Perbandingan Algoritma Naïve Bayes Dan Random Forest Dalam Klasifikasi Obesitas Berdasarkan Faktor Gaya Hidup

Rizkya Nanda Prakoso[1], Shidqi Ikmal Rochim[2], Ari Subarna[3], Muhammad Erfan K.[4]

[1,2,3,4] Progam Studi Informatika, Universitas Bina Sarana Informatika
[1]rizkyaprakoso4college@gmail.com
[2]shidqiikmalr@gmail.com
[3]arisubarnas.alstar36@gmail.com
[4]erfan121204@gmail.com

***Abstrak*— Obesitas merupakan salah satu permasalahan kesehatan yang berkaitan erat dengan pola hidup modern. Tujuan dari penelitian ini adalah untuk mengevaluasi dan membandingkan kinerja algoritma Naïve Bayes dan Random Forest dalam mengelompokkan tingkat obesitas berdasarkan pola gaya hidup. Dataset diperoleh dari platform Kaggle yang memuat berbagai atribut terkait kebiasaan hidup, seperti pola makan dan aktivitas fisik. Penelitian dimulai dengan tahapan *data preprocessing* meliputi penghapusan atribut yang tidak relevan, transformasi label kelas menjadi bentuk kategorikal, serta pembersihan data kosong. Data kemudian dibagi menjadi data latih dan data uji. Model klasifikasi dibangun menggunakan aplikasi RapidMiner dengan algoritma *Naïve Bayes* dan *Random Forest*. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *classification error*. Berdasarkan hasil pengujian, algoritma *Random Forest* menghasilkan akurasi 83,23%, *precision* 83,93%, dan *recall* 82,46% pada kelas *Obesity*, dengan *classification error* sebesar 16,77%. Sementara itu, *Naïve Bayes* mencatat akurasi 76,09%, *precision* 73,21% dan recall 71,93% pada kelas *Obesity*, sementara itu, hasil *classification error* sebesar 23,91%. Hasil analisis dari *Weight by Information Gain* menunjukkan bahwa atribut dengan bobot tertinggi adalah usia (0,290), diikuti frekuensi konsumsi sayuran (0,272) dan jumlah makan utama per hari (0,232) yang berperan penting dalam klasifikasi obesitas. Penelitian ini menyimpulkan bahwa algoritma *Random Forest* lebih unggul dibandingkan *Naïve Bayes* dalam memprediksi obesitas berdasarkan faktor gaya hidup dan faktor lainnya yang dapat memicu obesitas.**

***Kata Kunci*— Obesitas, Gaya Hidup, Naive Bayes, Random Forest, Rapidminer.**

I. PENDAHULUAN

Obesitas merupakan masalah kesehatan global yang terus meningkat. Data dari Organisasi Kesehatan Dunia (WHO) menunjukkan bahwa pada tahun 2022, lebih dari 39% populasi dewasa di dunia mengalami obesitas, yang berisiko menyebabkan penyakit kronis seperti diabetes tipe 2, hipertensi, dan gangguan metabolik, serta mengurangi kualitas hidup individu. Fenomena obesitas ini telah berkembang menjadi pandemi global yang perlu mendapatkan perhatian khusus, termasuk di Indonesia yang menghadapi tantangan serupa [1].

Faktor gaya hidup, seperti pola makan tidak sehat, kurangnya aktivitas fisik, dan kebiasaan lainnya, sangat berkontribusi terhadap meningkatnya prevalensi obesitas. Pola makan yang berisiko seperti konsumsi makanan tinggi lemak, karbohidrat sederhana, serta rendah serat dapat menyebabkan obesitas jika tidak diimbangi dengan pengeluaran energi melalui aktivitas fisik. Di sisi lain, teknologi modern dan kebudayaan yang memudahkan kehidupan sehari-hari sering kali menurunkan tingkat aktivitas fisik yang diperlukan untuk menjaga berat badan ideal [2].

Mengidentifikasi pola hubungan antara faktor gaya hidup dan obesitas bisa menjadi tantangan karena sifat data gaya hidup yang kompleks dan beragam. Oleh karena itu, analisis data berbasis teknologi seperti *data mining*, yang menggunakan algoritma klasifikasi, dapat memberikan wawasan yang lebih jelas dan membantu pengambilan keputusan yang lebih baik dalam pencegahan dan penanganan obesitas. Dalam hal ini, algoritma *Naïve Bayes* dan *Random Forest* merupakan dua metode yang sering digunakan dalam klasifikasi data kesehatan, termasuk obesitas.

Naïve Bayes merupakan algoritma yang cepat dalam membuat dan menilai model klasifikasi, cocok untuk menangani masalah klasifikasi biner dan multikelas secara efisien. Di sisi lain, *Random Forest* sebagai metode *ensemble learning* yang menggabungkan banyak pohon keputusan, mampu menangani data dengan kompleksitas tinggi dan menghasilkan prediksi yang lebih akurat dengan mengurangi risiko *overfitting* [3] [4]. Kedua algoritma ini memiliki potensi untuk digunakan dalam analisis faktor-faktor gaya hidup yang memengaruhi obesitas, namun penting untuk membandingkan kinerja keduanya untuk menentukan algoritma yang paling efektif dalam klasifikasi obesitas.

Penelitian ini bertujuan untuk membandingkan efektivitas algoritma *Naïve Bayes* dan *Random Forest* dalam mengklasifikasikan obesitas berdasarkan faktor gaya hidup. Dengan menggunakan aplikasi RapidMiner untuk analisis data, penelitian ini akan mengevaluasi kinerja kedua algoritma tersebut dalam memprediksi obesitas, serta memberikan rekomendasi untuk pencegahan dan intervensi obesitas yang lebih efektif.

