BAB III
METODE PENELITIAN

3.1. Desain Penelitian

Metode dalam penelitian ini mengusulkan pendekatan kuantitatif berbasis *machine learning* untuk memprediksi tingkat obesitas individu. Data penelitian diperoleh dari *Machine Learning Repository* yang berjumlah 2.112 sampel dengan 17 variabel yang mencakup kebiasaan makan, aktivitas fisik, serta karakteristik demografis. Seluruh data kemudian dibagi menggunakan teknik *stratified sampling*, yaitu 80% data *train* dan 20% data *test*, dengan proses validasi tambahan melalui 5-*fold cross-validation* guna memastikan model tidak mengalami *overfitting*.

Sebelum pemodelan dilakukan, peneliti melaksanakan tahap pra-pengolahan data berupa analisis korelasi untuk memahami hubungan antarvariabel dan mengidentifikasi potensi multikolinearitas. Hasil analisis korelasi tersebut kemudian menjadi dasar dalam proses pemilihan fitur. Selanjutnya, penelitian menerapkan metode SpFSR untuk menyaring fitur-fitur paling relevan terhadap prediksi obesitas, sekaligus mengurangi dimensi data agar model menjadi lebih efisien dan akurat.

Setelah fitur dipilih, berbagai algoritma klasifikasi diterapkan. Setiap model diuji dengan data *test* serta dinilai performanya dengan metrik evaluasi seperti akurasi, presisi, *recall*, dan *f1-score*. Gambar 3.1 menunjukkan alur penelitian yang dimulai dari pengumpulan *dataset*, tahap pra-pengolahan data dan analisis korelasi, hingga penerapan berbagai algoritma *machine learning* serta proses seleksi fitur menggunakan metode SpFSR. Alur penelitian disusun guna memperoleh model klasifikasi obesitas terbaik dengan performa prediksi yang maksimal.

3.2. Detail Metode dan Experimen

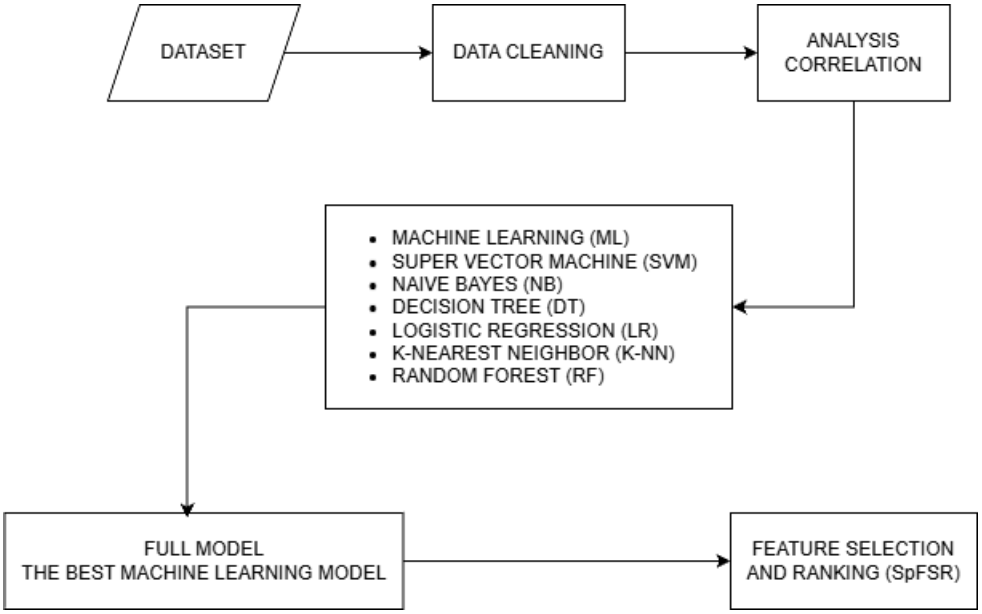
Pada penelitian ini, proses eksperimen dilakukan melalui beberapa tahapan utama yang dimulai dari tahap pengolahan data hingga evaluasi model klasifikasi. Dataset yang digunakan terlebih dahulu melalui tahap data preprocessing untuk memastikan kualitas data yang digunakan dalam penelitian. Tahap ini meliputi

pemeriksaan data, analisis korelasi antar variabel, serta penyesuaian format data agar dapat digunakan dalam proses pemodelan machine learning.

