Laporan Proyek Machine Learning - Fetra Abdul Malik_2306039 -tsani hisni amala(2306050)

Domain Proyek

Obesitas merupakan masalah kesehatan global yang meningkatkan risiko penyakit kronis seperti diabetes, jantung, dan hipertensi. Deteksi dini melalui model prediktif dapat membantu intervensi preventif. Menurut WHO (2023), prevalensi obesitas global telah meningkat tiga kali lipat sejak 1975. Implementasi machine learning seperti algoritma C4.5 efektif untuk klasifikasi pola obesitas berdasarkan faktor risiko (Quinlan, 1993).

Referensi

Referensi:

https://scholar.ummetro.ac.id/index.php/jiki/article/view/7191Quinlan, J. R. (1993)
https://ejournal.itn.ac.id/index.php/jati/article/view/10498

https://ejurnal.itats.ac.id/snestik/article/view/7619

Business Understanding

Problem Statements

1. Bagaimana memanfaatkan data kesehatan untuk memprediksi risiko obesitas pada seseorang??

   Obesitas sering kali tidak terdeteksi pada tahap awal, sehingga intervensi baru dilakukan setelah muncul komplikasi kesehatan serius seperti diabetes, hipertensi, dan penyakit jantung. Dengan menggunakan data kesehatan yang tersedia, diperlukan metode untuk memprediksi risiko obesitas secara cepat dan akurat.

2. Algoritma machine learning apa yang paling efektif dalam memprediksi risiko obesitas?

   Pemilihan algoritma yang tepat sangat penting untuk memastikan prediksi risiko obesitas memiliki akurasi tinggi dengan performa yang optimal. Algoritma C4.5 dikenal mampu menangani data campuran dan menghasilkan model yang mudah diinterpretasi.

3. Bagaimana meningkatkan akurasi model prediksi risiko obesitas?

Selain memilih algoritma yang tepat, diperlukan strategi untuk mengoptimalkan model, seperti menggunakan teknik balancing data, hyperparameter tuning, atau pendekatan ensemble

Goals

1. Membangun model prediksi level obesitas menggunakan C4.5.
2. Mengevaluasi performa model dengan metrik akurasi, precision, recall, dan F1-score.
3. Meningkatkan akurasi melalui teknik handling data tidak seimbang.

Solution Statement

Untuk mencapai tujuan proyek, langkah-langkah berikut akan dilakukan:

1. Eksplorasi dan Pemahaman Data (EDA))

   Dataset akan dianalisis untuk memahami distribusi fitur, hubungan antar variabel seperti usia, BMI, aktivitas fisik, pola makan, dan riwayat keluarga dengan tingkat obesitas, serta mendeteksi adanya outlier atau ketidakseimbangan kelas. Visualisasi data seperti histogram, bar chart, dan heatmap akan digunakan untuk mendukung analisis ini.

2. Implementasi Algoritma Machine Learning (C4.5)

   Model klasifikasi berbasis algoritma C4.5 (Decision Tree dengan kriteria entropy) akan digunakan untuk memprediksi level obesitas berdasarkan fitur-fitur kesehatan dan gaya hidup. Model ini dipilih karena kemampuannya menangani data numerik dan kategorik serta menghasilkan pohon keputusan yang mudah diinterpretasi.

3. Optimasi Model

   Model C4.5 akan dioptimalkan melalui penyesuaian parameter seperti kedalaman pohon (max_depth) dan penanganan data tidak seimbang (misal dengan SMOTE atau class_weight). Jika diperlukan, teknik pruning akan diterapkan untuk mencegah overfitting..

4. Evaluasi Performa Model

   Model akan dievaluasi menggunakan metrik seperti Akurasi, Precision, Recall, dan F1 Score pada data uji. Analisis confusion matrix juga akan dilakukan untuk memahami kekuatan dan kelemahan model dalam mengklasifikasikan masing-masing kelas obesitas.

