BAB II
LANDASAN TEORI

2.1. Tinjauan Teori

2.1.1. Obesitas sebagai Masalah Kesehatan Global

Obesitas adalah kondisi kesehatan kompleks yang dicirikan oleh penumpukan lemak tubuh secara berlebihan sehingga meningkatkan risiko terjadinya berbagai penyakit yang kronis. Organisasi Kesehatan Dunia (WHO) mengklasifikasikan status berat badan menggunakan *Body Mass Index* (BMI) ke dalam tujuh kategori, mulai dari *Insufficient Weight* (BMI < 18.5) hingga *Obesity Type III* (BMI ≥ 40), yang menjadi dasar klasifikasi *multi-class* pada penelitian yang dilakukan oleh C. Boutari dan C. S. Mantzoros [1]. Klasifikasi ini tidak hanya berfungsi sebagai indikator medis, tetapi juga menjadi kerangka kerja untuk pengembangan sistem prediksi berbasis *machine learning.*

Prevalensi obesitas terus meningkat hingga mencapai proporsi epidemik global, dengan Diperkirakan lebih dari satu miliar individu di seluruh dunia hidup dengan obesitas, terdiri dari 650 juta populasi dewasa, 340 juta kelompok remaja, serta 39 juta anak-anak [1]. Huang *et al.* menekankan bahwa proyeksi WHO memprediksi bahwa pada tahun 2025, 177 juta orang akan mengalami obesitas morbid, 1 miliar orang akan obesitas, dan Jumlah orang dewasa yang mengalami kelebihan berat badan diperkirakan mencapai 2,7 miliar [17]. Di Amerika Serikat, biaya kesehatan terkait obesitas diperkirakan mencapai USD 260 miliar per tahun atau sekitar 21% dari total pengeluaran kesehatan nasional [3], sementara di negara berkembang beban ini semakin berat karena keterbatasan fasilitas dan sumber daya kesehatan [4], [5].

Obesitas merupakan hasil interaksi kompleks antara faktor demografis (usia, jenis kelamin), kebiasaan makan (konsumsi makanan berkalori tinggi, frekuensi makan), dan aktivitas fisik (intensitas olahraga, gaya hidup sedentari). Penelitian menunjukkan bahwa interaksi non-linear antar faktor ini, seperti pengaruh usia terhadap metabolisme yang berinteraksi dengan pola aktivitas fisik, atau dampak jenis kelamin terhadap distribusi lemak yang dipengaruhi kebiasaan makan, membuat metode statistik konvensional seperti regresi linear sering kurang mampu menangkap

pola kompleks untuk prediksi yang akurat. Keterbatasan pendekatan tradisional ini menciptakan kebutuhan akan metodologi yang lebih canggih.

2.1.2. Konsep Dasar dan Definisi Obesitas

Machine Learning (ML) adalah bagian dari kecerdasan buatan (AI) yang berfokus pada perancangan algoritma sehingga sistem komputer dapat belajar dari data dan menghasilkan prediksi atau keputusan tanpa perlu diprogram secara khusus untuk tiap proses. Dalam prosesnya, ML menggunakan pendekatan statistik untuk menemukan pola dan hubungan dalam *dataset* sehingga sistem dapat meningkatkan performanya seiring bertambahnya data yang dipelajari.

Mitchell pada penelitiannya [84] memberikan definisi yang paling banyak digunakan dan bersifat formal, yaitu: "*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*." Artinya, suatu model ML dapat dikatakan "belajar" apabila kinerjanya meningkat saat diberikan pengalaman berupa data.

Dalam konteks penelitian prediksi dan klasifikasi obesitas, ML digunakan untuk menganalisis variabel fisik, perilaku, dan nutrisi, kemudian membangun model yang dapat mengklasifikasikan status BMI atau memprediksi risiko obesitas berdasarkan pola yang ditemukan dalam data.

