas fisik seperti berjalan ke halte atau stasiun.

EDA

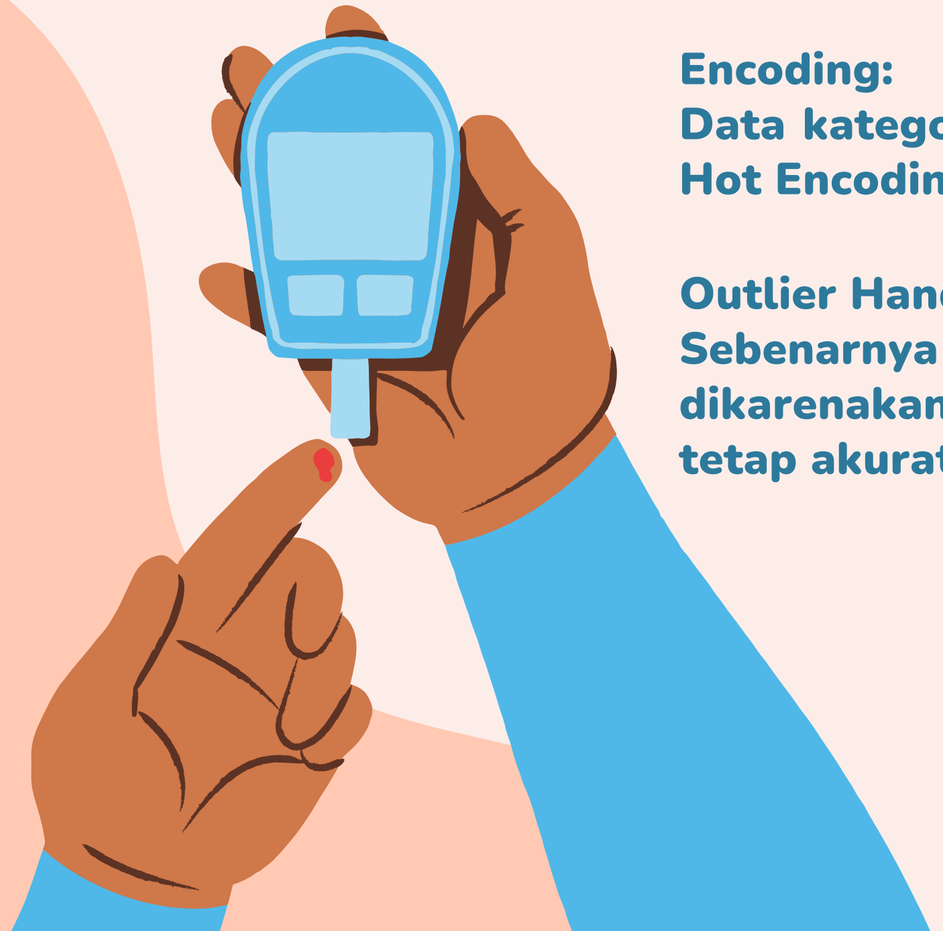


Grafik menunjukkan bahwa konsumsi makanan tinggi kalori berkontribusi signifikan terhadap obesitas, sementara frekuensi konsumsi sayuran tidak selalu berkorelasi dengan berat badan yang lebih sehat. Individu yang sering makan makanan tinggi kalori lebih cenderung mengalami obesitas tingkat tinggi, sedangkan konsumsi sayuran yang tinggi tidak selalu mengurangi risiko obesitas, kemungkinan karena faktor lain seperti pola makan keseluruhan dan aktivitas fisik.

PRA-PROCESSING DATA

Cek Missing Values:
Semua data sudah lengkap (no missing values).

Duplikasi Data:
Ditemukan 24 data duplikat dan dihapus untuk menjaga kualitas analisis.