II. TINJAUAN LITERATUR

A. *Obesitas*

Obesitas merupakan kondisi penumpukan lemak yang berlebihan di dalam tubuh yang terjadi akibat ketidakseimbangan antara asupan kalori dan energi yang

dibakar. Obesitas merupakan masalah kesehatan yang kompleks, dapat menyebabkan berbagai kondisi seperti diabetes mellitus, hipertensi, dislipidemia, dan berbagai jenis kanker. Faktor utama yang memengaruhi obesitas adalah gaya hidup, yang meliputi pola makan tidak sehat, kurangnya aktivitas fisik, dan kebiasaan-kebiasaan buruk lainnya. Di era modern, pola makan yang mengarah pada konsumsi makanan tinggi kalori seperti *fast food* dan makanan manis sangat berkontribusi pada peningkatan obesitas [5].

Studi lainnya juga menunjukkan bahwa kebiasaan makan yang tidak sehat dan gaya hidup *sedentary* (kurang gerak) menjadi faktor utama yang berperan dalam peningkatan prevalensi obesitas. Selain itu, menurut H.L. Blum, faktor-faktor seperti perilaku, lingkungan, pelayanan kesehatan, dan faktor genetik saling berinteraksi dan berpengaruh terhadap kesehatan seseorang, dengan perilaku dan gaya hidup menjadi penentu utama dalam tingkat kesehatan individu [6].

B. *Faktor Penyebab Obesitas*

Terdapat beberapa faktor yang dapat menyebabkan obesitas, diantaranya:

1) Faktor Perilaku:

Kebiasaan makan berlebihan merupakan salah satu penyebab obesitas. Kondisi ini muncul ketika asupan kalori melebihi jumlah kalori yang digunakan oleh tubuh. Pada dasarnya, tubuh memerlukan kalori untuk menunjang kehidupan dan melakukan aktivitas fisik, namun untuk mempertahankan berat badan yang ideal, dibutuhkan keseimbangan antara asupan dan pembakaran energi. Ketidakseimbangan antara keduanya dapat menyebabkan kelebihan berat badan hingga obesitas [6].

Selain pola makan, kurangnya aktivitas fisik turut berperan dalam meningkatkan risiko obesitas. Aktivitas fisik, baik melalui rutinitas harian maupun latihan yang terstruktur, sangat penting untuk menjaga keseimbangan energi tubuh. Jika kebiasaan ini diterapkan secara konsisten sejak usia dini hingga lanjut usia, maka dapat memberikan manfaat kesehatan jangka panjang dan membantu mencegah akumulasi lemak berlebih. Kurangnya aktivitas sejak masa anak-anak dapat memicu kelebihan berat badan yang berpotensi berlanjut hingga usia dewasa [6].

Di samping itu, kebiasaan merokok juga dapat memperburuk kondisi obesitas. Kandungan nikotin dalam rokok dikaitkan dengan rendahnya tingkat aktivitas fisik, kurangnya konsumsi buah dan sayur, serta tingginya konsumsi alkohol. Kebiasaan ini dalam jangka panjang dapat memicu penumpukan lemak, obesitas sentral, dan resistensi insulin [7].

2) Faktor Lingkungan:

Lingkungan memiliki peran penting dalam meningkatkan risiko obesitas. Aktivitas fisik yang dilakukan hanya untuk mengikuti tren, tanpa menjadi bagian dari gaya hidup sehari-hari, menyebabkan rendahnya tingkat aktivitas di masyarakat modern. Gaya hidup yang minim gerakan ini menjadi salah satu faktor signifikan yang memicu terjadinya obesitas karena tubuh tidak membakar kalori secara optimal.

Selain itu, kebebasan dalam memilih makanan juga dapat memberikan dampak negatif jika tidak diimbangi dengan keputusan yang bijak. Keberagaman pilihan makanan yang ada mendorong seseorang untuk lebih sering mengonsumsi makanan berkalori tinggi namun rendah kandungan gizi, seperti *fast food* dan minuman manis. Konsumsi berlebihan tanpa kontrol yang memadai berkontribusi pada peningkatan berat badan dan risiko obesitas yang lebih tinggi.

Faktor keluarga juga turut memainkan peran krusial dalam membentuk kebiasaan makan dan gaya hidup. Anggota keluarga umumnya memiliki kebiasaan makan dan aktivitas fisik yang mirip, sehingga mencerminkan kombinasi pengaruh genetik dan lingkungan. Anak-anak biasanya meniru perilaku orang tua mereka, termasuk kecenderungan mengonsumsi makanan tinggi kalori dan menjalani gaya hidup sedentari. Jika pola ini terus berlangsung tanpa perubahan, anak-anak memiliki risiko lebih besar mengalami obesitas dan permasalahan kesehatan terkait di kemudian hari [6].

3) Faktor Edukasi dan Pelayanan Kesehatan:

Pelayanan kesehatan memainkan peran penting dalam upaya pencegahan obesitas. Namun, implementasi program pencegahan masih menghadapi berbagai tantangan, salah satunya adalah rendahnya kesadaran masyarakat akan seriusnya masalah obesitas. Persepsi yang kurang terhadap dampak jangka panjang obesitas membuat banyak individu mengabaikan pentingnya pola hidup sehat dan pencegahan sejak dini.

Selain itu, keterbatasan akses terhadap informasi yang akurat dan edukasi kesehatan juga memperburuk kondisi ini. Penyuluhan yang efektif diharapkan mampu meningkatkan pengetahuan masyarakat tentang pentingnya menjaga pola makan yang sehat dan aktivitas fisik yang teratur. Melalui edukasi yang tepat, perilaku makan yang tidak sehat dapat diubah secara perlahan, dan kesadaran masyarakat untuk menjaga kesehatan pun dapat ditingkatkan [6].