2.1.3. Sistem Klasifikasi dan Pengukuran Obesitas *Body Mass Index* (BMI)

BMI merupakan indikator yang paling sering digunakan untuk mengklasifikasikan status berat badan pada populasi dewasa. BMI dihitung menggunakan rumus:

$$BMI = Berat\ Badan\ (kg) \ / \ [Tinggi\ Badan\ (m)]^2 \tag{2.1}$$

Melalui perhitungan pada formula (2.1), setiap individu dapat dikelompokkan ke dalam kelompok seperti berat badan yang kurang, normal, berat badan yang lebih, atau obesitas sesuai perhitungan berat badan yang dibagi dengan kuadrat tinggi badan, sehingga menghasilkan indikator numerik status berat badan individu. Klasifikasi obesitas berdasarkan BMI menurut WHO adalah:

Tabel 2.1. Klasifikasi Indeks Massa Tubuh (*Body Mass Index*/BMI)

No	Kategori	Rentang BMI (kg/m^2)
1	*Underweight* (Berat Badan Kurang)	< 18,5
2	*Normal Weight* (Berat Badan Normal)	18,5 – 24,9
3	*Overweight Level I*	25,0 – 29,9
4	*Overweight Level II*	30,0 – 34,9
5	*Obesity Type I* (Obesitas Tipe I)	35,0 – 39,9
6	*Obesity Type II* (Obesitas Tipe II)	40,0 – 44,9
7	*Obesity Type III* (Obesitas Tipe III)	≥ 45,0

Berdasarkan Tabel 2.1, klasifikasi status berat badan ditentukan dari nilai *Body Mass Index* (BMI) yang dikelompokkan ke dalam beberapa kategori, mulai dari *underweight* hingga *obesity type III* sesuai dengan rentang nilai BMI yang telah ditetapkan. Pembagian kategori tersebut menunjukkan bahwa semakin tinggi nilai BMI, semakin besar tingkat keparahan obesitas dan potensi risiko kesehatan yang menyertainya. Meskipun BMI merupakan alat skrining yang praktis dan mudah digunakan, indikator ini memiliki beberapa keterbatasan. Romero-Corral *et al.* pada penelitiannya menunjukkan bahwa BMI tidak memberikan informasi spesifik mengenai perbedaan massa otot dan massa lemak, sehingga dapat memberikan klasifikasi yang tidak akurat pada individu dengan massa otot tinggi seperti atlet [7]. Selain itu, BMI tidak memperhitungkan distribusi lemak tubuh yang memiliki implikasi kesehatan yang berbeda.

2.1.4. Epidemiologi Global Obesitas

a. Prevalensi dan Tren Global

Obesitas telah mencapai proporsi epidemi global dengan prevalensi yang meningkat secara konsisten selama kurun waktu beberapa dekade terakhir. Menurut Phelps *et al.*, analisis *pooled* dari 3.663 studi representatif populasi dengan 222 juta anak, remaja, dan dewasa menunjukkan tren peningkatan yang mengkhawatirkan dalam prevalensi obesitas di seluruh dunia dari tahun 1990 hingga 2022 [8].

Data epidemiologi menunjukkan bahwa obesitas mempengaruhi lebih dari 1,9 miliar orang secara global, dengan distribusi yang tidak merata melintasi berbagai negara dan kelompok sosiodemografi [1]. Di Amerika Serikat, prevalensi obesitas menunjukkan kenaikan sebesar tiga kali lipat selama lima dekade terakhir, dan proyeksi menunjukkan tren ini akan terus berlanjut meskipun berbagai upaya pengelolaan berat badan publik telah dilakukan [4].

b. Disparitas Demografis dan Sosial-ekonomi

Obesitas menunjukkan pola distribusi yang kompleks melintasi berbagai karakteristik demografis. Penelitian telah menunjukkan bahwa obesitas mempengaruhi individu dari semua latar belakang sosial-ekonomi, etnis, wilayah geografis, dan kelompok usia [12]. Namun, terdapat disparitas yang signifikan dalam prevalensi obesitas berdasarkan faktor-faktor antara lain tingkat pendapatan, pendidikan, dan ketersediaan layanan kesehatan [14].

2.1.5. Kesehatan dan Komorbiditas

a. Konsekuensi Metabolik Obesitas

Obesitas berkaitan erat dengan pengembangan berbagai gangguan metabolik yang dapat mengancam jiwa. Rohm *et al.* mengidentifikasi bahwa obesitas berkontribusi pada inflamasi kronis tingkat rendah hingga sedang, yang dapat memicu awal berbagai penyakit metabolik termasuk diabetes tipe 2, hipertensi, resistensi insulin, dislipidemia, dan kondisi kardiovaskular [20].

b. Risiko Mortalitas dan Morbiditas

Data epidemiologi menunjukkan bahwa obesitas berkontribusi terhadap sekitar 2.8 juta kematian global setiap tahunnya per September 2021 [16]. Obesitas juga dikaitkan dengan penyakit hati berlemak, risiko kanker, gangguan neurodegeneratif, dan penyakit jantung [15].

c. Beban Ekonomi Obesitas

Sistem Obesitas tidak hanya memiliki dampak kesehatan individual, tetapi juga menciptakan beban ekonomi yang substansial bagi sistem kesehatan global. Biaya yang tidak langsung dan langsung yang terkait dengan obesitas meliputi biaya perawatan medis, produktivitas yang hilang, dan kualitas hidup yang menurun [11].

