

# Pemodelan Prediksi Tingkat Obesitas Menggunakan Regresi Logistik: Studi Kasus pada Atribut Kesehatan dalam Populasi Tertentu

Amalia Najwannisa  
Fakultas Informatika  
Telkom University  
Bandung, Indonesia

amalianajwa@student.telkomuniversity.ac.id

Ammardito Syafa'at  
Fakultas Informatika  
Telkom University  
Bandung, Indonesia

ammarditos@student.telkomuniversity.ac.id

Aqeela Fathya Najwa  
Fakultas Informatika  
Telkom University  
Bandung, Indonesia

aqeelafn@student.telkomuniversity.ac.id

**Abstract**—Penelitian ini bertujuan untuk memprediksi risiko obesitas pada individu. Dataset yang digunakan terdiri dari 2111 baris data dan 17 kolom yang merupakan faktor-faktor yang dapat mempengaruhi prediksi terhadap obesitas. Metode regresi logistik dipilih untuk menjadi metode dalam penelitian ini dikarenakan relevansinya dalam memprediksi risiko obesitas. Model ini dapat berkontribusi signifikan dalam memberikan prediksi risiko yang akurat berdasarkan faktor-faktor yang tersedia.

**Keywords**—Obesitas, Regresi Logistik, Prediksi

## I. LATAR BELAKANG

Kesadaran masyarakat yang masih minim terhadap pola gaya hidup sehari-hari dapat berbahaya bagi munculnya risiko obesitas pada individu. Kebiasaan makan yang tidak teratur dan tidak sehat juga kurangnya aktivitas fisik yang dilakukan individu merupakan salah satu faktor terbesar yang menyumbangkan dalam perkembangan obesitas.

Dalam upaya untuk memprediksi kasus obesitas, informasi mengenai jenis kelamin, umur, tinggi badan, berat badan, riwayat keluarga penyandang obesitas, frekuensi konsumsi makanan berkalori tinggi, frekuensi mengonsumsi sayur, jumlah makanan utama per hari, camilan diantara waktu makan, merokok, jumlah konsumsi air per hari, pemantauan kalori, frekuensi aktivitas fisik, waktu menggunakan gadget, konsumsi alkohol, dan jenis transportasi yang digunakan sehari-hari sangat diperlukan untuk dapat menghasilkan prediksi yang akurat.

Penggunaan metode regresi logistik dinilai menjadi pilihan metode yang lebih relasional dan relevan dibanding dengan metode lainnya. Dengan menggunakan metode regresi logistik untuk penelitian ini dapat memberikan prediksi kemungkinan obesitas pada tingkat individu.

## II. STUDI LITERATUR

### A. Obesitas

Obesitas atau overnutrisi telah menjadi masalah yang mengganggu di seluruh dunia [1]. Obesitas merupakan penyakit multifaktorial yang ditandai oleh akumulasi jaringan lemak yang berlebihan dan dapat mengganggu kesehatan [2]. Obesitas merupakan kelainan kompleks dalam pengaturan nafsu makan dan metabolisme energi yang dikendalikan oleh beberapa faktor biologis spesifik [3].

### B. Regresi Logistik

Regresi Logistik adalah suatu metode analisis statistika yang digunakan untuk menjelaskan hubungan antara variabel terikat yang memiliki dua kategori atau lebih dengan satu atau lebih variabel bebas yang berskala kategori atau kontinu [4]. Regresi logistik bekerja sangat mirip dengan regresi linier, tetapi dengan variabel respons binomial [5].

Metode regresi logistik ini dapat menjadi metode yang sangat berguna di berbagai bidang, misalnya untuk memprediksi kejadian biner seperti ada atau tidaknya suatu penyakit. Dengan menggunakan regresi logistik biner, peneliti dapat memahami faktor-faktor yang mempengaruhi kemungkinan keberadaan suatu kejadian dan membuat prediksi tentang kemungkinan kejadian tersebut berdasarkan variabel prediktor yang tersedia.