Selanjutnya dilakukan proses seleksi fitur menggunakan metode Simultaneous Perturbation Feature Selection and Ranking (SpFSR). Metode ini digunakan untuk menentukan fitur-fitur yang paling relevan terhadap proses klasifikasi tingkat obesitas. Proses seleksi fitur bertujuan untuk mengurangi dimensi data sehingga model klasifikasi dapat bekerja lebih efisien serta meningkatkan performa prediksi.

Setelah fitur yang relevan diperoleh, tahap eksperimen dilakukan dengan menerapkan algoritma Random Forest sebagai metode klasifikasi. Model dilatih menggunakan data training untuk mempelajari pola hubungan antara variabel input dan kategori tingkat obesitas. Selanjutnya model diuji menggunakan data testing untuk mengetahui kemampuan model dalam melakukan prediksi terhadap data yang belum pernah digunakan sebelumnya.

Performa model kemudian dievaluasi menggunakan beberapa metrik evaluasi, yaitu accuracy, precision, recall, dan F1-score, serta analisis confusion matrix untuk melihat kemampuan model dalam mengklasifikasikan setiap kategori obesitas. Hasil evaluasi ini digunakan untuk mengetahui pengaruh penggunaan metode seleksi fitur SpFSR terhadap peningkatan performa model klasifikasi.



Gambar 3.1. Alur Penelitian

3.3. Tahapan Penelitian

Penelitian ini dimulai dengan pengumpulan data dari *Machine Learning Repository* yang berjumlah 2.112 sampel dengan 17 variabel terkait kebiasaan makan, aktivitas fisik, dan karakteristik individu. Data kemudian melalui proses pra-pengolahan, termasuk analisis korelasi untuk melihat hubungan antarvariabel dan memastikan tidak terjadi multikolinearitas. Setelah itu, *dataset* dibagi menggunakan *stratified sampling* 80% data pelatihan dan 20% data pengujian, serta divalidasi menggunakan 5-*fold cross-validation* untuk menjaga kestabilan model. Selanjutnya diterapkan metode SpFSR untuk memilih fitur-fitur penting yang paling berpengaruh dalam klasifikasi obesitas. Fitur terpilih kemudian digunakan untuk melatih beberapa algoritma *machine learning*, yaitu *Logistic Regression*, k-NN, *Decision Tree*, *Random Forest*, SVM, dan *Naïve Bayes*. Tahap akhir penelitian adalah penilaian performa setiap model dengan metrik untuk menemukan model terbaik dalam memprediksi tingkat obesitas.

3.4. Pengumpulan *Dataset*

Penelitian ini menggunakan *dataset* sekunder yang berjudul "*Dataset for Estimating Obesity Levels Based on Dietary Habits and Physical Conditions in Individuals from Colombia, Peru, and Mexico*" yang tersedia secara publik melalui UCI *Machine Learning Repository*. *Dataset* ini dikompilasi oleh Fabio Mendoza

Palechor dan Alexis de la Hoz Manotas yang pada *link* https://doi.org/10.1016/j.dib.2019.104344.

Dataset ini dipilih karena memiliki karakteristik yang representatif untuk analisis prediksi obesitas, mencakup berbagai aspek yang mempengaruhi tingkat obesitas individu, termasuk kebiasaan makan, kondisi fisik, dan karakteristik demografis. Data dikumpulkan dari tiga negara Amerika Latin (Kolombia, Peru, dan Meksiko) yang memberikan variasi geografis dan budaya yang cukup untuk analisis *machine learning* yang *robust*.

Pemilihan *dataset* ini juga didasarkan pada kualitas dan kelengkapan data yang telah melalui proses kurasi dan validasi sebelumnya, sehingga mengurangi risiko bias dan nilai yang hilang yang dapat mempengaruhi performa model prediksi. *Dataset* ini telah digunakan dalam berbagai penelitian sebelumnya di bidang kesehatan masyarakat dan *machine learning*, menunjukkan kredibilitas dan reliabilitasnya untuk penelitian ilmiah.

3.4.1. *Data Cleaning*

Proses data *cleaning* merupakan tahap awal yang penting sebelum data digunakan dalam analisis dan pemodelan *machine learning.* Tahap ini bertujuan untuk memastikan bahwa data berada dalam kondisi bersih, konsisten, dan layak olah sehingga tidak menimbulkan bias atau kesalahan pada hasil penelitian. Data *cleaning* diawali dengan pemeriksaan struktur *dataset* untuk memahami jenis variabel, jumlah observasi, serta mendeteksi kejanggalan data. Pada tahap ini dilakukan eksplorasi awal guna mengidentifikasi tipe data, distribusi nilai, serta kemungkinan adanya anomali.