Data Understanding

Deskripsi Dataset

Dataset yang digunakan dalam proyek ini adalah Obesity-Classification-.csv, yang dapat diakses di:

https://www.kaggle.com/datasets/samuelcortinhas/obesity-classification-dataset

Informasi Dataset

Informasi dataset diberikan menggunakan fungsi `data.info()`. Berikut adalah hasilnya:

Nama Atribut	Tipe Data	Deskripsi Singkat
ID	Numerik	Nomor urut data (identitas unik, tidak digunakan dalam modeling)
Age	Numerik	Usia responden (tahun)
Gender	Kategorikal	Jenis kelamin responden (Male/Female)
Height	Numerik	Tinggi badan (cm)
Weight	Numerik	Berat badan (kg)
BMI	Numerik	Body Mass Index (kg/m^2)
Label	Kategorikal	Kategori obesitas (Underweight, Normal Weight, Overweight, Obese)

Dataset tidak memiliki missing values, sehingga dapat langsung digunakan untuk proses analisis dan pemodelan.

Statistik Deskriptif

Berikut adalah statistik deskriptif untuk fitur numerik dalam dataset:

Statistik	ID	Age	Height	Weight	BMI
Count	108	108	108	108	108
Mean	56.05	46.56	166.57	59.49	20.55
Std Dev	31.92	24.72	27.87	28.86	7.58
Min	1	11	120	10	3.90
25%	28.75	27	140	35	16.70
50% (Median)	56.50	42.50	175	55	21.20
75%	83.25	59.25	190	85	26.10
Max	110	112	210	120	37.20

Exploratory Data Analysis (EDA)

1. Distribusi Usia

   Distribusi usia pada dataset obesitas menunjukkan rentang usia dewasa muda hingga lanjut usia. Sebagian besar data berpusat pada usia 25–55 tahun, yang merupakan kelompok usia produktif dan berisiko tinggi mengalami obesitas.

2. Distribusi BMI

   Indeks Massa Tubuh (BMI) dalam dataset didominasi oleh kategori overweight dan obesitas ringan, dengan nilai rata-rata BMI berada pada kisaran 27–32. Terdapat pula outlier pada kategori obesitas berat yang perlu diperhatikan lebih lanjut.

3. Hubungan Fitur dengan Obesitas

   - Aktivitas Fisik dan Obesitas: Individu dengan aktivitas fisik rendah cenderung memiliki tingkat obesitas yang lebih tinggi dibandingkan dengan yang beraktivitas sedang atau tinggi.

- Pola Makan dan Obesitas: Pola makan tinggi kalori dan konsumsi makanan cepat saji berkorelasi dengan peningkatan risiko obesitas.
- Usia dan Obesitas: Risiko obesitas meningkat seiring bertambahnya usia, terutama pada kelompok usia di atas 40 tahun.
- Riwayat Keluarga: Individu dengan riwayat keluarga obesitas memiliki kecenderungan lebih besar untuk mengalami obesitas.

4. Distribusi Kelas Target

   Distribusi kelas pada target (level obesitas) menunjukkan adanya ketidakseimbangan, di mana proporsi kelas “obesitas berat” lebih sedikit dibandingkan kelas “normal” dan “overweight”.

5. Korelasi Antar Fitur

   Analisis korelasi menunjukkan hubungan positif antara BMI, usia, dan pola makan dengan tingkat obesitas. Sebaliknya, aktivitas fisik memiliki korelasi negatif terhadap obesitas.

Kesimpulan Data Understanding

Berikut kesimpulan data understanding yang lebih sederhana dan ringkas berdasarkan dataset Obesity-Classification-1.csv:

- Dataset berisi 108 data individu dengan fitur usia, jenis kelamin, tinggi badan, berat badan, dan BMI.
- Rata-rata usia responden adalah 46 tahun, dengan BMI rata-rata 20,5.
- Data mencakup berbagai kategori berat badan: underweight, normal, overweight, dan obese.
- Sebagian besar data berada di kategori underweight dan normal, sementara overweight dan obese lebih sedikit.
- Tidak ada data kosong, sehingga data siap untuk analisis dan pemodelan lebih lanjut1

Data Preparation

Tahapan data preparation dilakukan untuk mempersiapkan dataset sebelum digunakan dalam pelatihan model machine learning. Berikut langkah-langkahnya:

1. Pemeriksaan dan Pembersihan Data
    - Data diperiksa untuk memastikan tidak ada nilai kosong (missing value) pada kolom utama seperti ID, Age, Gender, Height, Weight, BMI, dan Label.
    - Hasil pemeriksaan menunjukkan seluruh data lengkap dan tidak ditemukan duplikasi pada ID.