Encoding:
Data kategorikal dikonversi ke format numerik (Label Encoding dan One-Hot Encoding)

Outlier Handling:
Sebenarnya terdapat beberapa outlier, namun tidak saya bersihkan dikarenakan akan kehilangan terlalu banyak informasi, memastikan model tetap akurat dan representatif.

```
Jumlah outlier per fitur:
 Age        167
Height        1
Weight        1
FCVC          0
NCP         577
CH2O          0
FAF           0
TUE           0
```

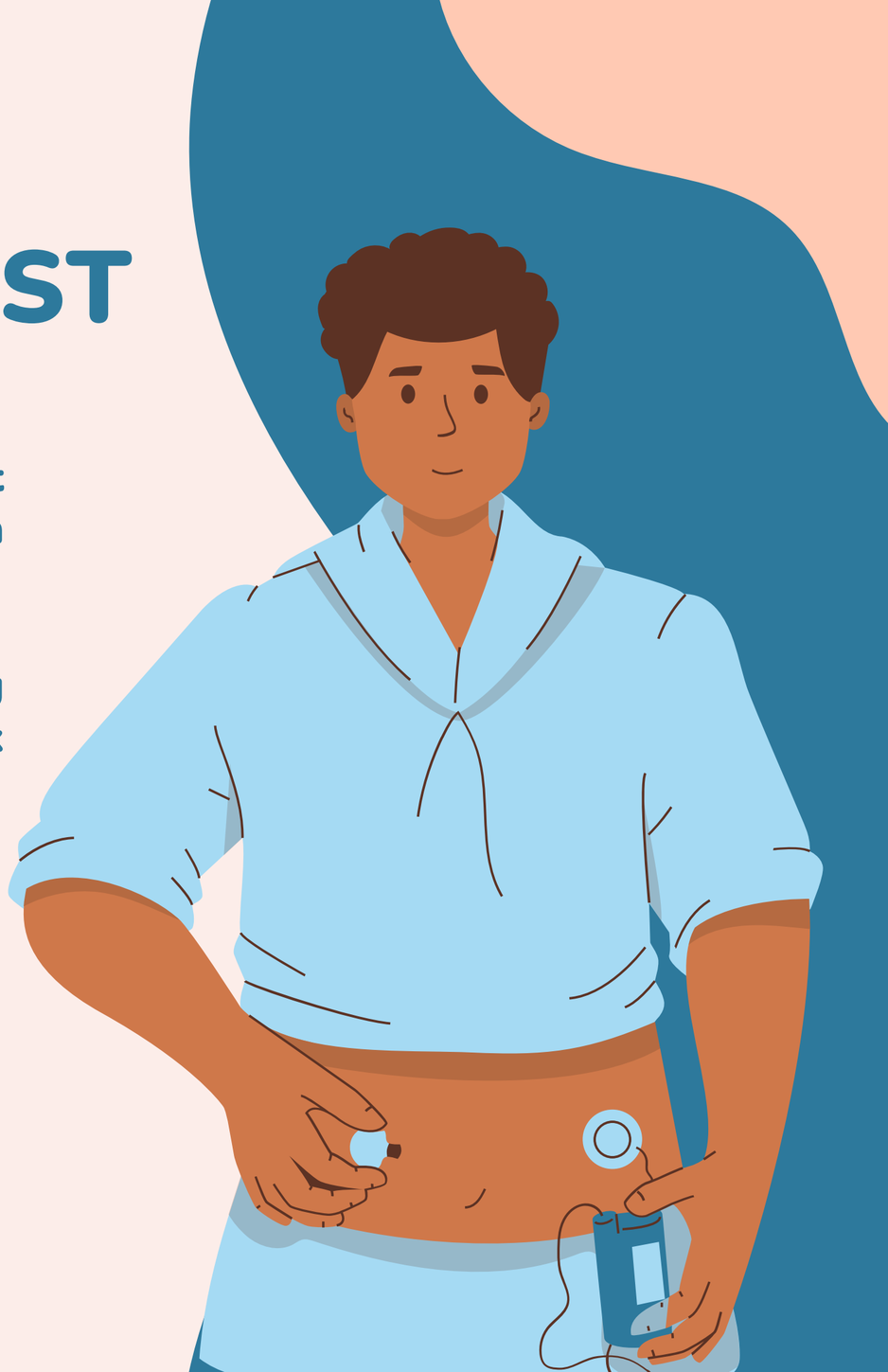
MODEL RANDOM FOREST

Keunggulan: Dapat menangani data numerik dan kategorikal, robust terhadap overfitting, mampu menangani missing values, dan memberikan hasil yang akurat.