Selanjutnya, dilakukan penanganan terhadap data hilang. Jika jumlah data hilang relatif kecil, data tersebut dapat dihapus. Namun apabila jumlahnya signifikan, dilakukan imputasi menggunakan metode statistik yang sesuai dengan jenis variabel. Proses berikutnya adalah penghapusan data duplikat yang dapat menyebabkan bias pada model, terutama dalam proses klasifikasi. Tahap berikutnya mencakup pemeriksaan konsistensi dan kewajaran nilai data. Peneliti memastikan bahwa format data seragam serta tidak terdapat nilai yang tidak masuk akal pada variabel numerik. Selain itu, dilakukan deteksi *outlier* untuk mengidentifikasi nilai ekstrem yang berpotensi memengaruhi performa model. *Outlier* yang disebabkan oleh kesalahan

pencatatan dihapus, sedangkan yang merepresentasikan kondisi nyata tetap dipertahankan.

Sebagai langkah akhir, dilakukan standardisasi atau normalisasi data numerik agar seluruh variabel berada pada skala yang sebanding, khususnya untuk algoritma yang sensitif terhadap perbedaan skala. *Dataset* yang telah dibersihkan kemudian disimpan dan didokumentasikan untuk memastikan proses dapat direplikasi dan siap digunakan pada tahap pemodelan *machine learning.*

3.5. *Analysis Correlation*

Tahap ini dilakukan guna memahami hubungan antar variabel dalam *dataset* sebelum proses pemodelan *machine learning.* Analisis ini bertujuan untuk memberikan gambaran mengenai pola keterkaitan antar fitur serta hubungannya dengan variabel target, sehingga peneliti dapat memahami struktur data secara lebih mendalam. Proses ini dilakukan setelah data melalui tahap pembersihan, sehingga nilai yang dianalisis berada dalam kondisi valid dan konsisten. Peneliti terlebih dahulu mengidentifikasi karakteristik setiap variabel, kemudian menghitung nilai korelasi guna mengetahui tingkat hubungan antar variabel. Nilai korelasi digunakan sebagai indikator kekuatan dan arah hubungan, positif maupun negatif, yang menggambarkan sejauh mana perubahan suatu variabel berhubungan dengan perubahan variabel lain.

Hasil perhitungan korelasi kemudian divisualisasikan dalam bentuk matriks korelasi untuk memudahkan interpretasi hubungan antar fitur. Melalui visualisasi ini, peneliti dapat dengan cepat mengidentifikasi variabel-variabel yang memiliki hubungan kuat, sedang, maupun lemah. Analisis ini membantu dalam mengenali fitur-fitur yang saling berkaitan erat serta fitur yang relatif berdiri sendiri. Selain melihat hubungan antara fitur dengan variabel target, analisis korelasi juga digunakan untuk mendeteksi adanya multikolinearitas antar fitur independen. Fitur-fitur yang memiliki korelasi sangat tinggi berpotensi menyebabkan redundansi informasi dan dapat memengaruhi kestabilan model. Oleh karena itu, hasil analisis korelasi digunakan sebagai dasar dalam menentukan apakah suatu fitur perlu dipertahankan, dikurangi, atau dipertimbangkan lebih lanjut pada tahap seleksi fitur.

Secara keseluruhan, tahap analisis korelasi berperan sebagai jembatan antara proses data *cleaning* dan pemodelan *machine learning.* Tahap ini membantu peneliti

memahami hubungan antar variabel, menyederhanakan struktur data, serta mendukung pengambilan keputusan dalam pemilihan fitur yang relevan.

3.6. *Machine Learning Implementation*

Tahap implementasi metode *machine learning* merupakan komponen inti dalam metodologi penelitian, di mana berbagai algoritma dengan karakteristik serta landasan teoretis yang berbeda diterapkan, disesuaikan, dan dievaluasi untuk menentukan solusi yang paling optimal bagi domain permasalahan yang diteliti. Pendekatan sistematis dalam pemilihan dan penerapan algoritma bertujuan untuk memastikan cakupan yang komprehensif terhadap berbagai paradigma pembelajaran, sehingga mampu menangkap beragam pola yang terdapat pada data. Setiap algoritma yang digunakan memiliki kelebihan dan keterbatasan masing-masing, yang menjadikannya sesuai untuk karakteristik data dan permasalahan tertentu.

a) *Support Vector Machine* (SVM)

SVM termasuk algoritma *supervised learning* yang membagi data ke dalam kelas tertentu dengan membentuk sebuah *hyperplane* optimal. Proses kerja SVM diawali dengan penyesuaian skala data melalui normalisasi atau standardisasi agar setiap fitur memiliki pengaruh yang seimbang. Selanjutnya, SVM memanfaatkan fungsi kernel untuk mentransformasikan data ke dalam ruang berdimensi lebih tinggi sehingga data yang sebelumnya tidak dapat dipisahkan secara linear menjadi dapat diklasifikasikan dengan lebih baik.