2.1.6. Teori *Machine Learning* dalam Klasifikasi Medis

a. Definisi dan Paradigma *Machine Learning* (ML)

ML adalah cabang kecerdasan buatan yang memungkinkan komputer belajar dari data dan melakukan prediksi atau pengambilan keputusan tanpa harus diprogram secara khusus untuk tiap tugas. Dalam konteks klasifikasi obesitas, ML menyediakan kerangka kerja untuk mengidentifikasi pola kompleks dalam data multidimensional

yang melibatkan faktor-faktor seperti demografi, gaya hidup, riwayat medis, dan karakteristik fisiologis.

b. *Supervised Learning* sebagai Pendekatan Utama

Supervised learning merupakan paradigma pembelajaran yang paling relevan untuk klasifikasi obesitas. Dalam *supervised learning*, algoritma dilatih menggunakan *dataset* berlabel dimana fitur masukan dan label keluaran yang sesuai (status obesitas) sudah diketahui. Tujuan utama adalah mengembangkan model yang dapat memprediksi label *output* untuk data baru berdasarkan pola yang dipelajari dari data pembelajaran.

Christodoulou *et al.* melakukan ulasan sistematis yang menunjukkan bahwa meskipun ML sering dipromosikan sebagai superior dibandingkan metode statistik tradisional, dalam banyak kasus model prediksi klinis, regresi logistik masih dapat memberikan performa yang kompetitif [48]. Temuan ini menekankan pentingnya pemilihan algoritma yang sesuai dengan karakteristik data serta tujuan penelitian.

c. Klasifikasi Multikelas dalam Konteks Obesitas

Klasifikasi obesitas melibatkan kelas-kelas yang berbeda (*insufficient weight, normal weight, overweight levels*, dan *obesity types*), yang membuat masalah ini lebih kompleks dibandingkan klasifikasi biner. Klasifikasi multi-kelas memerlukan algoritma yang dapat menangani Batas keputusan ganda dan interaksi fitur yang kompleks.

2.1.7. Algoritma Klasifikasi untuk Prediksi Obesitas

a. *Logistic Regression* (LR)

LR termasuk metode *supervised learning* yang memprediksi probabilitas hasil *binary* maupun *multinomial.* LR bekerja dengan memodelkan logit (*log-odds*) dari probabilitas kejadian sebagai fungsi *linear* dari fitur:

$$\log = \frac{\mathrm{p}}{(1+\mathrm{p})} = \mathrm{a} + \beta 1 + \mathrm{x}2 \ldots . + \beta \mathrm{i} + \mathrm{x} \tag{2.2}$$

Sesuai yang diformulasikan pada (2.2), dimana p adalah probabilitas hasil, x_i adalah fitur, dan β_i adalah koefisien. LR sederhana, mudah diinterpretasikan, dan cocok untuk data dengan hubungan *linear*. Namun, performa LR menurun jika hubungan antar fitur bersifat *non-linear* dan kompleks [48]-[50].

b. *Support Vector Machine* (SVM)

SVM membangun *hyperplane* optimal untuk membagi kelas data pada ruang multidimensi. *Hyperplane* dipilih untuk memaksimalkan margin antara kelas:

$$H = wT(x) + b = 0 \tag{2.3}$$

Formula (2.3) merepresentasikan fungsi keputusan SVM, di mana vektor bobot *w* dan bias *b* digunakan untuk membentuk *hyperplane* terbaik yang memisahkan data antar kelas dengan jarak margin terbesar. SVM efektif untuk klasifikasi biner, tetapi dapat diperluas ke multi-kelas menggunakan strategi *one-vs-one* atau *one-vs-all*. Kelebihan SVM adalah kemampuan menangani data berdimensi tinggi dan kompleksitas *non-linear*, sedangkan kelemahannya adalah sensitivitas terhadap pemilihan kernel dan parameter [51]-[53].