4) Faktor Genetik dan Biologis:

Faktor genetik menjadi salah satu penyebab utama obesitas, dengan usia sebagai salah satu aspek yang berkontribusi. Seiring bertambahnya usia, metabolisme tubuh cenderung melambat, sehingga proses pembakaran kalori menjadi kurang efisien. Akibatnya, kalori yang tidak terpakai lebih mudah disimpan sebagai lemak, yang pada akhirnya meningkatkan risiko obesitas.

Selain usia, jenis kelamin juga memengaruhi risiko obesitas. Perempuan memiliki risiko lebih tinggi mengalami obesitas dibanding laki-laki karena tingkat metabolisme basal mereka sekitar 10% lebih rendah. Umumnya, tubuh perempuan lebih banyak menyimpan makanan sebagai lemak, sementara tubuh laki-laki lebih cenderung memanfaatkannya untuk membentuk otot dan menghasilkan energi. Perbedaan ini menyebabkan

perempuan lebih mudah mengalami penimbunan lemak, terutama jika kurang melakukan aktivitas fisik.

Faktor genetik dari orang tua juga memiliki pengaruh besar terhadap risiko obesitas pada anak. Anak yang memiliki orang tua dengan berat badan berlebih memiliki kemungkinan 40–50% untuk mengalami obesitas. Risiko ini meningkat hingga 70–80% jika kedua orang tua mengalami obesitas. Selain faktor genetik, pola makan dan gaya hidup dalam keluarga juga cenderung diturunkan, sehingga memperkuat risiko obesitas [6].

C. *Data Mining*

Data mining merupakan suatu proses yang memungkinkan pengguna untuk mengakses dan menganalisis data dalam jumlah besar secara efisien. Proses ini memanfaatkan teknik statistik, kecerdasan buatan, dan *machine learning* untuk mengekstraksi informasi tersembunyi yang berguna dalam pengambilan Keputusan [8]. Proses data mining melibatkan serangkaian teknik untuk menemukan pola tersembunyi dalam data yang telah dikumpulkan. Teknik ini membantu mengungkap pengetahuan yang tidak dapat diidentifikasi secara manual, menjadikannya alat yang sangat berguna dalam pengolahan database besar [8].

Sebagai proses yang semi-otomatis, data mining terdiri dari beberapa tahap yang bersifat interaktif. Tahapan ini melibatkan kombinasi metode statistik dan kecerdasan buatan untuk mengidentifikasi informasi potensial. Dalam proses ini, pengguna dapat berinteraksi langsung atau melalui perantaraan *knowledge base* untuk mendapatkan hasil yang relevan [9].

D. *Algoritma Klasifikasi: Naïve Bayes dan Random Forest*

Klasifikasi merupakan proses untuk menemukan model, pola, atau fungsi yang dapat menggambarkan serta membedakan data sehingga dapat dikelompokkan ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Proses ini dimulai dengan mengidentifikasi karakteristik suatu objek, lalu objek dengan ciri serupa dimasukkan ke dalam kelas yang telah ditentukan sebelumnya. Dataset yang digunakan dalam klasifikasi harus memiliki label atau atribut target. Akurasi dari proses ini diukur berdasarkan persentase data yang diklasifikasikan ke kelas yang benar. Tujuan utama klasifikasi adalah memprediksi kelas dari suatu objek yang sebelumnya belum diketahui kelasnya [10].

1) Naïve Bayes:

Naive Bayes Classifier merupakan salah satu metode klasifikasi yang berlandaskan pada teorema Bayes, yang diperkenalkan oleh ilmuwan asal Inggris, Thomas Bayes. Metode ini mengandalkan konsep probabilitas dan statistik untuk memperkirakan kemungkinan kejadian di masa depan berdasarkan data atau pengalaman sebelumnya. Ciri khas dari *Naive Bayes* adalah asumsi independensi antara kondisi atau kejadian yang ada, yang membuatnya disebut "naive" atau naif [9].

Keunggulan utama dari metode *Naive Bayes* terletak pada kemampuannya untuk berfungsi secara optimal meskipun menggunakan data pelatihan yang relatif sedikit. Hal ini disebabkan karena algoritma ini hanya memerlukan varians dari setiap variabel dalam satu kelas untuk melakukan klasifikasi, tanpa harus menghitung seluruh matriks kovarians. Dengan demikian, Naive Bayes lebih efisien dibandingkan metode lain yang umumnya memerlukan volume data yang lebih besar [9]. Selain itu, *Naive Bayes* dapat digunakan untuk data kuantitatif maupun kualitatif, tidak memerlukan *data training* yang banyak, memiliki waktu perhitungan yang cepat, dan sangat mudah dipahami serta diterapkan dalam berbagai situasi.

Namun, meskipun memiliki banyak kelebihan, algoritma *Naive Bayes* juga memiliki beberapa kekurangan. Salah satunya adalah ketika probabilitas kondisionalnya bernilai nol, maka hasil prediksi juga akan menjadi nol. Selain itu, asumsi independensi antar variabel dapat mengurangi akurasi model, karena dalam kenyataannya banyak variabel yang saling berkorelasi.

Akurasi dari *Naive Bayes* juga tidak dapat diukur hanya dengan satu probabilitas saja, dan keputusan yang diambil sangat bergantung pada pengetahuan awal yang dimiliki. Hal ini membuat algoritma ini kurang efektif jika tidak ada informasi awal yang cukup. Selain itu, *Naive Bayes* lebih dirancang untuk mendeteksi kata-kata dalam teks dan tidak cocok untuk menganalisis data berupa gambar [9].