Proses: Random Forest adalah ensemble method yang menggabungkan beberapa pohon keputusan (decision trees) untuk meningkatkan performa.

Parameter yang Digunakan:

- **n_estimators: 100 pohon keputusan.**
- **max_depth: Tidak dibatasi, otomatis disesuaikan oleh model.**



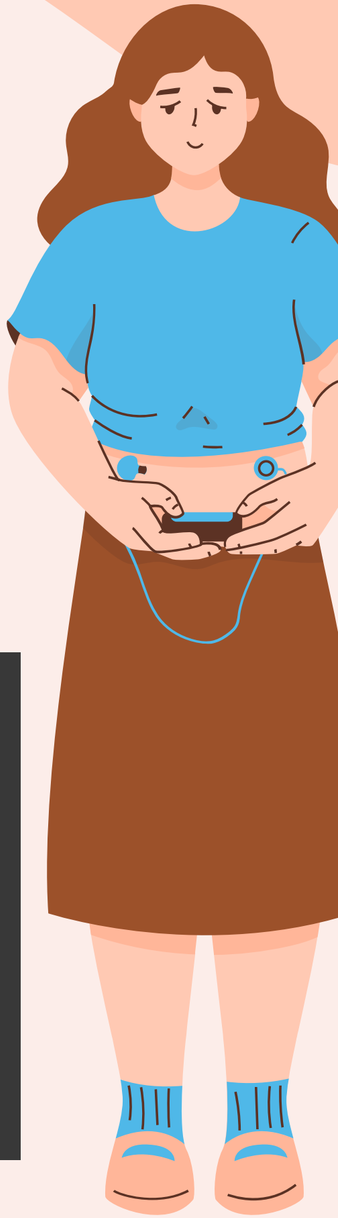
EVALUASI MODEL

```
Accuracy (Random Forest): 0.9450
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        53
           1       0.77      0.95      0.85        57
           2       0.99      0.96      0.97        70
           3       1.00      1.00      1.00        60
           4       1.00      0.98      0.99        65
           5       0.92      0.85      0.89        55
           6       0.96      0.91      0.94        58

    accuracy                           0.94       418
   macro avg       0.95      0.94      0.94       418
weighted avg       0.95      0.94      0.95       418
```

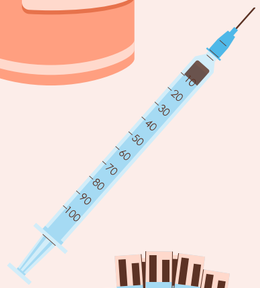
Model memiliki akurasi sebesar 94.50%, yang berarti dari seluruh prediksi yang dilakukan, 94.5% di antaranya benar. Ini menunjukkan bahwa model bekerja dengan sangat baik dalam mengklasifikasikan data.

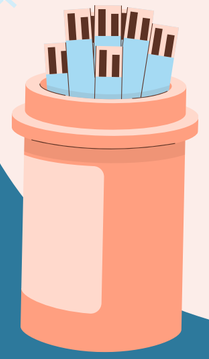
Walaupun ada beberapa kesalahan klasifikasi, jumlahnya sangat kecil dibandingkan total data, menunjukkan model bekerja dengan sangat baik.