c. *k-Nearest Neighbors* (k-NN)

k-NN adalah metode non-parametrik yang mengklasifikasikan data baru berdasarkan kedekatan dengan k tetangga terdekat. Ukuran jarak yang umum digunakan adalah *Euclidean Distance*:

$$d(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{2.4}$$

Fomula (2.4) menghitung jarak *Euclidean* antara data *train* dan data *test* pada algoritma k-NN, di mana nilai jarak tersebut menjadi dasar dalam menentukan k tetangga terdekat yang paling berpengaruh terhadap proses klasifikasi. Kelebihan k-NN adalah kesederhanaan dan fleksibilitas, tetapi performanya tergantung pada pemilihan parameter k dan distribusi data. k-NN cenderung rentan terhadap data *noisy* atau *outlier* [54]-[57].

d. *Decision Tree* (DT)

DT memodelkan keputusan melalui struktur pohon dengan membagi *dataset* berdasarkan fitur untuk meminimalkan impuritas, biasanya menggunakan entropi:

$$dE(S) = \sum_{i=1}^{c} - P_i log_2 P_i \quad (2.5)$$

Formula (2.5) menghitung nilai entropi sebagai ukuran ketidakmurnian data pada setiap node, sehingga pemilihan fitur pemisah dapat meningkatkan kualitas struktur pohon keputusan. DT mudah diinterpretasikan, memberikan wawasan visual mengenai pengaruh fitur, namun rentan terhadap *overfitting* terutama pada data berdimensi tinggi [61]-[63].

e. *Random Forest* (RF)

RF adalah metode *ensemble learning* yang membangun banyak pohon keputusan secara acak (*bagging*). Prediksi akhir dihasilkan dengan mengombinasikan hasil dari seluruh pohon:

$$y = (h1(x), h2(x), \ldots., hi(x)) \quad (2.6)$$

Formula (2.6) merepresentasikan mekanisme prediksi RF, di mana keluaran akhir diperoleh melalui agregasi hasil prediksi dari sejumlah pohon keputusan $hi(x)$, sehingga mampu meningkatkan akurasi model klasifikasi multi-kelas. RF efektif dalam menangani variabel yang berkorelasi, meningkatkan akurasi, dan mengurangi variansi model. Metode ini sangat sesuai untuk prediksi multi-kelas seperti obesitas [58]-[60].

f. Naïve Bayes (NB)

NB menggunakan asumsi independensi antar fitur dan teorema Bayes:

$$p(x) = \frac{p(h).p(h)}{p(x)} \quad (2.7)$$

Seperti pada formula (2.7) yang menggambarkan dasar probabilistik NB yang memanfaatkan teorema Bayes untuk menghitung probabilitas suatu kelas berdasarkan fitur masukan, dengan asumsi bahwa setiap fitur bersifat saling independen dalam proses klasifikasi. NB sederhana, cepat, dan sering digunakan pada klasifikasi teks dan data medis. Kelemahannya muncul jika fitur saling berkorelasi tinggi, yang dapat menurunkan akurasi [64]-[66].

2.1.8. Seleksi Fitur dengan Sparse Feature Selection and Ranking (SpFSR)

Seleksi fitur merupakan tahap krusial dalam *machine learning* untuk mengidentifikasi variabel paling relevan. Dalam penelitian ini, digunakan metode *Sparse Feature Selection and Ranking* (SpFSR). Metode ini bekerja dengan prinsip *Simultaneous Perturbation Stochastic Approximation* (SPSA) untuk mengevaluasi kontribusi setiap fitur melalui gangguan acak terkendali.

Fitur dinilai berdasarkan pengaruhnya terhadap pengurangan *error* prediksi, kemudian diurutkan dan dipilih *subset* teratas (top-k). Keuntungan penggunaan SpFSR antara lain:

1. Fitur menurunkan dimensi dataset sehingga proses pelatihan model menjadi lebih cepat serta mengurangi kompleksitas komputasi.
2. Memilih fitur unik dan relevan, mengurangi redundansi akibat korelasi antar fitur seperti Berat dan Tinggi.
3. Meningkatkan akurasi dan interpretabilitas model, sehingga dapat diaplikasikan dalam analisis epidemiologi dan sistem pendukung keputusan kesehatan.