2) Random Forest:

Random Forest adalah sekumpulan pohon keputusan yang digunakan untuk klasifikasi dan prediksi data, dengan proses yang dimulai dari akar pohon hingga mencapai daun. Hasil klasifikasi ditunjukkan melalui bentuk setiap pohon yang terbentuk, sementara hasil prediksi diperoleh dari rata-rata keluaran seluruh pohon. Metode ini dikembangkan dari algoritma *Classification and Regression Tree* (CART) dengan menerapkan teknik *bagging* serta pemilihan fitur secara acak. [11].

Keunggulan *Random Forest* terletak pada kemampuannya mengatasi *noise* dan *missing values* serta fleksibilitasnya dalam menangani data besar. Meski demikian, interpretasi hasilnya bisa sulit dan algoritma ini memerlukan *tuning model* yang tepat untuk memperoleh hasil yang optimal. Meskipun demikian, *Random Forest* tetap menjadi pilihan populer dalam klasifikasi karena tingkat akurasinya yang tinggi [12].

E. *Teknik Pengujian Menggunakan Rapidminer:*

RapidMiner merupakan perangkat lunak analisis data yang memanfaatkan algoritma dan teknik *data mining* untuk menemukan pola dari data berukuran besar. Dengan mengintegrasikan metode statistik, kecerdasan buatan, dan sistem basis data, RapidMiner mempermudah pengguna dalam mengolah data melalui penggunaan elemen yang disebut operator. Operator ini dapat dikoneksikan ke node output untuk menampilkan hasil yang diinginkan.

RapidMiner dikembangkan menggunakan bahasa pemrograman Java dan mengandalkan file XML yang berisi rangkaian proses kerja berupa susunan operator. Terdapat lebih dari 500 operator yang tersedia, mencakup berbagai fungsi seperti pemuatan dan transformasi data, pembersihan data

(preprocessing), visualisasi, pembuatan model, serta evaluasi performa model.

Untuk menggunakan Rapidminer, pertama-tama data perlu diimpor ke dalam aplikasi. Setelah dataset terpasang, pengguna dapat menghubungkan beberapa operator, seperti operator untuk menangani data yang hilang. Setelah data lengkap, *modeling* atau analisis lainnya dapat dilakukan. Selama proses, pengguna akan dibawa ke halaman desain, di mana data dan operator yang telah disambungkan dapat dilihat. Hasil dari desain yang telah dijalankan dapat diperiksa pada halaman hasil [13].

F. *Penelitian Terkait*

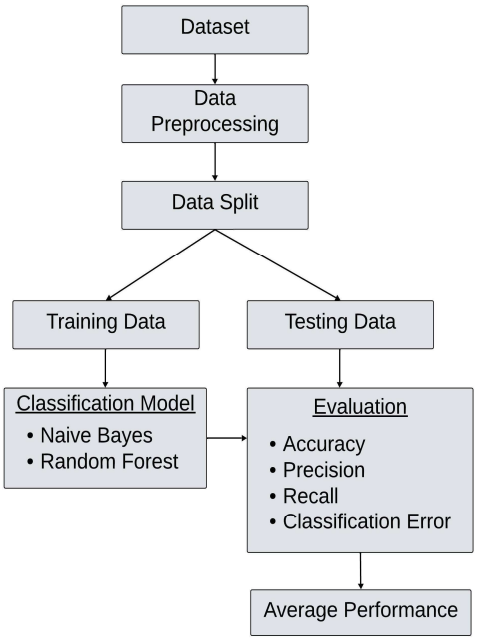
Penelitian sebelumnya memberikan landasan penting dalam pengembangan dan penerapan algoritma klasifikasi terhadap masalah obesitas. Salah satu studi relevan dilakukan dalam jurnal berjudul "Penggunaan Data Mining untuk Prediksi Tingkat Obesitas di Meksiko Menggunakan Metode Random Forest" [14]. Penelitian ini bertujuan mengembangkan model prediksi obesitas yang akurat guna mengidentifikasi individu dengan risiko tinggi. Meskipun fokus utamanya adalah pada algoritma *Random Forest*, penelitian ini juga membahas implementasi algoritma *K-Nearest Neighbors* (KNN) untuk klasifikasi, dengan pemanfaatan fungsi-fungsi di Python. Hasil penelitian ini menegaskan efektivitas algoritma *Random Forest* dalam menangani data obesitas [14].

Penelitian lain berjudul "Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest" [15] juga memberikan kontribusi signifikan. Meskipun fokus utamanya adalah prediksi *stroke*, penelitian ini membandingkan tiga algoritma klasifikasi, termasuk *Naïve Bayes* dan *Random Forest* yang juga digunakan dalam penelitian ini. Berdasarkan hasil pengujian terhadap 5.110 data responden, algoritma *Decision Tree* menunjukkan akurasi tertinggi sebesar 95,13%, diikuti oleh *Random Forest* dan *Naïve Bayes*. Studi ini menekankan pentingnya pemilihan algoritma yang tepat untuk meningkatkan akurasi klasifikasi pada data kesehatan [15].

Selain itu, jurnal "Efektivitas Algoritma AdaBoost dan XGBoost pada Dataset Obesitas Populasi Dewasa" [16] meneliti efektivitas dua algoritma *ensemble* dalam memprediksi obesitas. Hasil penelitian menunjukkan bahwa XGBoost memiliki performa yang lebih baik, dengan nilai akurasi, *precision*, dan *recall* sebesar 92%, dibandingkan AdaBoost yang hanya mencapai sekitar 40%. Temuan ini menggambarkan bahwa algoritma *ensemble* secara umum, termasuk *Random Forest*, cenderung menunjukkan performa tinggi dalam klasifikasi obesitas yang kompleks. Dengan merujuk pada studi studi tersebut, penelitian ini memperkuat relevansi perbandingan antara algoritma *Naïve Bayes* dan *Random Forest* dalam klasifikasi obesitas, khususnya yang berkaitan dengan faktor gaya hidup [16].

