PROPOSAL RISET INFORMATIKA

“Penerapan Metode XGBoost dengan Optimasi SMOTE untuk Klasifikasi Tingkat Obesitas Berdasarkan Gaya Hidup”

Disusun Oleh:
Chesa Saskia Rafika / 22081010211

PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”
JAWA TIMUR

2025

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Obesitas merupakan salah satu permasalahan kesehatan global yang terus meningkat dari tahun ke tahun. Berdasarkan data World Health Organization (WHO), prevalensi obesitas telah meningkat lebih dari tiga kali lipat sejak tahun 1975 dan menjadi salah satu penyebab utama penyakit tidak menular seperti diabetes tipe 2, hipertensi, serta penyakit jantung (WHO, 2024). Di Indonesia, hasil Riset Kesehatan Dasar (Riskesdas) menunjukkan peningkatan angka obesitas baik pada laki-laki maupun perempuan dewasa setiap tahunnya (Kemenkes RI, 2023). Kondisi ini menunjukkan pentingnya upaya deteksi dini terhadap risiko obesitas untuk mencegah komplikasi lebih lanjut. Gaya hidup memiliki peran besar terhadap tingkat obesitas, termasuk pola makan, aktivitas fisik, durasi tidur, dan kebiasaan merokok. Identifikasi hubungan antara gaya hidup dengan tingkat obesitas dapat membantu dalam pengambilan keputusan medis dan edukasi masyarakat. Namun, proses identifikasi secara manual sulit dilakukan ketika jumlah data sangat besar dan bersifat kompleks.

Perkembangan teknologi *machine learning* memberikan solusi efektif untuk menganalisis pola dan memprediksi tingkat obesitas berdasarkan data gaya hidup. Salah satu algoritma yang banyak digunakan karena performanya yang tinggi adalah Extreme Gradient Boosting (XGBoost). Algoritma ini mampu menangani data dengan dimensi besar dan kompleksitas tinggi serta memiliki kemampuan generalisasi yang baik.

Namun, permasalahan utama dalam klasifikasi obesitas adalah ketidakseimbangan kelas (imbalanced data) misalnya, jumlah individu dengan obesitas berat jauh lebih sedikit dibandingkan kategori normal. Ketidakseimbangan ini dapat menurunkan akurasi model karena algoritma cenderung bias terhadap kelas mayoritas. Untuk mengatasi hal ini, digunakan metode SMOTE (Synthetic Minority Oversampling Technique) yang bekerja dengan menambah sampel sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang. Dengan demikian, penelitian ini berfokus pada penerapan metode XGBoost

dengan optimasi SMOTE untuk meningkatkan performa klasifikasi tingkat obesitas berdasarkan gaya hidup. Diharapkan, hasil penelitian ini dapat menjadi kontribusi dalam bidang kesehatan prediktif berbasis *machine learning* serta memberikan insight terhadap faktor gaya hidup yang paling berpengaruh terhadap obesitas.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana penerapan metode XGBoost untuk melakukan klasifikasi tingkat obesitas berdasarkan faktor gaya hidup?
2. Bagaimana pengaruh penerapan SMOTE terhadap peningkatan performa model klasifikasi obesitas?
3. Seberapa baik akurasi, presisi, dan *recall* yang dihasilkan oleh model XGBoost dengan optimasi SMOTE dibandingkan dengan tanpa SMOTE?

1.3 Identifikasi Masalah

Dari uraian di atas, dapat diidentifikasi beberapa permasalahan utama, yaitu:

1. Belum diketahui seberapa besar pengaruh setiap faktor gaya hidup terhadap tingkat obesitas.
2. Diperlukan model machine learning yang mampu menangani data tidak seimbang dan memiliki performa tinggi dalam klasifikasi.
3. Ketidakseimbangan jumlah data antar kelas tingkat obesitas menyebabkan hasil klasifikasi kurang optimal.
4. Belum ada penelitian yang secara khusus menerapkan kombinasi XGBoost dan SMOTE untuk kasus klasifikasi obesitas berbasis gaya hidup di konteks data Indonesia.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menerapkan metode XGBoost untuk klasifikasi tingkat obesitas berdasarkan data gaya hidup individu.
2. Mengoptimalkan hasil klasifikasi dengan menggunakan teknik SMOTE untuk menyeimbangkan distribusi data.
3. Menganalisis performa model dengan membandingkan hasil sebelum dan sesudah penerapan SMOTE.