## III. DATA DAN METODE PENELITIAN

### A. Dataset

Dataset yang digunakan merupakan dataset yang berasal dari website Kaggle yang terdiri dari 2111 baris dengan 17 kolom berbeda. Pada setiap baris dilabel dengan suatu variabel kelas "nobeyesdad" yang merupakan tingkat obesitas dengan mengklasifikasikan ke dalam 7 kelompok/tingkat: "Insufficient Weight", "Normal Weight", "Overweight Level I", "Overweight Level II", "Obesity Type I", "Obesity Type II", dan "Obesity Type III".

Data diperoleh secara langsung dari pengguna melalui platform web sebesar 23% dan sisanya 77% dibuat secara sintesis dengan menggunakan alat Weka dan filter SMOTE.

### B. Tahapan Penelitian

Dalam tahapan penelitian dapat menggambarkan bagaimana prediksi obesitas dengan metode regresi logistik dilakukan. Tahapan-tahapan tersebut antara lain:

#### 1. Pemahaman Masalah

Dalam hal ini, masalah yang ada adalah memprediksi kemungkinan seseorang mengalami obesitas berdasarkan beberapa faktor risiko yang ada.

#### 2. Pengumpulan Data

Pengumpulan data tersebut bisa mencakup berbagai variabel, seperti usia, jenis kelamin, tinggi badan, berat badan, riwayat penyakit keluarga, pola makan, tingkat aktivitas fisik, dan lain-lain.

3. **Pembersihan dan Persiapan Data**  
Data yang telah dikumpulkan kemungkinan besar akan memerlukan pembersihan dan persiapan sebelum digunakan untuk analisis. Tahapan ini meliputi menanganai nilai yang hilang, menghilangkan data yang tidak relevan atau tidak valid, dan melakukan transformasi data.
4. **Eksplorasi Data**  
Tujuan dari eksplorasi data ini adalah untuk memahami hubungan antara variabel-variabel yang ada dan melihat pola-pola yang mungkin ada dalam data.
5. **Pemodelan Regresi Logistik**  
Tahapan ini melibatkan pembuatan model regresi logistik. Model ini akan mengambil variabel-variabel prediksi, yaitu: jenis kelamin, umur, tinggi badan, berat badan, riwayat keluarga penyandang obesitas, frekuensi konsumsi makanan berkalori tinggi, frekuensi mengonsumsi sayur, jumlah makanan utama per hari, camilan diantara waktu makan, merokok, jumlah konsumsi air per hari, pemantauan kalori, frekuensi aktivitas fisik, waktu menggunakan gadget, konsumsi alkohol, dan jenis transportasi yang digunakan sehari-hari dan menggunakan mereka untuk memprediksi kemungkinan obesitas. Proses ini melibatkan penggunaan algoritma regresi logistik untuk menyesuaikan model dengan data.
6. **Validasi Model**  
Setelah model dibangun, penting untuk menguji seberapa baik model tersebut memprediksi obesitas. Validasi model dapat dilakukan dengan membagi data menjadi subset pelatihan dan pengujian dan kemudian menguji kinerja model pada subset pengujian. Evaluasi model yang digunakan adalah akurasi dan confusion matrix.
7. **Implementasi dan Penyajian Hasil**  
Setelah model dikonstruksi dan divalidasi, langkah terakhir adalah mengimplementasikan model ke dalam aplikasi atau sistem yang sesuai. Hasil dari model disajikan dalam bentuk web aplikasi yang dibangun melalui Streamlit.

### C. Regresi Logistik

Regresi logistik adalah jenis model klasifikasi parametrik yang digunakan ketika variabel responsnya adalah tipe kategorikal. Ide dasar di balik regresi logistik adalah untuk menemukan hubungan antara fitur dan probabilitas hasil tertentu. Dalam regresi logistik, fungsi sigmoid digunakan untuk memetakan prediksi ke probabilitas.

Persamaan umum yang menyatakan model regresi logistik adalah sebagai berikut:

$$p(X) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

di mana  $p(X)$  adalah output yang diprediksi,  $b_0$  adalah istilah intersep dan  $b_1$  adalah koefisien untuk nilai input tunggal ( $x$ ) [6].