Dalam konteks penelitian obesitas, SpFSR berhasil mengidentifikasi fitur utama seperti riwayat keluarga obesitas, tinggi badan, usia, frekuensi konsumsi makanan utama, dan pola makan, yang relevan secara klinis dan perilaku. Dengan kombinasi RF dan SpFSR, model prediksi menjadi lebih efisien, akurat, dan interpretatif, mendukung pengembangan strategi pencegahan obesitas di masyarakat.

2.1.9. *Preprocessing Data* dan Analisis Korelasi Antar Fitur

Dalam penelitian *machine learning*, *preprocessing data* merupakan tahap awal yang sangat penting untuk menjamin kualitas data, meningkatkan akurasi model, dan mempermudah interpretasi hasil. Salah satu langkah utama dalam *preprocessing* adalah analisis korelasi antar fitur. Korelasi antar fitur menunjukkan sejauh mana dua variabel memiliki hubungan linear, dengan nilai berkisar pada -1 dan 1. Korelasi bernilai 1 menunjukkan hubungan positif sempurna, −1 mencerminkan hubungan negatif sempurna, dan 0 berarti tidak terdapat korelasi. Analisis korelasi memiliki beberapa tujuan penting dalam konteks *machine learning*:

a) Identifikasi Multikolinearitas

Multikolinearitas terjadi ketika dua atau lebih fitur memiliki hubungan yang sangat kuat. Kondisi ini dapat mempersulit model untuk menilai pengaruh masing-masing fitur secara independen, sehingga prediksi menjadi kurang stabil dan interpretasi menjadi bias. Misalnya, dalam penelitian obesitas, variabel Berat dan Tinggi berkorelasi positif moderat sebesar 0,46, tinggi badan berpengaruh terhadap berat badan, tetapi keduanya juga memiliki kontribusi unik terhadap prediksi obesitas. Analisis korelasi membantu peneliti mengidentifikasi variabel yang bersifat renungan sehingga dapat dipertimbangkan untuk digabungkan atau dihapus.

b) Seleksi Fitur Efektif

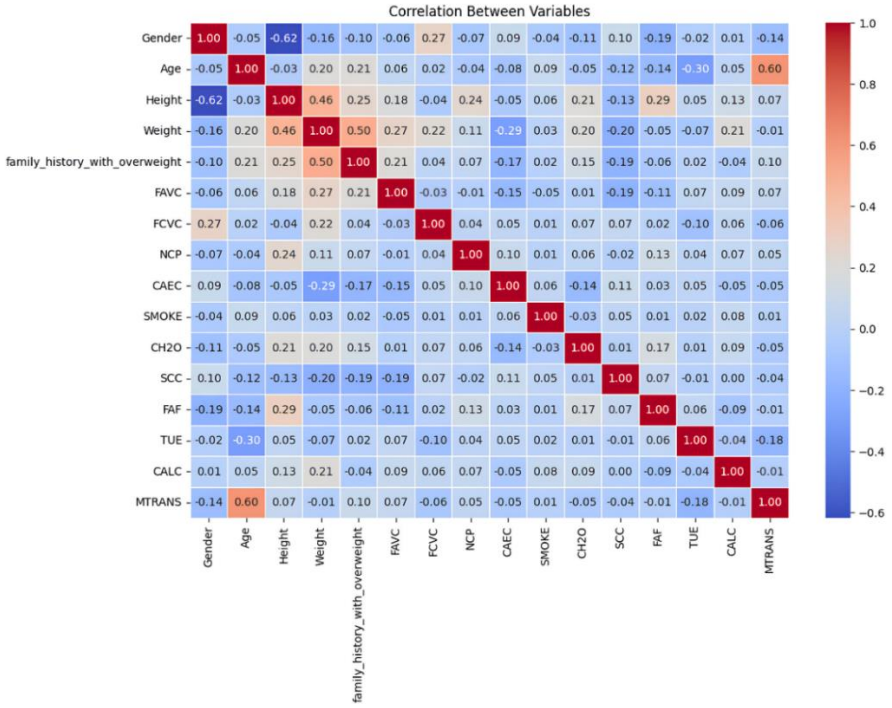
Hanya beberapa fitur yang memberikan kontribusi signifikan terhadap kemampuan prediksi model. Dengan menilai korelasi antar fitur, variabel yang tidak relevan atau redundan dapat dihapus, sehingga dimensi data berkurang, model menjadi lebih sederhana, dan performa prediksi tetap optimal. Misalnya, korelasi negatif antara Jenis Kelamin dan Tinggi Badan (-0,62) menunjukkan adanya pola distribusi tertentu yang perlu diperhatikan sebelum memasukkan fitur ini ke dalam model.