BAB 2

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian mengenai klasifikasi tingkat obesitas dengan machine learning telah banyak dilakukan, terutama menggunakan data gaya hidup sebagai faktor utama. Salah satu penelitian berjudul *"Optimasi AdaBoost dan XGBoost untuk Klasifikasi Obesitas Menggunakan SMOTE"* memanfaatkan dataset publik dari Kaggle yang berisi 2.111 data individu dengan 17 atribut gaya hidup seperti usia, berat badan, pola makan, aktivitas fisik, dan kebiasaan merokok. Hasil penelitian tersebut menunjukkan bahwa metode XGBoost memberikan performa paling baik dibandingkan AdaBoost, dengan akurasi meningkat dari 92,4% menjadi 94,5% setelah penerapan teknik SMOTE untuk menyeimbangkan data. Fitur yang paling berpengaruh terhadap hasil klasifikasi adalah berat badan dan usia (Chen & Guestrin, 2016; Chawla et al., 2002).

Penelitian lain berjudul *"Enhancing Obesity Prediction through SMOTE-based Classification Models: A Comparative Study"* juga menggunakan dataset yang sama dan membandingkan beberapa algoritma seperti Decision Tree, Random Forest, AdaBoost, dan XGBoost. Hasilnya menunjukkan bahwa XGBoost dengan penerapan SMOTE memperoleh akurasi tertinggi sebesar 98,6%, melampaui algoritma lainnya. Dari hasil tersebut dapat disimpulkan bahwa XGBoost merupakan metode paling unggul untuk klasifikasi tingkat obesitas berbasis gaya hidup, terutama ketika dikombinasikan dengan teknik optimasi SMOTE. Namun, sebagian penelitian sebelumnya masih berfokus pada perbandingan algoritma tanpa analisis mendalam terhadap parameter XGBoost, sehingga penelitian ini mencoba mengoptimalkan metode tersebut untuk meningkatkan akurasi prediksi obesitas.

2.2 Landasan Teori

1. Obesitas berdasarkan faktor gaya hidup

   Obesitas merupakan kondisi medis ketika terjadi penumpukan lemak tubuh secara berlebihan yang dapat meningkatkan risiko berbagai penyakit kronis seperti diabetes, hipertensi, dan penyakit jantung. Faktor utama penyebab obesitas tidak hanya bersifat genetik, tetapi juga dipengaruhi oleh gaya hidup individu. Faktor-faktor tersebut meliputi pola makan tinggi kalori, kurangnya aktivitas fisik, konsumsi alkohol, kebiasaan merokok, serta durasi penggunaan teknologi yang tinggi. Perubahan pola hidup modern yang ditandai dengan aktivitas sedentari dan konsumsi makanan cepat saji turut mempercepat peningkatan angka obesitas di berbagai negara. Oleh karena itu, analisis hubungan antara gaya hidup dan tingkat obesitas penting dilakukan untuk mendukung upaya pencegahan dan pengendalian obesitas melalui pendekatan data dan teknologi.

2. Penjelasan algoritma XGBoost

   Extreme Gradient Boosting (XGBoost) adalah salah satu algoritma *ensemble learning* berbasis *gradient boosting* yang dikembangkan oleh Chen dan Guestrin. Algoritma ini bekerja dengan membangun sejumlah *decision tree* secara bertahap, di mana setiap pohon baru berusaha memperbaiki kesalahan prediksi dari pohon sebelumnya. Keunggulan XGBoost terletak pada efisiensi komputasi, kemampuan menangani data yang tidak seimbang, serta penerapan regularisasi untuk menghindari *overfitting*. Selain itu, XGBoost memiliki parameter optimasi seperti *learning rate*, *max depth*, dan *subsample*, yang dapat disesuaikan untuk meningkatkan performa model dalam klasifikasi obesitas berbasis data gaya hidup.

3. Penjelasan metode SMOTE

   SMOTE adalah teknik *oversampling* yang digunakan untuk menyeimbangkan distribusi kelas pada dataset yang tidak seimbang. Metode ini bekerja dengan membuat data sintetis baru pada kelas minoritas melalui interpolasi antara sampel yang berdekatan dalam ruang fitur. Dengan demikian, model tidak akan bias terhadap kelas mayoritas dan

dapat menghasilkan prediksi yang lebih akurat. Dalam konteks klasifikasi tingkat obesitas, penerapan SMOTE membantu algoritma XGBoost mengenali pola pada kategori obesitas yang jumlah datanya relatif sedikit.