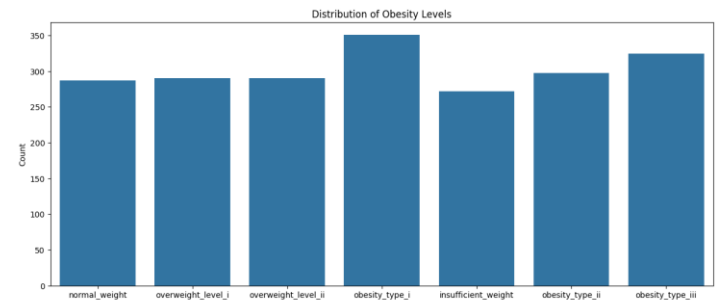
## IV. HASIL DAN PEMBAHASAN

Berdasarkan pada data mengenai obesitas yang telah dikumpulkan, penting bagi data tersebut untuk melalui

pengecekan mengenai *missing value*, hasil pengecekan tersebut dapat dilihat pada tabel berikut:

atribut	jumlah missing value
id	0
gender	0
age	0
height	0
weight	0
family_history_with_overweight	0
favc	0
fcvc	0
ncp	0
caec	0
smoke	0
ch2o	0
scc	0
faf	0
tue	0
calc	0
mtrans	0
nobeyesdad	0
bmi	0

Berdasarkan dari tabel di atas, dapat terlihat bahwa dataset yang tersedia tidak memiliki *missing value* di dalamnya. Data kemudian melalui proses eksplorasi data dengan tujuan ingin melihat apakah data target yang dipilih, yaitu data pada kolom “nobeyesdad”, memiliki ketimpangan jumlah kemunculan data per levelnya. Hasil dari eksplorasi tersebut dapat dilihat pada histogram “Distribution of Obesity Level” di bawah ini:



Dari histogram yang ditampilkan di atas, dapat dikatakan bahwa setiap level obesitas dalam dataset memiliki jumlah kemunculan dengan selisih yang tidak beda jauh tiap levelnya atau dapat dikatakan tidak ada ketimpangan antara jumlah per level obesitas yang muncul.

Dikarenakan tidak adanya selisih yang signifikan berdasarkan jumlah kemunculan diantara setiap level, maka dapat langsung dilakukan pemilihan fitur yang dimasukkan menjadi data  $x$  dan data target untuk proses prediksi yang akan dilakukan. Dari dataset yang tersedia, data  $x$  merupakan kolom jenis kelamin, umur, tinggi badan, berat badan, riwayat keluarga penyandang obesitas, frekuensi konsumsi makanan berkalori tinggi, frekuensi mengonsumsi sayur, jumlah makanan utama per hari, camilan diantara waktu makan, merokok, jumlah konsumsi air per hari, pemantauan kalori, frekuensi aktivitas fisik, waktu menggunakan gadget, konsumsi alkohol, dan jenis transportasi yang digunakan sehari-hari, sedangkan targetnya merupakan kolom “nobeyesdad”.

Setelah adanya data x dan target, dilakukan pembagian data, menjadi data pelatihan dan data pengujian. Pembagian ini menggunakan rasio 80:20, dimana 80% merupakan data pelatihan sedangkan 20% merupakan data pengujian.

Model regresi logistik akan mengolah data pelatihan dan data pengujian tersebut hingga dapat menghasilkan suatu prediksi tentang risiko obesitas yang terjadi. Hasil prediksi dengan model regresi logistik ini dapat dilihat keakuratannya melalui *confusion matrix* dan nilai akurasi. Visualisasi dari *confusion matrix* yang tercipta dari hasil prediksi yang telah dilakukan yaitu seperti di bawah ini:

True Label	Insufficient Weight	54	2	0	0	0	0	0
	Normal Weight	4	43	0	0	0	13	2
	Obesity Type 1	0	0	62	8	3	0	5
	Obesity Type 2	0	0	1	57	0	0	0
	Obesity Type 3	0	0	0	0	63	0	0
	Overweight Level 1	0	6	1	0	0	41	8
	Overweight Level 2	0	0	2	0	0	7	41
		Insufficient Weight	Normal Weight	Obesity Type 1	Obesity Type 2	Obesity Type 3	Overweight Level 1	Overweight Level 2
		Predicted Label						

Dari *confusion matrix* di atas, dapat terlihat bahwa dalam diagram utama matriks merupakan nilai jumlah prediksi yang tepat sasaran sesuai labelnya, misalnya: label "Insufficient Weight" memiliki nilai 54 di diagonal utamanya yang berarti ada 54 prediksi tentang label tersebut yang benar dan sesuai dengan label aslinya. Untuk angka di luar diagonal utama merupakan jumlah prediksi yang salah atau tidak sesuai dengan label aslinya. Model yang baik akan memiliki angka yang tinggi di diagonal utama dan angka kecil di luar diagonal utama.