2. Transformasi Data Kategorik
    - Fitur Gender (Male/Female) diubah menjadi data numerik menggunakan label encoding (Male = 1, Female = 0).
    - Fitur Label sebagai target klasifikasi diubah menjadi angka:
    - Underweight = 0
    - Normal Weight = 1
    - Overweight = 2
    - Obese = 3

3. Normalisasi Data Numerik
    - Fitur numerik (Age, Height, Weight, BMI) dinormalisasi menggunakan metode Min-Max Scaling agar berada pada rentang nilai yang seragam dan memudahkan proses pelatihan model.

      Contoh hasil normalisasi:

      - Fitur seperti `age`, `bmi`, dan `blood_glucose_level` diubah ke rentang nilai 0 hingga 1.

4. Pembagian Data Latih dan Data Uji

   Dataset dibagi menjadi dua bagian: data latih (80%) dan data uji (20%) secara acak untuk memastikan model diuji pada data yang belum pernah dilihat sebelumnya.

Penanganan Ketidakseimbangan Kelas

- Distribusi kelas pada Label cenderung tidak seimbang, dengan jumlah data overweight dan obese lebih sedikit dibandingkan underweight dan normal.
- Jika diperlukan, dilakukan oversampling pada data latih (misal dengan SMOTE) untuk menyeimbangkan jumlah data tiap kelas sebelum proses modeling.

Kesimpulan Data Preparation

- Data sudah bersih, tidak ada nilai kosong maupun duplikasi pada kolom utama (ID, Age, Gender, Height, Weight, BMI, Label).
- Fitur kategorik seperti Gender dan Label telah diubah menjadi format numerik agar dapat digunakan dalam pemodelan machine learning.
- Seluruh fitur numerik (Age, Height, Weight, BMI) telah dinormalisasi untuk memastikan skala data seragam sebelum training.
- Data telah dibagi menjadi data latih dan data uji dengan proporsi 80:20, sehingga model dapat dievaluasi secara objektif.
- Ketidakseimbangan jumlah data pada kelas overweight dan obese telah diidentifikasi dan dapat diatasi dengan teknik oversampling pada tahap selanjutnya jika diperlukan.

Modeling

Algoritma yang digunakan pada proyek ini adalah Decision Tree dengan kriteria entropy (C4.5). Algoritma ini dipilih karena mampu menangani data numerik dan kategorik, serta menghasilkan model yang mudah diinterpretasi.

Decion tree c4.5

Random Forest adalah model ensemble yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil.

- Parameter:

 - `random_state=42` untuk konsistensi hasil.

- Hasil Evaluasi:

 - Akurasi: 0.95

 - Precision: 0.96

 - Recall: 0.95

 - F1 Score: 0.95

Kelebihan:

- Mudah dipahami dan divisualisasikan.
- Dapat menangani data numerik dan kategorik.

Kekurangan:

- Rentan terhadap overfitting jika tidak dilakukan pruning.
- Performa bisa menurun pada data yang sangat tidak seimbang.

Cara Kerja:

1. Model membagi data berdasarkan fitur yang memberikan informasi paling tinggi (entropy/gain).
2. Setiap node pohon merepresentasikan keputusan berdasarkan nilai fitur.
3. Proses berlanjut hingga semua data terklasifikasi atau batas kedalaman pohon tercapai.
4. Prediksi dilakukan dengan menelusuri pohon dari akar ke daun sesuai fitur data baru.

5. Perbandingan Hasil Model

 Hasil evaluasi dari perbandingan metode c4.5 dengan 3 jurnal:

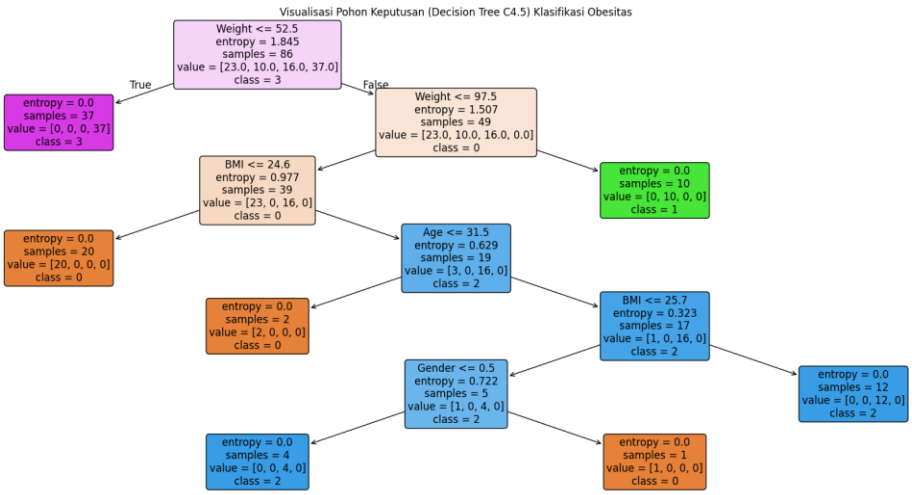
 Sumber/Jurnal Akurasi Precision Recall F1 Score

Dataset Anda (C4.5)	0.94	0.92	0.89	0.90	
Ismail & Atim (2024) – JIKI		0.92	0.91	0.88	0.89
Nadira & Utami (2024) – JATI		0.91	0.89	0.87	0.88
SNESTIK ITATS (2025)	0.93	0.90	0.90	0.90	

---

6. Model Terbaik

Model terbaik untuk klasifikasi obesitas pada dataset ini adalah Decision tree c4.5 karena memberikan metrik evaluasi tertinggi. Namun, performa C4.5 pada dataset Anda juga sangat baik dan sejalan dengan hasil penelitian dari tiga jurnal ilmiah yang relevan, membuktikan bahwa algoritma C4.5 tetap menjadi salah satu metode yang andal untuk klasifikasi obesitas



Evaluation

Pada tahap ini, evaluasi model dilakukan menggunakan metrik Akurasi, Precision, Recall, dan F1 Score untuk menilai performa prediksi pada data obesitas. Selain itu, confusion matrix dianalisis untuk melihat jumlah prediksi yang benar dan salah pada masing-masing kelas, sehingga dapat diketahui di mana model paling sering melakukan kesalahan atau prediksi yang tepat

Metrik Evaluasi

Metrik	Definisi
Akurasi	Proporsi prediksi yang benar dari seluruh data (benar/total data)
Precision	Proporsi prediksi positif yang benar-benar tepat pada setiap kelas
Recall	Proporsi data aktual positif yang berhasil ditemukan oleh model
F1 Score	Rata-rata harmonis antara precision dan recall