III. *METODE PENELITIAN*

Penelitian ini dilakukan melalui serangkaian tahapan sistematis untuk mengevaluasi dan membandingkan kinerja algoritma *Naive Bayes* dan *Random Forest* dalam mengklasifikasikan obesitas berdasarkan faktor-faktor gaya hidup. Gambar di bawah ini menunjukkan alur tahapan penelitian yang dimulai dari pengumpulan dataset, kemudian dilanjutkan dengan proses *data preprocessing* untuk membersihkan dan menyiapkan data. Setelah itu, data dibagi menjadi *data training* dan *data testing*. Model klasifikasi dibangun menggunakan algoritma *Naive Bayes* dan *Random Forest* pada data latih, lalu diuji menggunakan data uji. Evaluasi kinerja dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *classification error*. Tahap akhir adalah analisis rata-rata kinerja (average performance) untuk menentukan algoritma yang paling optimal.



Gbr. 1 Alur Penelitian.

A. *Persiapan Dataset*

Penelitian ini menggunakan dataset "Obesity Dataset" yang diunduh dari platform Kaggle dengan link https://www.kaggle.com/datasets/suleymansulak/obesity-dataset. Dataset ini terdiri dari 1.610 data dan memiliki beberapa atribut yang berkaitan dengan obesitas, gaya hidup, dan faktor lainnya yang dapat menjadi pemicu obesitas. Atribut yang terdapat dalam dataset diantaranya *Sex, Age, Height, Overweight/Obese Families, Consumption of Fast Food, Frequency of Consuming Vegetables, Number of Main Meals Daily, Food Intake Between Meals, Smoking, Liquid Intake Daily, Calculation of Calorie Intake, Physical Excercise, Schedule Dedicated to Technology, Type of Transportation Used,* dan *Class*. Berikut ini adalah 10 contoh dataset sebelum dilakukannya *data cleaning*:

Age	Height	Overweight_Obese_Family	Consumption_of_Fast_Food	Frequency_of_Consuming_Vegetables	Number_of_Main_Meals_Daily	Food_Intake_Between_Meals	Smoking	Liquid_Intake_Daily	Calculation_of_Calorie_Intake	Physical_Excercise	Schedule_Dedicated_to_Technology	Type_of_Transportation_Used	Class
18	155	2	2	3	1	3	2	1	2	3	3	4	2
18	158	2	2	3	1	1	2	1	2	1	3	3	2
18	159	2	2	2	1	3	2	3	2	2	3	4	2
18	162	2	2	2	2	2	2	2	2	1	3	4	2
18	165	2	1	2	1	3	2	1	2	3	3	2	2
18	176	1	1	1	1	4	2	2	2	4	3	4	2
19	152	2	2	3	1	1	2	3	2	2	3	2	2
19	158	2	2	3	2	4	2	1	1	1	3	3	2
19	159	2	2	2	2	2	2	1	2	2	3	4	2
19	162	2	2	2	2	2	2	2	1	2	3	4	3

Gbr. 2 Contoh 10 Dataset Sebelum *Data Cleaning*.

B. *Data Preprocessing*

Data Preprocessing merupakan tahap awal dalam *machine learning*, di mana data diolah dan dikonversi ke dalam bentuk tertentu agar dapat dengan mudah dibaca atau diproses oleh mesin [17]. Proses ini mencakup beberapa tahapan penting, yaitu *data cleaning*, transformasi data, dan reduksi data. Tahap *data cleaning* dilakukan dengan menghapus atribut-atribut yang tidak relevan seperti *Height, Liquid Intake Daily, Schedule Dedicated to Technology,* dan *Type of Transportation Used*.

Selanjutnya, dilakukan transformasi data dengan mengganti nilai numerik pada atribut *Class* (1/2/3/4) menjadi label kategorikal yang lebih representatif, yaitu *Underweight, Normal, Overweight,* dan *Obesity*. Tahap terakhir adalah reduksi data, yaitu menghapus data kosong atau tidak relevan guna meningkatkan kualitas dataset sehingga analisis yang dilakukan dapat menghasilkan output yang lebih akurat. Setelah melalui tahapan ini, dataset siap digunakan dalam proses klasifikasi. Berikut adalah 10 contoh dataset yang telah melalui proses *data cleaning*:

Sex	Age	Overweight_Obese_Family	Consumption_of_Fast_Food	Frequency_of_Consuming_Vegetables	Number_of_Main_Meals_Daily	Food_Intake_Between_Meals	Smoking	Calculation_of_Calorie_Intake	Physical_Excercise	Class
2	18	2	2	3	1	3	2	2	3	Normal
2	18	2	2	3	1	1	2	2	1	Normal
2	18	2	2	2	1	3	2	2	2	Normal
2	18	2	2	2	2	2	2	2	1	Normal
2	18	2	1	2	1	3	2	2	3	Normal
2	18	1	1	1	1	4	2	2	4	Normal
2	19	2	2	3	1	1	2	2	2	Normal
2	19	2	2	3	2	4	2	1	1	Normal
2	19	2	2	2	2	2	2	2	2	Normal
2	19	2	2	2	2	2	2	1	2	Overweight

Gbr. 3 Contoh 10 Dataset Sesudah *Data Cleaning*.

C. *Data Split*

Data split atau pembagian data adalah metode umum dalam validasi model, di mana dataset dipisahkan menjadi dua bagian terpisah, yaitu data pelatihan dan data pengujian, yang tidak saling tumpang tindih [18]. Dataset dalam penelitian ini dibagi dengan rasio 80:20 guna meminimalkan risiko *overfitting*, yaitu situasi ketika model menunjukkan kinerja sangat baik pada data pelatihan tetapi gagal menghasilkan prediksi akurat saat diuji pada data baru [19]. Sebanyak 80% data (1.288 data) digunakan untuk pelatihan, sementara 20% sisanya (322 data) digunakan untuk pengujian.