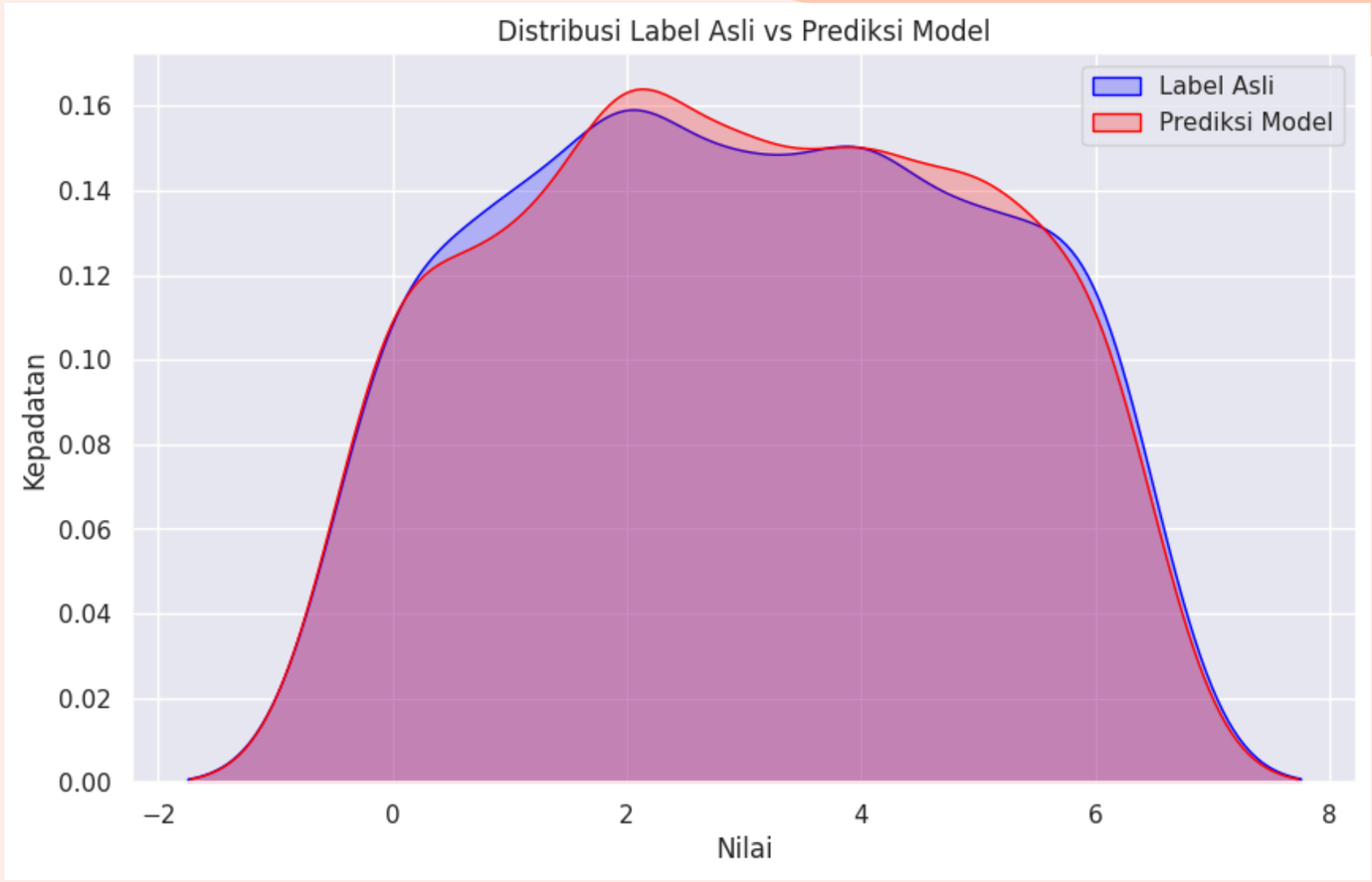
```
Confusion Matrix (Random Forest):
[[50  3  0  0  0  0  0]
 [ 0 54  0  0  0  3  0]
 [ 0  1 67  0  0  0  2]
 [ 0  0  0 60  0  0  0]
 [ 0  0  1  0 64  0  0]
 [ 0  8  0  0  0 47  0]
 [ 0  4  0  0  0  1 53]]
```



LABEL VS MODEL







Visualisasi ini membandingkan distribusi label asli dengan prediksi model menggunakan grafik kepadatan untuk menilai seberapa baik model memprediksi data aktual. Tumpang tindih yang tinggi menunjukkan performa model yang baik, sedangkan perbedaan signifikan menandakan perlunya optimasi lebih lanjut.

THANK YOU!