Pada tahap pelatihan, SVM membangun model berdasarkan data dari masing-masing kelas yang berada paling dekat dengan *hyperplane*, yang disebut *support vectors*. Model yang terbentuk kemudian diuji menggunakan data uji untuk mengevaluasi kemampuannya dalam mengklasifikasikan data baru. Melalui pendekatan ini, SVM mampu menghasilkan pemisahan kelas yang stabil dan akurat, khususnya pada dataset berdimensi tinggi dan memiliki pola data yang kompleks.

b) *Naive Bayes* (NB)

Proses kerja *Naïve Bayes* diawali dengan menghitung probabilitas awal (*prior*) dari setiap kelas obesitas pada data pelatihan. Selanjutnya, dihitung probabilitas

kemunculan setiap fitur terhadap masing-masing kelas. Pada tahap prediksi, NB menentukan kelas dengan probabilitas posterior tertinggi berdasarkan kombinasi fitur yang dimiliki oleh suatu individu.

Meskipun asumsi independensi fitur jarang terpenuhi sepenuhnya dalam data kesehatan, *Naïve Bayes* tetap mampu memberikan hasil klasifikasi yang cukup baik dengan komputasi yang sederhana dan efisien. Oleh karena itu, algoritma ini digunakan sebagai model pembanding dalam mengevaluasi kinerja metode klasifikasi obesitas pada penelitian ini.

c) *Decision Tree* (DT)

Proses klasifikasi obesitas menggunakan *Decision Tree* dimulai dengan memasukkan seluruh data latih yang telah melalui tahap praproses. Algoritma kemudian mengevaluasi setiap fitur, seperti kebiasaan makan, aktivitas fisik, dan karakteristik individu, untuk menentukan atribut yang paling efektif dalam memisahkan kelas obesitas. Data dibagi secara bertahap ke dalam cabang-cabang berdasarkan hasil evaluasi tersebut hingga terbentuk struktur pohon keputusan. Pada tahap akhir, setiap data diklasifikasikan ke dalam kategori obesitas tertentu berdasarkan jalur keputusan yang dilalui dari akar hingga simpul daun.

d) *Logistic Regression* (LR)

Pada *Logistic Regression*, proses dimulai dengan mempelajari hubungan antara variabel input dan kelas obesitas melalui pendekatan *probabilistik*. Setiap fitur diberikan bobot untuk merepresentasikan pengaruhnya terhadap kemungkinan seseorang termasuk dalam kategori obesitas tertentu. Model menghitung nilai kemungkinan setiap kelas, kemudian menentukan kelas akhir berdasarkan kemungkinan tertinggi. Proses ini memungkinkan *Logistic Regression* mengidentifikasi kecenderungan risiko obesitas secara sistematis dan terukur.

e) *K-Nearest Neighbor* (K-NN)

Pada tahap penerapan algoritma K-NN, proses dimulai dengan memastikan bahwa seluruh data telah melalui tahap pembersihan dan normalisasi. Normalisasi dilakukan karena K-NN mengandalkan perhitungan jarak antar data, sehingga perbedaan skala antar fitur dapat memengaruhi hasil klasifikasi. Setelah data siap, ditentukan nilai k sebagai nominal tetangga terdekat yang dihitung pada proses

prediksi. Nilai *k* ini ditentukan melalui beberapa pengujian untuk memperoleh performa yang paling stabil. Berbeda dengan algoritma lain, K-NN tidak melakukan proses pelatihan model secara eksplisit karena bersifat *lazy learner,* sehingga model hanya menyimpan data latih untuk digunakan pada tahap prediksi.