c) Mencegah *Overfitting*

Fitur yang sangat berkorelasi dapat menyebabkan model *overfitting*, di mana akurasi tinggi tercapai pada data pelatihan tetapi menurun pada data baru. Dengan menghapus atau menyesuaikan fitur yang berkorelasi tinggi, model menjadi lebih *robust* dan generalisasi prediksi meningkat. Dalam penelitian ini, fitur *MTRANS* dan Umur memiliki korelasi positif 0,60, sedangkan *TUE* dan Umur memiliki korelasi negatif -0,30, yang perlu diperhitungkan untuk meminimalkan bias dalam prediksi.

Analisis korelasi merupakan langkah penting untuk memahami pola dalam dataset, memandu seleksi fitur, dan membangun model yang lebih interpretatif serta efisien. Literatur sebelumnya menegaskan bahwa *preprocessing* dan penghapusan multikolinearitas dapat meningkatkan akurasi prediksi dan stabilitas model [41]-[45]. Berdasarkan Gambar 2.1, matriks korelasi menunjukkan hubungan antar fitur kebiasaan makan, kondisi fisik, dan variabel demografis yang memiliki tingkat korelasi berbeda-beda, baik positif maupun negatif. Pola korelasi tersebut memberikan

gambaran awal mengenai fitur-fitur yang berpotensi redundan atau berkontribusi signifikan terhadap model, sehingga dapat dijadikan dasar dalam proses seleksi fitur dan pencegahan *overfitting*.



Gambar 2.1. Matriks Korelasi Kebiasaan Makan, Kondisi Fisik, dan Variabel Demografis

Berdasarkan hasil visualisasi correlation matrix pada gambar 2.1, dapat dilihat hubungan antar variabel yang digunakan dalam dataset penelitian. Nilai korelasi berada pada rentang -1 hingga 1, di mana nilai mendekati 1 menunjukkan hubungan positif yang kuat, sedangkan nilai mendekati -1 menunjukkan hubungan negatif antar variabel.

Dari hasil analisis korelasi terlihat bahwa beberapa variabel memiliki hubungan yang cukup kuat. Salah satu hubungan yang terlihat cukup signifikan adalah antara height dan weight dengan nilai korelasi sekitar 0.46, yang menunjukkan bahwa semakin tinggi seseorang maka cenderung memiliki berat badan yang lebih besar. Hal ini merupakan hubungan yang wajar secara biologis.

Selain itu, variabel *family_history_with_overweight* juga menunjukkan hubungan positif dengan variabel weight, yang menunjukkan bahwa faktor riwayat keluarga dapat berpengaruh terhadap kondisi berat badan seseorang.

Sebagian besar variabel lainnya menunjukkan nilai korelasi yang relatif rendah, yang menandakan bahwa antar variabel memiliki hubungan yang tidak terlalu kuat. Kondisi ini cukup baik untuk proses pemodelan machine learning karena dapat mengurangi kemungkinan terjadinya *multicollinearity* antar fitur.

Hasil analisis korelasi ini kemudian digunakan sebagai dasar dalam proses feature selection menggunakan metode *SpFSR*, sehingga hanya fitur yang paling relevan yang digunakan dalam proses klasifikasi tingkat obesitas.

2.1.10. Relevansi dengan Penelitian Obesitas

Dimensionality Obesitas merupakan kondisi multifaktorial yang dipengaruhi oleh genetik, demografis, dan faktor gaya hidup, antara lain kebiasaan harian, aktivitas fisik, dan, pola makan. Integrasi *preprocessing*, korelasi fitur, klasifikasi *machine learning*, dan seleksi fitur canggih seperti SpFSR memungkinkan penelitian ini untuk:

1. Memahami pola dan hubungan antar fitur yang signifikan.
2. Membangun model prediksi yang *robust* dan dapat digeneralisasikan.
3. Memberikan wawasan praktis bagi tenaga kesehatan dan pembuat kebijakan dalam pencegahan obesitas.