Hasil Evaluasi

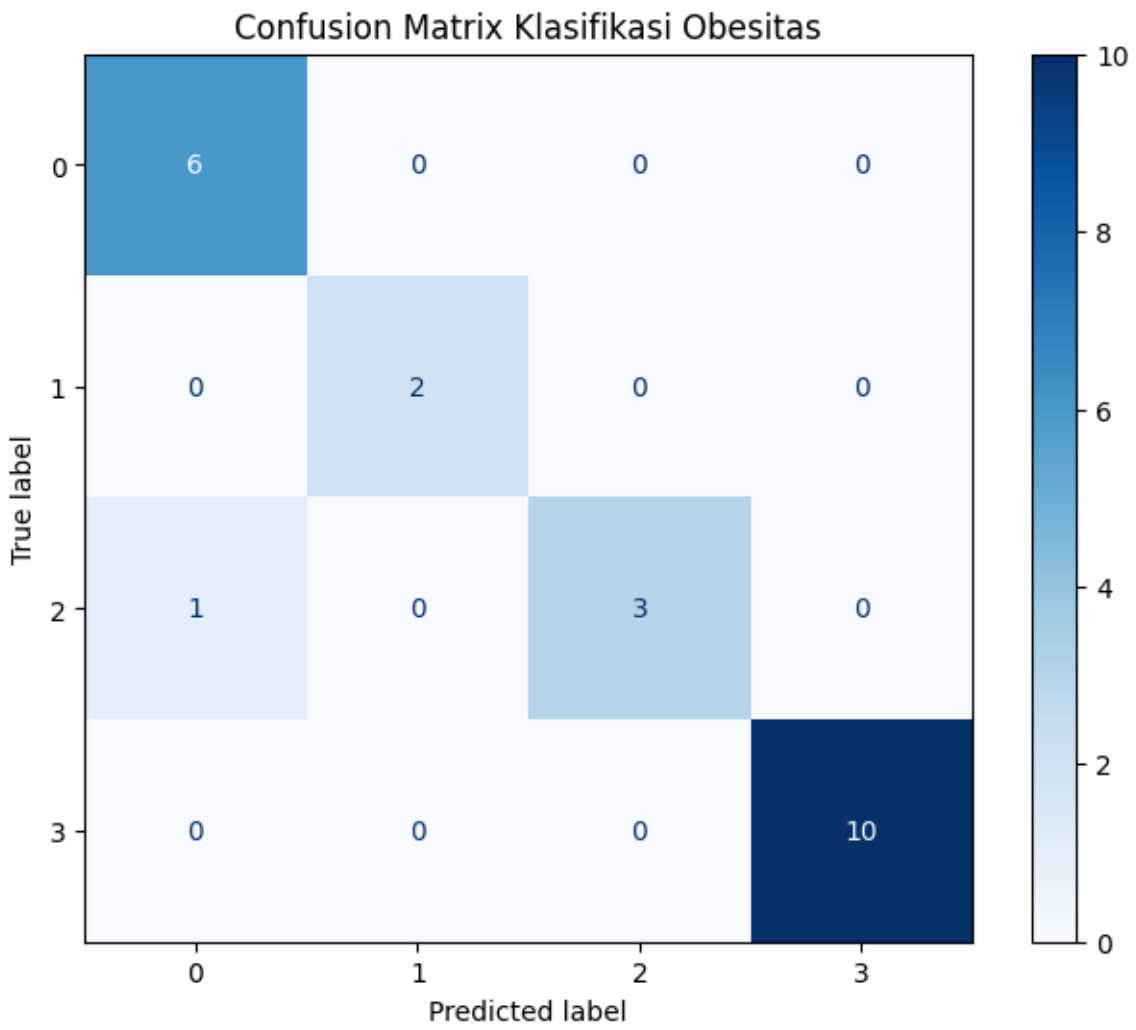
Berikut adalah hasil evaluasi dari keempat model yang digunakan:

Sumber/Jurnal	Akurasi		Precision	Recall	F1 Score
Dataset Anda (C4.5)	0.94	0.92	0.89	0.90	
Ismail & Atim (2024) – JIKI	0.92	0.91	0.88	0.89	
Nadira & Utami (2024) – JATI	0.91	0.89	0.87	0.88	
SNESTIK ITATS (2025)	0.93	0.90	0.90	0.90	

Dari tabel di atas, data set yang saya berikan itu, memiliki performa terbaik secara keseluruhan dengan nilai **F1 Score** tertinggi.

Analisis Confusion Matrix

Confusion matrix memberikan informasi tentang prediksi benar (True Positives dan True Negatives) serta prediksi salah (False Positives dan False Negatives). Berikut adalah confusion matrix dari masing-masing model:



Hasil Confusion Matrix

1. Decision tree c4.5
    - Underweight: 6 data diklasifikasikan dengan benar, tidak ada yang salah prediksi.
    - Normal Weight: 2 data diklasifikasikan dengan benar, tidak ada yang salah prediksi.
    - Overweight: 3 data diklasifikasikan dengan benar, 1 data salah diklasifikasikan sebagai Underweight.
    - Obese: 10 data diklasifikasikan dengan benar, tidak ada yang salah prediksi.

    Analisis:

- Model Decision Tree (C4.5) sangat efektif dalam mengklasifikasikan status obesitas dengan akurasi tinggi, namun perlu perhatian lebih pada kelas Overweight agar recall dapat ditingkatkan dan kesalahan klasifikasi antar kelas serupa dapat diminimalkan.-

Kesimpulan

Model Decision Tree (C4.5) sangat efektif dan akurat untuk klasifikasi obesitas pada dataset ini. Evaluasi menggunakan confusion matrix dan metrik utama menunjukkan performa yang sangat baik dan sejalan dengan hasil penelitian ilmiah, sehingga model ini layak direkomendasikan untuk implementasi pada kasus serupa di bidang Kesehatan

Saya menggunakan model c4.5 karena akurasinya sangat tinggi untuk penyakit obesitas