Pada tahap pengujian, setiap data baru dihitung jaraknya terhadap seluruh data latih menggunakan metrik jarak tertentu, seperti *Euclidean Distance.* Selanjutnya, *k* data dengan jarak terdekat dipilih sebagai tetangga, dan kelas obesitas ditentukan berdasarkan kelas yang paling dominan di antara tetangga tersebut. Proses ini dilakukan untuk seluruh data uji guna memperoleh hasil klasifikasi. Berdasarkan hasil prediksi, performa model kemudian dievaluasi menggunakan metrik. Apabila hasil yang diperoleh belum optimal, peneliti dapat menyesuaikan nilai *k* atau metode perhitungan jarak untuk meningkatkan kinerja model.

f) *Random Forest* (RF)

Random Forest memulai proses klasifikasi dengan membangun banyak pohon keputusan secara acak menggunakan *subset* data dan fitur yang berbeda. Setiap pohon melakukan klasifikasi obesitas secara mandiri berdasarkan pola yang dipelajari dari data latih. Hasil prediksi dari seluruh pohon kemudian digabungkan menggunakan mekanisme pemungutan suara mayoritas. Pendekatan ini meningkatkan stabilitas dan akurasi model dalam mengklasifikasikan tingkat obesitas.

3.7. *Full Model (The Best Machine Learning Model)*

Tahapan penentuan *Full Model* Terbaik *Machine Learning* dimulai setelah seluruh algoritma yang digunakan, yaitu *Support Vector Machine, Naïve Bayes, Decision Tree, Logistic Regression,* dan *Random Forest,* menyelesaikan proses pelatihan dan pengujian menggunakan *dataset* yang sama. Pada tahap ini, peneliti mengumpulkan hasil performa masing-masing model berdasarkan metrik evaluasi. Seluruh hasil tersebut disusun dalam bentuk tabel perbandingan untuk memudahkan identifikasi model dengan performa terbaik. Selain itu, analisis *confusion matrix* dilakukan untuk memahami pola kesalahan prediksi, termasuk kemampuan model dalam mengenali setiap kelas obesitas dan keseimbangan performa antar kelas, terutama pada kondisi data yang tidak seimbang.

Setelah evaluasi awal dilakukan, peneliti melanjutkan ke tahap validasi menggunakan teknik *cross-validation* guna memastikan kestabilan dan kemampuan generalisasi model. Pada tahap ini, setiap algoritma dilatih dan diuji secara berulang dengan pembagian data yang berbeda, lalu menghitung rata-rata performa serta deviasi standar dari masing-masing model. Model yang menunjukkan performa konsisten dengan nilai rata-rata tinggi dan deviasi standar rendah diinterpretasikan sebagai model yang paling stabil. Berdasarkan hasil evaluasi dan validasi tersebut, satu algoritma ditetapkan sebagai *Full Model* Terbaik *Machine Learning*, yang selanjutnya digunakan sebagai model utama dalam penelitian dan menjadi dasar untuk analisis lanjutan serta penarikan kesimpulan.

3.8. *Feature Selection and Ranking (SpFSR)*

Tahapan SpFSR dimulai setelah proses pembersihan data, analisis korelasi, dan pengujian awal model *machine learning* selesai dilakukan. Pada tahap ini, peneliti berfokus pada penyaringan fitur untuk mengurangi variabel yang tidak relevan, redundan, atau memiliki kontribusi rendah terhadap performa model. Proses awal dilakukan dengan mengidentifikasi hubungan antara fitur dan variabel target, serta mendeteksi adanya multikolinearitas antar fitur melalui analisis statistik dan korelasi. Tujuan dari tahap ini adalah memastikan bahwa fitur yang dipertahankan benar-benar memberikan informasi penting bagi proses klasifikasi obesitas dan tidak menambah kompleksitas model secara tidak perlu.

Setelah fitur awal tersaring, SpFSR dilanjutkan dengan proses pemeringkatan fitur berdasarkan tingkat kepentingannya. Setiap fitur diberikan skor Tingkat kepentingan fitur menggunakan metode penilaian tertentu, kemudian diurutkan dari yang paling berpengaruh hingga yang paling rendah kontribusinya. Berdasarkan hasil pemeringkatan tersebut, peneliti menetapkan sejumlah fitur terbaik untuk digunakan dalam pemodelan lanjutan. Model kemudian diuji kembali menggunakan fitur terpilih untuk mengevaluasi perubahan performa sebelum dan sesudah penerapan SpFSR. Jika terjadi peningkatan akurasi, efisiensi komputasi, atau penurunan *overfitting*, maka fitur hasil seleksi ditetapkan sebagai fitur akhir dan digunakan pada tahap pemodelan *final*, sekaligus didokumentasikan sebagai bagian dari hasil penelitian.