Studi ini menekankan bahwa metode ML dapat digunakan tidak hanya untuk prediksi akurat, tetapi juga untuk memahami determinan utama obesitas, mendukung strategi intervensi berbasis data. Gambar 2.2 mrupakan algoritma seleksi fitur menggunakan Random Forest dilakukan dengan memanfaatkan nilai *feature importance* yang dihitung pada setiap lipatan *cross-validation* untuk memperoleh rata-rata kontribusi masing-masing fitur. Pendekatan ini memungkinkan pemilihan fitur yang paling relevan secara objektif, sehingga model klasifikasi obesitas yang dibangun menjadi lebih efisien, *robust*, dan memiliki kemampuan generalisasi yang lebih baik.

Algorithm 1: Feature Selection using Random Forest

```
1   INITIALIZE: RF ← RandomForestClassifier
2   I ← ZEROS(cv_folds, m)
3   // Feature importance matrix
4   FOR fold = 1 TO cv_folds DO
5       (X_train, X_val, y_train, y_val) ←
6       KFoldSplit(X, y, fold)
7       RF.fit(X_train, y_train); I[fold] ←
```

```
        RF.feature_importances_
    END FOR
    importance_mean ← MEAN(I)
    selected_idx ← ARGSORT(importance_mean,
    descending=True)[1:k]
    X_selected ← X[:, selected_idx]
    RETURN: X_selected, selected_idx,
    importance_mean[selected_idx]
```

Gambar 2.2. Algoritma Seleksi Fitur Menggunakan Random Forest

2.2. Tinjauan Penelitian Terdahulu (*Literature Review*)

Bagian ini meninjau penelitian terdahulu yang berkaitan dengan penggunaan *machine learning* (ML) untuk memprediksi tingkat obesitas serta penerapan metode seleksi fitur (*feature selection)* dalam mengoptimalkan akurasi model. Penelitian sebelumnya telah menunjukkan bahwa seleksi fitur berperan dalam meningkatkan performa model ML dengan cara menghapus fitur yang tidak relevan dan mengurangi beban komputasi. Elisseeff dan Guyon menegaskan bahwa pemilihan fitur mampu memperbaiki kinerja model dengan menghilangkan *input* yang tidak diperlukan [33]. Bolón-Canedo *et al.* juga menekankan bahwa *feature selection* membantu meningkatkan kemampuan interpretasi model, yang penting dalam pengambilan keputusan klinis [34]. Chandrasekhar dan Sahin menyebutkan bahwa reduksi dimensi dapat mengurangi beban komputasi [35], sedangkan Garibaldi dan Jonathan menyoroti nilai penting model sederhana untuk mempermudah penggunaan di bidang kesehatan [36]-[37].

Selain itu, beberapa penelitian terkait klasifikasi obesitas antara lain seperti yang dilakukan oleh Montañez *et al.*. Dengan menggunakan data genetik publik dengan lebih dari 6.622 varian genetik (SNPs) untuk memprediksi risiko obesitas, hasilnya menegaskan bahwa algoritma SVM menghasilkan performa optimal dengan nilai AUC 90,5%, sensitivitas 88,24%, dan spesifisitas 86,96%. Akan tetapi, memiliki kekurangan pada ukuran sampel yang kecil (N = 164) dan hanya menggunakan dua kategori kelas (normal dan risiko), yang dapat mengurangi detail klinis dari status BMI [28].

Jindal *et al.* menggabungkan tiga model analisis yaitu *Generalized Linear Model* (GLM), *Random Forest* (RF), dan *Partial Least Squares* (PLS) untuk

mengklasifikasikan tingkat obesitas berdasarkan data antropometri. Model ini mencapai akurasi rata-rata 89,68%, menunjukkan potensi besar dalam pengambilan keputusan berbasis data. Keterbatasannya adalah penggunaan variabel standar dan belum diuji pada populasi yang lebih beragam [29].

Marcos-Pasero *et al.* melakukan penelitian eksploratif terhadap 221 anak berusia 6-9 tahun dengan 190 variabel multidomain menggunakan algoritma *Random Forest* (RF) dan *Gradient Boosting Machine* (GBM). Faktor dominan yang mempengaruhi obesitas anak adalah status gizi keluarga, total energi, dan BMI orang tua. Keterbatasannya terletak pada desain penelitian *cross-sectional* dan data laporan mandiri yang berpotensi bias [30].