D. *Implementasi Algoritma*

Algoritma *Naive Bayes* dan *Random Forest* diterapkan pada dataset yang telah melalui proses *data preprocessing* dengan menggunakan aplikasi RapidMiner. Setelah proses klasifikasi selesai, hasil dari kedua algoritma dibandingkan untuk mengetahui metode mana yang memberikan kinerja terbaik dalam pengujian model.

E. *Evaluasi Kinerja*

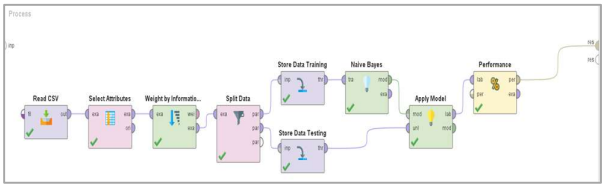
Evaluasi performa algoritma dilakukan dengan menggunakan beberapa metrik, yaitu akurasi, presisi, *recall*, dan *classification error.* Metrik-metrik ini digunakan untuk menilai seberapa baik model dapat mengklasifikasikan data secara tepat.

F. *Average Performance*

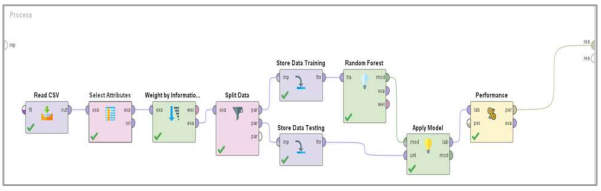
Interpretasi terhadap hasil evaluasi dilakukan untuk mendukung pencapaian tujuan penelitian. Tahap ini bertujuan untuk menganalisis dan menyimpulkan algoritma mana yang lebih optimal dalam mengklasifikasikan obesitas berdasarkan faktor gaya hidup.

IV. HASIL DAN PEMBAHASAN

Berikut disajikan proses pengujian model pada aplikasi Rapidminer:



Gbr. 4 Pengujian Model Menggunakan Algoritma Naive Bayes.



Gbr. 5 Pengujian Model Menggunakan Algoritma Random Forest.

Gambar di atas menunjukkan bahwa penelitian ini memanfaatkan sejumlah operator dalam aplikasi Rapidminer. Adapun beberapa operator yang digunakan antara lain:

1. *Read CSV*
   Operator ini digunakan untuk membaca file dataset berformat CSV yang menjadi sumber data utama penelitian. Setelah file berhasil diimpor, langkah selanjutnya adalah menetapkan peran atribut, yaitu menentukan atribut mana yang berfungsi sebagai label (kelas target) dan mengatur tipe data masing-masing atribut, seperti *polynominal, binominal*, atau *integer*, sesuai dengan karakteristik data yang dimiliki.

2. *Select Attributes*
   Operator ini digunakan untuk menyaring kolom yang akan digunakan dalam proses pelatihan model dan menghapus atribut yang tidak dibutuhkan.
3. *Weight by Information Gain*
   Operator ini digunakan untuk menentukan atribut mana yang paling berpengaruh terhadap target (label) dengan menggunakan metode *Information Gain*, sehingga membantu dalam proses *feature selection*.
4. *Split Data*
   Operator ini digunakan untuk memisahkan *data training* sebesar 80% dan *data testing* sebesar 20%.
5. *Store Data Training* dan *Data Testing*
   Operator ini digunakan untuk menyimpan *data training* dan *testing* secara permanen untuk keperluan pelatihan dan evaluasi model.
6. *Algoritma Naïve Bayes dan Random Forest*
   Operator ini membangun model klasifikasi berbasis algoritma *Naïve Bayes* dan *Random Forest* berdasarkan pola dari *data training* untuk memprediksi label.
7. *Apply Model*
   Operator ini digunakan untuk menghasilkan prediksi dari model berdasarkan *data testing* yang belum pernah dilihat sebelumnya.
8. *Performance (Classification)*
   Operator ini digunakan untuk menghitung metrik evaluasi seperti akurasi, *precision*, *recall*, *classification error*, dan metrik lainnya untuk mengukur kinerja model klasifikasi.

Setelah seluruh operator yang telah dijelaskan sebelumnya dimasukkan ke dalam alur proses, langkah selanjutnya adalah menjalankan proses untuk menguji performa model yang telah dibangun. Berikut ini adalah hasil dari evaluasi performa model berdasarkan pengujian yang dilakukan:

accuracy: 76.09%

	true Normal	true Overweight	true Obesity	true Underweight	class precision
pred. Normal	114	22	1	9	78.08%
pred. Overweight	13	84	15	0	75.00%
pred. Obesity	3	12	41	0	73.21%
pred. Underweight	2	0	0	6	75.00%
class recall	86.36%	71.19%	71.93%	40.00%	

Gbr. 6 Hasil Pengujian Model Menggunakan Algoritma Naive Bayes.

accuracy: 83.23%

	true Normal	true Overweight	true Obesity	true Underweight	class precision
pred. Normal	120	18	1	5	83.33%
pred. Overweight	11	91	9	0	81.98%
pred. Obesity	0	9	47	0	83.93%
pred. Underweight	1	0	0	10	90.91%
class recall	90.91%	77.12%	82.46%	66.67%	

Gbr. 7 Hasil Pengujian Model Menggunakan Algoritma Random Forest

Setelah kedua model dijalankan, model dengan algoritma *Naïve Bayes* menghasilkan akurasi sebesar 76,09% sedangkan model dengan algoritma *Random Forest* menghasilkan akurasi sebesar 83,23%. Perbandingan ini menunjukkan bahwa algoritma *Random Forest* memberikan hasil klasifikasi yang lebih akurat dibandingkan algoritma *Naïve Bayes*.