Selain dengan *confusion matrix*, keakuratan model juga diuji dengan fungsi *accuracy*. Dari model yang telah diterapkan, didapatkan hasil akurasi yang dapat dilihat pada tabel di bawah:

	precision	recall	f1-score	support
insufficient_weight	0.92	0.96	0.94	56
normal_weight	0.87	0.66	0.75	62
obesity_type_i	0.94	0.77	0.85	78
obesity_type_ii	0.86	0.98	0.92	58
obesity_type_iii	0.95	1.00	0.98	63
overweight_level_i	0.67	0.73	0.70	56
overweight_level_ii	0.70	0.84	0.76	50
<b>accuracy</b>			<b>0.85</b>	423
macro avg	0.85	0.85	0.84	423
weighted avg	0.85	0.85	0.85	423

Dari tabel di atas dapat terlihat bahwa perolehan akurasi dari model yang digunakan sebesar 0.85 atau 85%. Besar persentase ini menunjukkan bahwa model memiliki performa yang baik dalam memprediksi tingkat obesitas seseorang berdasarkan dari atribut yang telah dipilih sebelumnya.

## V. KESIMPULAN

Penelitian yang bertujuan untuk memprediksi risiko obesitas pada individu ini menggunakan metode regresi logistik berdasarkan beberapa faktor risiko yang ada untuk memprediksi kemungkinannya. Faktor-faktor tersebut diantaranya jenis kelamin, usia, tinggi badan, berat badan, riwayat keluarga penyandang obesitas, pola makan, tingkat aktivitas fisik, dan kebiasaan gaya hidup lainnya. Faktor-faktor tersebut diperoleh dari dataset yang diambil dari website Kaggle yang memiliki 2111 baris data dan 17 kolom di dalamnya.

Tahapan dari penelitian ini mencakup pemahaman masalah, pengumpulan data, pembersihan dan persiapan data, eksplorasi data, pemodelan regresi logistik, validasi model, dan implementasi dan penyajian hasil.

Model regresi logistik yang dibangun dalam penelitian ini memiliki persentase akurasi sebesar 0.85 atau 85%. Persentase ini dapat menunjukkan bahwa model memiliki performa yang baik dalam memprediksi level obesitas seseorang berdasarkan faktor yang ada. Untuk validasi model dari penelitian ini juga menggunakan *confusion matrix* yang menunjukkan bahwa model memiliki kemampuan yang cukup tinggi untuk memprediksi secara akurat level obesitas individu.

Dengan demikian, dapat ditarik Kesimpulan bahwa model regresi logistik menjadi metode atau alat prediksi yang efektif dan relevan untuk memprediksi risiko obesitas pada individu.

## VI. REFERENCES

- [1] M. M. Putra, N. M. Raningsih and N. N. I. Saraswati, "Pola Hidup Dengan Kejadian Obesitas: Literature Review," *Jurnal Ilmu Keperawatan Medikal Bedah*, vol. 5, 2022.
- [2] H. Akbar, "Pemberian Edukasi Mengenai Obesitas pada Remaja di Madrasah Aliyah Negeri 1 Indramayu," *Community Engagement & Emergence Journal*, vol. 2, pp. 1-5, 2021.
- [3] A. F. Syam, I. Alwi, S. Setiati, M. S. K., B. Setiyohadi and A. W. Sudoyo, *Buku Ajar Ilmu Penyakit Dalam Jilid II Edisi VI*, Interna Publishing, 2015.
- [4] D. Hosmer and L. S., *Applied Logistic Regression*, John Wiley and Sons Inc, Canada, 2000.
- [5] S. Sperandei, "Understanding Logistic Regression Analysis," *Biochemia Medica*, 2014.
- [6] D. Saini, T. Chand, D. K. Chouhan and M. Prakash, "A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 419-444, 2021.