Thamrin *et al.* melakukan penelitian klasifikasi obesitas pada populasi dewasa Indonesia menggunakan dataset nasional RISKESDAS 2018 dengan tiga algoritma ML (CART, LR, dan *Naïve Bayes*). Penelitian ini menggunakan 21 variabel kategorikal dan menerapkan teknik SMOTE untuk mengatasi ketidakseimbangan data. Hasilnya menunjukkan bahwa *Logistic Regression* (LR) memiliki akurasi 72,2%, sedangkan CART memiliki sensitivitas tertinggi (82,7%). Kelemahan penelitian ini adalah belum menggunakan Teknik seleksi fitur dan model *ensemble* yang lebih kompleks untuk mengurangi bias prediksi [25]. Tabel 2.2 berikut merangkum berbagai studi terdahulu yang relevan, mencakup jenis *dataset* yang digunakan, algoritma *machine learning* yang diterapkan, teknik seleksi fitur yang digunakan, serta metrik evaluasi performa yang dilaporkan.

Tabel 2.2. Ringkasan Algoritma Optimal dari Studi yang Menggunakan Pembelajaran Mesin untuk Prediksi Obesitas

Referensi	Algoritma	*Dataset*	Faktor Resiko	Hasil
Montañez *et al.* [28]	*Gradient Boosting Machine*, RF, SVM *with Radial Basis Function Kernel*, k-NN, dan *Generalized Linear Models with Elastic Net Regularization*	800	*Single Nucleotide Polymorphisms* (13SNPs), *Gender*, dan *Genetic*	SVM: AUC = 90,5%.
Jindal *et al.* [29]	*Random Forest* dan *Ensemble*	1893	*Variations*, *Age*, *BMI*, *Weight*, dan *Height*	*Ensemble*: RF *accuracy* = 89,68%

Marcos-Pasero *et al.* [30]	*Gradient Boosting Machine* dan *Random Forest*	221	*Perception of Family Nutritional Status, TEI, TEE, BMI, Mother's Meal Frequency,* dan *IPAC*	RF: *accuracy* = 55,07%
Thamrin *et al.* [25]	*Classification* dan *Regression Trees, Logistic Regression,* dan *Naïve Bayes*	200	*Sweet drinks, Quick foods, Salty foods, Energy drinks, Sugary foods, Fatty/oily foods, Marital status, Age group, Work category meals, Grilled foods, Education, Smoking, Seasoning powders, Pre-served foods, Soft/carbonated alcoholic drinks, Diagnosed hypertension, Beverages, Mental emotional problems, Physical activity, Fruit* dan *vegetable intake* RISKESDAS	AUC = 79.79% LR: *accuracy* = 72,22%,

Berdasarkan hasil penelitian terdahulu, menunjukkan bahwa pemilihan fitur memberikan dampak besar terhadap peningkatan akurasi model *machine learning* untuk klasifikasi obesitas. Namun, sebagian besar penelitian sebelumnya masih terbatas pada jumlah sampel, variasi variabel, dan kurangnya penerapan teknik seleksi fitur canggih.

Berdasarkan hal tersebut, penelitian ini mengusulkan penerapan metode *Simultaneous Perturbation Feature Selection and Ranking* (SpFSR) yang diintegrasikan dengan algoritma *Random Forest*, untuk meningkatkan kinerja model dan akurasi dalam mendeteksi tingkat obesitas.

2.3. Objek Penelitian

Dataset yang digunakan menjadi objek penelitian ini yaitu "*Estimation of Obesity Levels Based on Eating Habits and Physical Condition*" yang dikompilasi oleh Palechor dan Manotas [40] dari UCI *Machine Learning Repository*. *Dataset* ini mencakup data dari 2,112 individu yang dikumpulkan dari Colombia, Peru, dan Mexico dengan rentang usia 14-61 tahun [32].

Dataset terdiri dari 17 variabel prediktor yang dikategorikan dalam tiga domain, yaitu kebiasaan makan (6 variabel), kondisi fisik (4 variabel), dan karakteristik demografis (4 variabel). Target variabel adalah tingkat obesitas yang diklasifikasikan menjadi tujuh kategori mulai dari *Insufficient Weight* hingga *Obesity Type III* dengan distribusi yang relatif seimbang.

Data yang dikumpulkan mencakup informasi gaya hidup, demografis, dan behavioral yang relevan untuk pengembangan model *machine learning* guna mengklasifikasikan tingkat obesitas berdasarkan faktor risiko yang beragam.