classification_error: 23.91%

	true Normal	true Overweight	true Obesity	true Underweight	class precision
pred. Normal	114	22	1	9	78.08%
pred. Overweight	13	84	15	0	75.00%
pred. Obesity	3	12	41	0	73.21%
pred. Underweight	2	0	0	6	75.00%
class recall	86.36%	71.19%	71.93%	40.00%	

Gbr. 8 Hasil *Classification Error* Algoritma Naive Bayes.

classification_error: 16.77%

	true Normal	true Overweight	true Obesity	true Underweight	class precision
pred. Normal	120	18	1	5	83.33%
pred. Overweight	11	91	9	0	81.98%
pred. Obesity	0	9	47	0	83.93%
pred. Underweight	1	0	0	10	90.91%
class recall	90.91%	77.12%	82.46%	66.67%	

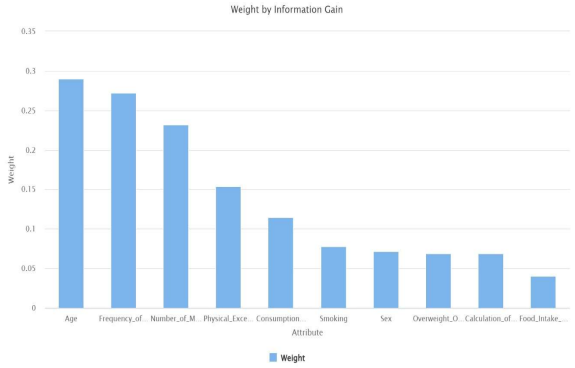
Gbr. 9 Hasil *Classification Error* Algoritma Random Forest.

Model dengan algoritma *Naïve Bayes* menghasilkan *classification error* sebesar 23,91%, sedangkan *Random Forest* sebesar 16,77%. Ini menunjukkan bahwa *Random Forest* memiliki tingkat kesalahan klasifikasi yang lebih rendah dibandingkan *Naïve Bayes*.

attribute	weight
Age	0.290
Frequency_of_Consuming_Vegetables	0.272
Number_of_Main_Meals_Daily	0.232
Physical_Excercise	0.154
Consumption_of_Fast_Food	0.115
Smoking	0.078
Sex	0.072
Overweight_Obese_Family	0.069
Calculation_of_Calorie_Intake	0.069
Food_Intake_Between_Meals	0.040

Gbr. 10 Hasil Weight by Information Gain.

Hasil perhitungan bobot menggunakan metode *Weight by Information Gain* menunjukkan bahwa atribut yang paling berpengaruh terhadap label klasifikasi adalah *Age* dengan nilai bobot tertinggi sebesar 0,290. Atribut ini memberikan informasi paling signifikan dalam membedakan individu yang mengalami obesitas dan yang tidak. Selanjutnya, atribut *Frequency_of_Consuming_Vegetables* menempati urutan kedua dengan bobot 0,272, diikuti oleh *Number_of_Main_Meals_Daily* pada urutan ketiga dengan bobot sebesar 0,232. Nilai-nilai bobot tersebut mengindikasikan bahwa frekuensi konsumsi sayuran dan jumlah makan utama per hari juga berperan dalam prediksi obesitas, meskipun kontribusinya tidak sebesar atribut usia.



Gbr. 11 Visualisasi Bobot Atribut Berdasarkan Weight by Information Gain.

Diagram ini menggambarkan kontribusi relatif masing-masing atribut dalam proses klasifikasi obesitas. Atribut dengan nilai bobot tertinggi berperan signifikan dalam membedakan antara data obesitas dan non-obesitas. Secara umum, hasil evaluasi menunjukkan bahwa model memiliki performa yang cukup baik.

Algoritma	Naïve Bayes	Random Forest
Accuracy	76,09%	83,23%
Class Precision	Underweight : 75,00%	Underweight : 90,91%
	Normal : 78,08%	Normal : 83,33%
	Overweight : 75,00%	Overweight : 81,98%
	Obesity : 73,21%	Obesity : 83,93%
Class Recall	Underweight : 40,00%	Underweight : 66,67%
	Normal : 86,36%	Normal : 90,91%
	Overweight : 71,19%	Overweight : 77,12%
	Obesity : 71,93%	Obesity : 82,46%
Classification Error	23,91%	16,77%
Kappa	0,635	0,745
Weighted Mean Precision	75,32%	85,04%
Weighted Mean Recall	67,37%	79,29%

Gbr. 12 Perbandingan Hasil Dari Algoritma Naive Bayes dan Random Forest.

Selain akurasi dan *classification error*, *Random Forest* unggul dalam *precision* dan *recall* di seluruh kelas. Pada kelas *Obesity*, *precision* mencapai 83,93%, lebih tinggi dari *Naïve Bayes* yang hanya 73,21%. Untuk kelas *Underweight*, *recall Random Forest* sebesar 66,67%, jauh melebihi *Naïve Bayes* sebesar 40%. Meski fokus utama pada empat metrik evaluasi, nilai *kappa, weighted mean precision*, dan *recall* turut memperkuat bahwa *Random Forest* lebih akurat dan konsisten dibandingkan *Naïve Bayes*.