4. Konsep evaluasi model (akurasi, precision, recall, f1-score).

Evaluasi model merupakan tahap penting untuk mengukur kinerja algoritma klasifikasi. Beberapa metrik yang umum digunakan yaitu:

- Akurasi, yaitu rasio jumlah prediksi benar terhadap total data yang diuji.
- Precision, yaitu tingkat ketepatan model dalam memprediksi kelas positif yang benar.
- Recall, yaitu kemampuan model mendeteksi seluruh data positif secara benar.
- F1-Score, yaitu rata-rata harmonik antara precision dan recall yang memberikan gambaran keseimbangan antara keduanya.

BAB 3

METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini bersifat eksperimental dengan pendekatan kuantitatif, karena melibatkan proses pengujian model algoritma *machine learning* untuk memperoleh hasil berbasis data numerik. Pendekatan ini digunakan untuk menganalisis hubungan antara faktor gaya hidup dan tingkat obesitas melalui penerapan algoritma XGBoost dengan optimasi SMOTE. Tujuan dari penelitian ini adalah untuk memperoleh model klasifikasi yang akurat dan stabil dalam memprediksi tingkat obesitas berdasarkan pola hidup individu.

3.2 Sumber Data

Data yang digunakan dalam penelitian ini berasal dari dataset publik berjudul "*Obesity Levels Dataset*", yang tersedia di platform Kaggle. Dataset ini juga dilampirkan dalam bentuk file Excel dengan nama "Obesity_Dataset.xlsx" sebagai lampiran penelitian. Dataset tersebut terdiri atas 2.111 baris data dan 17 atribut yang menggambarkan berbagai faktor gaya hidup, pola makan, serta kebiasaan individu yang memengaruhi tingkat obesitas.

Beberapa atribut yang digunakan antara lain: Gender, Age, Height, Weight, FAVC (frekuensi konsumsi makanan berlemak), FCVC (konsumsi sayur/buah), NCP (jumlah makanan per hari), CAEC (konsumsi kalori tambahan), SMOKE (kebiasaan merokok), CH2O (konsumsi air putih), FAF (aktivitas fisik), TUE (penggunaan teknologi), CALC (konsumsi alkohol), SCC (pemantauan kalori), dan MTRANS (moda transportasi). Atribut target yang akan diklasifikasikan adalah NObeyesdad, yaitu tingkat obesitas yang terbagi menjadi beberapa kategori seperti *Insufficient Weight, Normal Weight, Overweight Level I & II, Obesity Type I–III*.

3.3 Tahapan Penelitian

Tahapan penelitian ini disusun secara sistematis agar proses klasifikasi tingkat obesitas dapat dilakukan secara optimal. Adapun langkah-langkah penelitian meliputi:

1. Pengumpulan Data
   Mengambil dataset *Obesity Levels Dataset* dari sumber resmi (Kaggle) yang berisi data gaya hidup dan tingkat obesitas individu.
2. Pra-pemrosesan (data cleaning, encoding, normalisasi)
   Melakukan *data cleaning* untuk menghapus data kosong atau duplikat, kemudian mengubah data kategorikal menjadi numerik melalui proses *encoding*, serta menormalkan skala data dengan teknik *normalization* agar model tidak bias terhadap variabel tertentu.
3. Penyeimbangan data dengan SMOTE
   Menerapkan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk menyeimbangkan agar model tidak condong terhadap kelas mayoritas.
4. Pembagian Data (Train-Test Split)
   Dataset dibagi menjadi dua bagian, yaitu 80% data latih dan 20% data uji untuk mengukur performa model secara objektif.
5. Penerapan metode XGBoost
   Membangun model klasifikasi menggunakan algoritma Extreme Gradient Boosting (XGBoost), kemudian melakukan tuning parameter untuk mendapatkan hasil terbaik.

3.4 Evaluasi Hasil

Evaluasi model dilakukan menggunakan Confusion Matrix dan empat metrik utama, yaitu accuracy, precision, recall, dan F1-score.

- Accuracy digunakan untuk melihat persentase prediksi benar terhadap seluruh data uji.
- Precision menunjukkan tingkat ketepatan model dalam mengklasifikasikan data dengan benar.
- Recall menunjukkan kemampuan model untuk mendeteksi seluruh data dalam suatu kelas.
- F1-score menggambarkan keseimbangan antara precision dan recall.