V. KESIMPULAN

Berdasarkan hasil penelitian, algoritma *Naïve Bayes* dan *Random Forest* menunjukkan perbedaan performa yang signifikan dalam klasifikasi obesitas berdasarkan faktor gaya hidup. *Naïve Bayes* lebih efisien secara komputasi, namun kurang optimal dalam mengenali pola non-linear antar atribut. Sebaliknya, *Random Forest* menunjukkan kinerja lebih unggul dengan akurasi sebesar 83,23% dan *classification error* sebesar 16,77%, dibandingkan Naïve Bayes yang memiliki akurasi 76,09% dan *classification error* 23,91%. Dari segi evaluasi klasifikasi, *Random Forest* mencatat nilai *kappa* sebesar 0,745, lebih tinggi dibandingkan *Naïve Bayes* sebesar 0,635, yang menunjukkan tingkat kesesuaian prediksi yang lebih baik.

Selain itu, *Random Forest* juga memiliki *weighted mean precision* sebesar 85,04% dan *weighted mean recall* sebesar 79,29%, lebih tinggi dibandingkan *Naïve Bayes* yang masing-masing hanya mencapai 75,32% dan 67,37%. Atribut yang paling berpengaruh dalam prediksi obesitas adalah usia (0,290), diikuti frekuensi konsumsi sayuran (0,272) dan jumlah makan utama per hari (0,232), yang menunjukkan bahwa frekuensi konsumsi sayuran dan jumlah makan utama berperan dalam klasifikasi obesitas, meskipun pengaruhnya tidak sebesar faktor usia.

REFERENSI

[1] D. Hermawan *et al.*, *Mengenal Obesitas*. Penerbit Andi, 2020.
[2] T. Sudargo, H. Freitag, N. A. Kusmayanti, and F. Rosiyani, *Pola Makan dan Obesitas*. UGM PRESS, 2018.
[3] I. Maryani and I. Irmayansyah, "Penerapan Algoritma Naïve Bayes Untuk Penentuan Diagnosa Obesitas Pada Peserta Sosialisasi Deteksi Dini Penyakit Tidak Menular (PTM)," *TeknoIS J. Ilm. Teknol. Inf. dan Sains*, vol. 13, no. 2, pp. 234–248, Jul. 2023, doi: 10.36350/jbs.v13i2.200.
[4] A. Rifaldi, I. D. Satrio, and L. A. M. Huda, "Implementasi Algoritma K-nearest neighbor (KNN), Random Forest, Naive Bayes dan Decision Tree untuk mengklasifikasikan tingkat obesitas," *Researchgate.Net*, pp. 01–05, 2022, doi: http://dx.doi.org/10.13140/RG.2.2.29928.03841.
[5] A. B. Putri and A. Makmun, "Pola Makan terhadap Obesitas," *Indones. J. Heal.*, vol. 02, no. 01, pp. 68–76, 2021, doi: 10.33368/inajoh.v2i1.39.
[6] S. K. Saraswati *et al.*, "Literature Review : Faktor Risiko Penyebab Obesitas," *MEDIA Kesehat. Masy. Indones.*, vol. 20, no. 1, pp. 70–74, Feb. 2021, doi: 10.14710/mkmi.20.1.70-74.
[7] A. Susanto, E. N. Sari, and R. S. Prastiwi, "Analisis Hubungan Perilaku Merokok dengan Obesitas Sentral Pada Orang Dewasa Sehat di Suradadi Kabupaten Tegal," *PREPOTIF J. Kesehat. Masy.*, vol. 5, no. 2, pp. 1193–1198, Oct. 2021, doi: 10.31004/prepotif.v5i2.2461.
[8] C. Zai, "Implementasi Data Mining Sebagai Pengolahan Data," *J. Portal Data*, vol. 2, no. 3, pp. 1–12, 2022.
[9] A. F. Watratan, A. Puspita. B, and D. Moeis, "IMPLEMENTASI AlGORITMA NAIVE BAYES UNTUK MEMPREDIKSI TINGKAT PENYEBARAN COVID," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 1, pp. 7–14, 2020, doi: 10.55606/jurritek.v1i1.127.
[10] E. Adhi Guna *et al.*, "Implementasi Algoritma Decision Tree untuk Klasifikasi Data Evaluation Car Menggunakan Python," *J. Sist. Inf. dan Ilmu Komput.*, vol. 1, no. 4, pp. 167–177, 2023, doi: https://doi.org/10.59581/jusiik-widyakarya.v1i4. 1793.
[11] S. Mahmuda, "Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube," *J. Jendela Mat.*, vol. 2, no. 01, pp. 21–31, 2024, doi: 10.57008/jjm.v2i01.633.
[12] N. Novianti, S. P. A. Alkadri, and I. Fakhruzi, "Klasifikasi Penyakit Hipertensi Menggunakan Metode Random Forest," *Progresif J. Ilm. Komput.*, vol. 20, no. 1, p. 380, 2024, doi: 10.35889/progresif.v20i1.1663.
[13] Ainurrohma, "Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka," *Prism. Pros. Semin. Nas. Mat.*, vol. 4, pp. 493–499, 2021.
[14] E. Dwi *et al.*, "Penggunaan Data Mining untuk Prediksi tingkat Obesitas di Meksiko Menggunakan Metode Random Forest," *Agustus*, vol. 8, pp. 2549–7952, 2024.
[15] Y. Aulia, A. Andriyansyah, S. Suharjito, and S. W. Nensi, "Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest," *J. Ilmu Komput. dan Inform.*, vol. 3, no. 2, pp. 89–98, 2024, doi: 10.54082/jiki.90.

[16] C. E. Sukmawati, A. Fitri, N. Masruriyah, and A. R. Juwita, "Efektivitas algoritma AdaBoost dan XGBoost pada dataset obesitas populasi dewasa," vol. 6, no. 2, pp. 101–111, 2024, doi: 10.37905/jji.

[17] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.

[18] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.

[19] V. Diukarev and Y. Starukhin, "Proposed Methods for Preventing Overfitting in Machine Learning and Deep Learning," vol. 17, no. 10, pp. 85–94, 2024.